Tweeting Islamophobia: Islamophobic hate speech amongst followers of UK political parties on Twitter

Bertram Vidgen

Wolfson College, University of Oxford

Thesis submitted in partial fulfilment of the requirement for the degree of DPhil in Information, Communication and the Social Sciences in the Oxford Internet Institute at the University of Oxford

Supervisors

Dr Taha Yasseri (Oxford Internet Institute & Alan Turing Institute)

Prof Helen Margetts OBE (Oxford Internet Institute & Alan Turing Institute)

January 2019

Words: 99,609

Acknowledgments

I always thought that writing the Acknowledgments would be tricky – Who should I include? How can I express my gratitude in so few words? What specifically should I thank everyone for? But it turns out my concerns were misplaced. Although many people have been incredibly helpful, writing these acknowledgments has been very easy.

I would like to express genuine gratitude to my supervisors, Dr. Taha Yasseri and Prof. Helen Margetts OBE, for your help, support and insights, and for your roles in my academic development. It has been a privilege working with you, and through your supervision I have learnt more than I ever anticipated when I started my PhD. Taha, you have provided incredible and incisive input across all of the methods, analysis and interpretation – the computational aspect of this work would simply not be possible without you. Helen, you have helped to mould this PhD and drive the theoretical arguments. Your critical insights and ability to identify the real value in any analysis are second to none. I appreciate all of the time you have each put in to this project, and I sincerely look forward to working with both of you in the future.

I would like to thank my three assessors from Transfer, Confirmation and Viva: Prof. Matt Williams, Dr. Jonathan Bright and Dr. Michael Biggs. I have hugely appreciated discussing my work with each of you and your input has been incredibly valuable in developing and honing the focus of my PhD. A special thanks also goes out to Dr. Scott Hale, who has not been formally involved in this work but has offered some excellent advice and guidance throughout, and Prof. Marie Gillespie, who has really helped to shape a jumble of ideas into a full thesis.

Many peers have helped me with this PhD, from annotating tweets to advising on style to identifying relevant papers to read. I would like to thank each of you for your help and thoughts – I hope I have, and can again in the future, reciprocate! In particular, thanks to Suzanne van Geuns for reading the fifth chapter and providing such discerning feedback. But over the past year one person stands out for having been unceasingly constructive, supportive and helpful: Margie Cheesman, your thoughts and insights have been transformative. I cannot express how wonderful you have been.

The Oxford Internet Institute (OII) is made up of many excellent researchers – many of whom I respect and have learnt a great deal from. In particular, Prof. Vicky Nash and Prof. Gina Neff have been motivational and inspiring, as both leaders and researchers. The OII is also made up of many excellent non-research staff who have truly made my PhD far easier and more enjoyable than it would have been otherwise. There are too many to list – but, in particular, Duncan, Emily, David, Ornella, Tim, Victoria, Laura and Jordan, thank you so much for all of your work. Thanks also to the ESRC for their generous funding, the Alan Turing Institute for funding and technical assistance, and to Wolfson College for its support.

I'd especially like to thank my family. Theo, you have been fantastic in your support and interest, and I really hope to one day repay all of your kindness. Dad, you have been instrumental in my journey towards completing a PhD – from encouraging me to do one in the first place to helping at some pivotal moments in the research and writeup. Mum, at 27 years old it still amazes me how much you help me with every part of my life, PhD included! I cannot thank you all enough. As much as anything else, this PhD is a product of your (collective) unwavering support.

Completing a PhD has been one of the biggest undertakings of my life. There have been some (i.e. many) difficult times and some (i.e. very many!) immensely gratifying moments. Throughout all of it, I have felt extremely privileged to study at the Oxford Internet Institute and the Alan Turing Institute, both of which are fantastic interdisciplinary institutions. I could not imagine completing this research anywhere else.

Thank you to everyone for all of your incredible help!

Dedication

This thesis is dedicated to my parents, Richard and Heather.

Abstract

The great promise of social media platforms such as Twitter is to connect people separated across time and space. This has had far-ranging consequences for politics by changing discursive, participative and organisational practices. However, despite much early techno-optimism about platforms like Twitter, concerns are growing that they enable harmful, hateful and divisive behaviours. In this thesis, I focus on one of the most concerning and harmful behaviours on Twitter and in politics more broadly: Islamophobic hate speech. The socio-political consequences of hate speech are deeply concerning, and include causing harm to targeted victims, spreading divisiveness, and normalizing dangerous and extremist ideas.

The aim of this thesis is to enhance our understanding of the nature and dynamics of Islamophobic hate speech amongst followers of UK political parties on Twitter. I study four parties from across the political spectrum: the BNP, UKIP, the Conservatives and Labour. I make three main contributions. First, I define Islamophobia in terms of negativity and generality, thus making a robust, theoretically-informed contribution to the study of a deeply contested concept. This argument informs the second contribution, which is methodological: I create a multi-class supervised machine learning classifier for Islamophobic hate speech. This distinguishes between weak and strong varieties and can be applied robustly and at scale.

My third contribution is theoretical. Drawing together my substantive findings, I argue that Islamophobic tweeting amongst followers of UK parties can be characterised as a *wind system* which contains Islamophobic *hurricanes*. This analogy captures the complex, heterogeneous dynamics underpinning Islamophobia on Twitter, and highlights its devastating effects. I also show that Islamist terrorist attacks drive Islamophobia, and that this affects followers of all four parties studied here. I use this finding to extend the theory of cumulative extremism beyond extremist groups to include individuals with mainstream affiliations. These contributions feed into ongoing academic, policymaking and activist discussions about Islamophobic hate speech in both social media and UK politics.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Aikake Information Criterion
ANOVA	Analysis of Variance
API	Application Programming Interface
ARIMA	Auto-Regressive Moving Average
AUC	Area Under the Curve
BBC	British Broadcasting Company
BIC	Bayesian Information Criterion
BNP	The British National Party
BOW	Bag of Words
CBOW	Continuous Bag of Words
CPS	Crown Prosecution Service
CSEW	Crime Survey for England & Wales
CST	Community Security Trust
ECRI	European Commission against Racism and Intolerance
EDL	English Defence League
EU	European Union
GIF	Graphics Interchange Format
GloVe	Global Vectors
HM	Her Majesty's
ISIS	Islamic State in Iraq and Syria
LM	Latent Markov
MANOVA	Multivariate Analysis of Variance
MEND	Muslim Engagement & Development
MEP	Member of the European Parliament
MP	Member of Parliament
NASA	National Aeronautics and Space Administration
NB	Naïve Bayes
NLP	Natural Language Processing
OLS	Ordinary Least Squares

REST	Representational State Transfer
RQ	Research Question
SNP	Scottish National Party
SVM	Support Vector Machines
UK	United Kingdom
UKIP	United Kingdom Independence Party
USA	United States of America

Chapter 1 | Introduction

The great promise of social media platforms such as Twitter is to connect people separated across time and space by enabling them to share, observe and interact with multimedia and text-based content quickly and easily. This has had far-reaching consequences on politics, changing discursive, participatory and organisational practices and reconfiguring collective action (Castells, 2015; Chadwick & Stromer-Galley, 2016; Margetts, 2017a; Margetts, John, Hale, & Yasseri, 2015). Twitter, in particular, is one of the most important and widely used platforms for political activities (Cihon & Yasseri, 2016). Elected politicians use it to communicate with the public (Jungherr, 2016) and citizens use it to engage in political talk more broadly (Wright, Graham, & Jackson, 2017) and to form deliberative publics (Mckelvey, Digrazia, & Rojas, 2014).

Despite much early techno-optimism about the positive impact of Twitter on contemporary politics, concerns that it enables harmful, hateful and divisive behaviours are growing (Crilley & Gillespie, 2019; Hemsley, Jacobson, Gruzd, & Mai, 2018; Howard & Parks, 2012). In this thesis, I focus on a deeply concerning and harmful behaviour on Twitter and in politics more generally: Islamophobic hate speech. This is one of many forms of identity-based hate which have raised concerns in political discourses, including misogyny, anti-LGBTQA, xenophobic prejudice and racism. All forms of hate should be studied, monitored, challenged and countered, not least as it is likely that they share many affinities, such as how they are articulated, who they impact, their political logics and their causes.

In studying Islamophobic hate, I start from a simple premise – that words matter. Things can be done with words; harm can be inflicted, support can be provided, and ideologies and identities constructed (Laclau, 2005b; Searle, 1969). Islamophobic hate speech is a

behaviour which *does things* in contemporary politics. It causes huge harm to targeted victims and communities, spreads divisiveness, and normalizes dangerous and extremist ideas, as with other forms of hate speech (Matsuda, Lawrence, Delgado, & Crenshaw, 1993). Islamophobic hate speech on social media might also be a precursor to individuals engaging in other harmful and extremist behaviours, including offline hate crimes (Awan & Zempi, 2017; Müller & Schwarz, 2017) and right wing terrorism (Gill et al., 2017) – although more evidence is needed to substantiate this fully.

In most previous research, Islamophobia has been almost exclusively associated with the far right, despite growing anecdotal evidence that it exists amongst supporters, voters, followers and representatives of mainstream parties such as the Conservatives and Labour. Islamophobia can cause harm wherever it manifests, and the potential for mainstream parties to engage in Islamophobia should not be ignored. In this thesis, I depart from previous research by studying parties from across the UK political spectrum: the British National Party (BNP), the United Kingdom Independence Party (UKIP), the Conservatives and Labour. Specifically, I study these parties' Twitter followers. These can be understood, ontologically, as 'digitally native' political actors, which are of growing importance in contemporary politics.

Islamophobic hate speech is also an understudied aspect of UK politics (All Party Parliamentary Group on British Muslims, 2018), and its presence on social media platforms is particularly concerning. More work is needed to define and monitor Islamophobic hate speech, and also to critically understand its nature, dynamics and drivers within contemporary politics. Academic research will contribute to ongoing policymaking and activist discussions, assist with work to counter the harmful effects of Islamophobia and provide support to victims. Accordingly, and reflecting the central role played by Twitter in contemporary politics, this thesis addresses a single aim:

To understand the nature and dynamics of Islamophobic hate speech amongst followers of UK political parties on Twitter

From this research aim, I identify five research questions (RQs) and an additional research goal (RG):

- RQ 1: What is the conceptual basis of Islamophobia?
- RQ 2: To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?
- RQ 3: To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?
- RQ 4: To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?
- RQ 5: Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?
- RG: To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia

By addressing these research questions and goal, I make three main contributions. First, I make a *conceptual* contribution to the study of Islamophobia. Islamophobia is a deeply contested concept in the social sciences and I offer an account which is robust and rigorous, defining Islamophobia in terms of negativity and generality. I use this argument to construct a framework for the second, *methodological*, contribution: creating a multi-class supervised machine learning classifier for Islamophobic hate speech. This distinguishes between weak and strong varieties and can be applied robustly and at scale.

My third contribution is *theoretical* and relates to the nature and dynamics of Islamophobia within UK politics. Drawing the findings of the research together, I pose a meta-argument: Islamophobic hate speech amongst followers of UK political parties on Twitter can be best conceptualised as analogous with a meteorological *wind system*. This analogy highlights the complexity of Islamophobia and how it operates heterogeneously, varying in terms of strength, users, time and political party followership. Islamist terrorist events precipitate Islamophobic *hurricanes*, which are large but temporary spikes in Islamophobic hate speech.

This meta-argument is based on my substantive findings about the nature of Islamophobia within UK politics. Through my empirical analysis, I provide insight into the nature of UKIP and the BNP, characterising UKIP as a *halfway house* between the mainstream and the far right. I show that there is considerable heterogeneity in the strength, magnitude and trajectory of Islamophobic hate speech expressed by followers of the BNP. I also argue that there is a *twin threat* of Islamophobia in UK politics, from the strong but relatively rare Islamophobia of the far right to the weak (but more widespread and insidious) Islamophobia of followers of mainstream parties. This points to the existence of an *Islamophobia gap*; mainstream parties officially reject Islamophobia and support policies and discourses for greater tolerance and respect, but their Twitter followers nonetheless engage in Islamophobic hate speech.

I extend Eatwell's theory of cumulative extremism (Eatwell, 2006). Currently, this theory is situated at just the level of inter-group dynamics, hypothesizing that different extremisms feed off and magnify each other. Through investigating the impact of Islamist terrorist attacks on hate speech, I argue that cumulative extremism operates at the level of individuals, including the followers of mainstream political parties. I also outline a four-step process of escalation and de-escalation following Islamist terror attacks.

Initially, there is a very short period in which the volume of Islamophobic hate speech rapidly increases, followed by a long two-phase de-escalation period. Then, the level of Islamophobia returns to a stable baseline level. The increase in Islamophobia during terrorist attacks is associated with an increase in the number of users who tweet Islamophobically for the first time (who I call *one-off Islamophobes*). However, the vast majority of the increase in Islamophobic tweets during such periods is driven by existing Islamophobes sending a higher number of Islamophobic tweets. I also question the theory of cumulative extremism by showing that the baseline level of Islamophobia does not increase following terror attacks. I use this to argue that extremisms do not accumulate but *reacts* to each other, usually in a way that is only short and temporary.

Each of the findings in this thesis substantively advances knowledge of Islamophobic hate speech amongst Twitter followers of UK political parties and, as I discuss in Chapter 8, are also relevant for understanding behaviour (i) in UK politics and (ii) on social media more broadly. Although I focus exclusively on Islamophobic hate speech, the findings are also relevant for understanding other forms of hate, such as racism, misogyny and xenophobia, given well-documented affinities between them. Further research is required to understand how different manifestations of hate overlap and coincide.

The scope of this work can be understood in light of Boellstorff's work apropos the relationship between digital and virtual research. He argues that digital objects of study should be addressed 'in [their] own terms' but that researchers should recognise 'the direct and indirect ways online sociality points at the physical world and vice versa' (Boellstorff, 2012, p. 40). Thus, the work here is specifically focused on studying the digital but recognises the inextricable connection between online and the offline behaviour. Nonetheless, I still recommend caution in generalising the results to other

contexts such as other social media platforms, other Internet sites, and the offline world, as socio-technical affordances, relations and practices differ.

Finally, in this thesis I use a mix of quantitative and qualitative methods within a complementary computational social science research design. My source of data is the digital traces left by users and which are now increasingly available to researchers through platforms' APIs. This enables me to unobtrusively investigate the actual behaviour of Twitter users, rather than their anticipated, preferred or remembered behaviours (Margetts, 2017a). The use of computational methods alongside rigorous theory and qualitative conceptual work contributes to the ongoing refinement of 'big data' applications in the social sciences by showing the research potential in integrating them fully within a single research design.

1.1 | Structure

This thesis consists of 8 chapters. Chapters 2 and 3 comprise the literature review and research design overview. In Chapters 4, 5, 6 and 7 I report on empirical findings. This comprises qualitative thematic work in Chapter 4, the creation of a supervised machine learning classifier in Chapter 5 and statistical analyses of users' behaviour in Chapters 6 and 7. Chapter 8 is a discussion chapter in which I synthesize the results and conclude the research.

In Chapter 2, I review existing literature relevant to the research aim, drawing together three overlapping fields of research – (i) Islamophobia, (ii) politics on social media, specifically Twitter, and (iii) UK party politics. Through this, I generate the research questions and additional research goal outlined above.

In Chapter 3, I outline and justify the research approach and design, which can be understood as a form of 'complementary' computational social science (Blok &

Pedersen, 2014). I explain the data collection process, discuss relevant ethical issues, and then provide a brief overview of the different methods used in each chapter. Note that methods are discussed in greatest detail within the corresponding empirical chapters.

In Chapter 4, I conduct an in-depth philosophical and thematic investigation of Islamophobia by examining a dataset of tweets from far right Twitter accounts. I use this analysis to critique five widely used conceptualisations of Islamophobia, identified from relevant academic literature. Building on this, I argue that Islamophobia can be conceptualised in terms of two dimensions: (i) negativity and (ii) generality. This answers the first research question: 'what is the conceptual basis of Islamophobia?' I then use the conceptual arguments to create a framework for distinguishing between weak and strong varieties of Islamophobic hate speech.

In Chapter 5, I build on the framework created in Chapter 4 to develop two machine learning classifiers. The first is a binary classifier which identifies whether tweets are Islamophobic or not. The second is a multi-class classifier which distinguishes whether tweets are non-Islamophobic, weak Islamophobic or strong Islamophobic. This realises the additional research goal: 'To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work into the concept of Islamophobia'. To create the classifiers, I annotate several thousand tweets (alongside two other expert annotators), extract relevant input features through extensive testing (primarily, a word embeddings model), select an optimal algorithm (Support Vector Machines (SVM)) and then evaluate. Both classifiers are suitable for use in the proceeding empirical chapters.

In Chapter 6, I use the multi-class machine learning classifier to study the behaviour of all followers of the BNP in order to address the second research question: 'To what extent does Islamophobic behaviour vary across followers of UK far right parties on Twitter?'

I show that there is considerable heterogeneity amongst followers of the BNP, and identify, using a ground-up statistical method (latent Markov modelling), the existence of six distinctive user trajectories. Based on this, I argue that it is inadequate to use a 'broad brush' to characterise the far right and highlight the need for more nuanced characterizations.

Chapter 7 is split into two parts. First, I study differences in the prevalence and strength of Islamophobic tweeting across followers of the four political parties (BNP, UKIP, the Conservatives and Labour). I show that followers of the BNP send the most Islamophobic tweets, followed by UKIP and then the Conservatives and Labour. Second, I examine the impact of Islamist terrorist attacks on the prevalence of Islamophobic hate speech. Here, I (i) identify a four-phase process of escalation and de-escalation of Islamophobic hate speech during Islamist terrorist attacks and (ii) show that the attacks impact followers of all four parties. From these results, I argue that Islamophobia constitutes a *twin threat* in UK politics. I extend the theory of cumulative extremism to include individuals who are not part of extremist groups and also challenge the notion of 'accumulation' by showing that the baseline level of Islamophobia does not increase following attacks. I also examine the types of users who tweet Islamophobically during terrorist attacks and find that the increase in volume is not driven by 'one-off' Islamophobes but, rather, by existing Islamophobes sending a higher volume of Islamophobic tweets.

In Chapter 8, I discuss and synthesize findings and conclude the research. I make the meta argument, already discussed, that Islamophobia amongst followers of UK political parties on Twitter should be conceptualised as a *wind system*, in which Islamist terror events are *hurricanes*. I then discuss the implications of my findings for the followers of UK party politics more widely, regarding (i) the ideological position of UKIP, which I argue constitutes a *halfway house* between mainstream parties and the far right, and (ii)

the existence of an *Islamophobia gap* between mainstream parties' official positions and the behaviour of their followers. I consider the role of social media more widely, contributing to ongoing discussions about its effects within society. I also consider the policy contributions of my work in terms of (i) defining Islamophobia, (ii) monitoring and predicting Islamophobia, (iii) providing support to victims, (iv) countering Islamophobia and (v) understanding radicalization pathways. Finally, I discuss the thesis' limitations and outline future steps to maximize the research's impact. Limitations are also discussed at the end of each chapter.

Chapter 2 | Literature review

In this literature review, I start from the research aim defined in the Introduction:

> *To understand the nature and dynamics of Islamophobia amongst followers of UK political parties on Twitter*

From this, I identify five research questions from the existing literature, as well as an additional research goal.

In the first section, I examine the nature and concept of Islamophobia. I highlight that at present there is no consensus as to how Islamophobia should be defined, which is a considerable obstacle to empirically investigating it. I then consider Islamophobia within UK politics, discussing its role within both far right and mainstream political parties. I argue that current approaches are too simplistic, failing to adequately consider (i) the internal heterogeneity of Islamophobia within the far right and (ii) manifestations of Islamophobia within mainstream parties. In the second section, I then outline and critically discuss five of the most applicable theoretical explanations of Islamophobic behaviour. I argue that the theory of cumulative extremism is the most relevant, and link it with empirical research on how terrorist attacks drive Islamophobia. I also highlight several potential extensions to the theory. I then discuss the methodological choices involved in studying Islamophobic hate speech on social media and justify the use of a machine learning classifier in this thesis. In the final section, I survey machine learning classifiers in previous research and identify several limitations; chiefly, they are insufficiently informed by relevant *conceptual* social scientific research into Islamophobia. As such, I argue that a theoretically informed and contextually specific Islamophobic hate speech classifier is needed for this project.

2.1 | Islamophobia

Islamophobia can be understood as a 'new name for an old concept' (Bleich, 2011, p. 1582). Whilst the term itself only entered public consciousness with the Runnymede Trust's 1997 report, 'Islamophobia: a threat for us all' (Runnymede Trust, 1997), it describes a phenomenon which has long existed especially in the West; systematic and casual derogation, subjugation, exploitation and exclusion of Muslims. Western research in this area largely originates with Said's landmark 1978 text *Orientalism*, in which he argued the West has long viewed Islam, and the so-called 'East' more generally, as subordinate and inferior. Through a close reading of Western cultural artefacts, primarily English literature, Said found evidence of 'subtle and persistent Eurocentric prejudice against Arabo-Islamic people and their culture.' (Said, 1978, p. 56). Others have since built on Said's work to argue that Islam is routinely viewed as a 'threatening other' in the West (Poole, 2002, p. 33), that the West has a long history of 'gain[ing] cultural and civilising power over Muslim populations' (Ingham-Barrow, 2018, p. 11) and that in many discourses Muslims are constructed to be 'alien and foreign to western society' (Lowe, 1985, p. 55). Thus, Dunn et al. argue that whilst the term 'Islamophobia' is relatively new, the tropes and behaviours associated with it are 'well-rehearsed' and familiar to people living in Western societies (Dunn, Klocker, & Salabay, 2007, p. 564).

Islamophobia has a hugely detrimental impact on victims and targeted communities, as well as wider society (Ingham-Barrow, 2018; Runnymede Trust, 2017; Tell Mama, 2015). There is evidence that the prevalence of criminal Islamophobic behaviour is increasing. In 2016/2017 the Home Office reported that there was a 35% increase in

religiously motivated hate crimes to 5,949 incidences[1] (HM Government, 2017b). Not all religious hate crimes are necessarily anti-Muslim in nature, and noticeably Jewish people are often targets of religiously motivated hate crime (CST, 2018). Nonetheless, given initial evidence that the recorded hate crimes were driven by prominent events, such as Islamist terror attacks (HM Government, 2017b), it is likely that a considerable amount of the recorded anti-religious hate crime is anti-Islamic in nature. Figures released by the mayor of London, Sadiq Khan, show also that there was a considerable increase in Islamophobic hate crime following the London Bridge Islamist terror attack in June 2017 (Dodd & Marsh, 2017).

The Crime Survey for England and Wales (CSEW) also provides evidence that the prevalence of Islamophobic crime is increasing in the UK. The CSEW is a face-to-face victimization survey conducted by the Office for National Statistics in which a representative sample of 35,000 people resident in England and Wales report on their experiences of crime over the 12 months leading up to the survey (CSEW, 2018). Although the reported values in the CSEW are estimates, the survey is useful because it is more comprehensive than other methods and the face-to-face methodology helps to minimize reporting biases. The most recent figures (for 2012/2013) estimate that Race and Religion are the most prevalent focuses of hate-motivated crimes. Out of 278,000 estimated hate crimes in 2012/2013, 154,000 related to race and 70,000 to religion, with

[1]Over the same period there was a 27% increase in racial/ethnic hate crimes to 62,685 incidences. Crimes are recorded under multiple identities and so it is likely that most Islamophobic attacks which took place were reported as religious hate crimes HMGovernment (2017). Hate Crime, England and Wales, 2016/2017. H. Office. London, Home Office.

an increase in both cases on the prior year (HM Government, 2017b). Other evidence suggests that, specifically, online Islamophobic hate crimes are increasing. Tell Mama reported a 47% increase in 2016 in the number of offline Islamophobic 'incidents' compared with the previous year (n = 642). The proportion of attacks which were directly abusive and violent had also increased (Tell Mama, 2017). That said, as Tell Mama collects only self-reported data and its collection practices change over time, such figures should be treated with some caution.

In many cases, Islamophobia overlaps with other prejudices (such as misogyny, xenophobia and racism), leading to intersectional experiences for victims (Burnap & Williams, 2016; Mccall, Crenshaw, & Cho, 2013). For instance, several studies show that Muslim women often suffer from heightened levels of prejudice and aggression because they are more likely to wear clothing associated with the Muslim faith, such as the niqab or burqa (Bilge, 2010; Mirza, 2013). Studying intersectional forms of Islamophobia is important for understanding how individuals experience Islamophobia in their daily lives, however it is not the primary concern of this work.

In a recent report, the All Party Parliamentary group on British Muslims' makes the argument that, 'Islamophobia is rooted in racism and is a type of racism that targets expressions of Muslimness or perceived Muslimness' (All Party Parliamentary Group on British Muslims, 2018). This definition, which was adopted by the Labour Party but rejected by the Conservative Party in 2019, raises a long-running tension in discussions of Islamophobia: is it a type of racism or is it rooted in (i.e. similar to but distinct from) racism? The issue here is that Islamophobia is, analytically, concerned only with religion but in practice it usually manifests against racially identified targets. And, as the APPG's definition also alludes, the racial identification of targets is a question of perception. In their report, they show that Sikhs, Hindus, and even Catholic Italians, have all been

victims of Islamophobia. In terms of how victims and communities are impacted, which is the main focus of the APPG's work, Islamophobia is therefore closely linked to the racial and cultural performances of Muslim identity: race and culture are closely intertwined with religion. There are also clear political benefits in linking Islamophobia (an often-ignored issue within policymaking) to racism (a near universally reviled prejudice). However, if we consider how Islamophobia is articulated then the race and religion can be more easily separated. Islamophobic hate speech may, and often does, involve a racial dimension – but unless it also involves targeting against Muslims *qua* religion then it cannot be easily discerned as Islamophobia. That is, Islamophobia must *necessarily* involve an attack against the religious identity. It may often also involve a racial or cultural component, and which individuals are the targets of Islamophobia is almost certainly racially biased, but for it to be Islamophobia it must involve a religious attack. This is a necessary (and, indeed, sufficient) component. Thus, without ignoring the hugely important role played by racism and cultural superiority in how Islamophobia manifests across society, it can still be viewed as a distinct analytical category of articulation. In this thesis, I focus specifically on articulations of Islamophobia *qua* religion rather than Islamophobia *qua* race.

Contemporary Islamophobia is complex and multifaceted, operating across many different modalities and settings, including workplace discrimination (Panayi, 2014), legally defined hate crimes (CPS, 2017), hate speech (Burnap et al., 2014; Sponholz, 2016), micro aggressions (Haque, Tubbs, Kahumoku-Fessler, & Brown, 2018; Nadal et al., 2012), 'everyday' practices (Dunn & Hopkins, 2016; Moosavi, 2015), physical assault (Tell Mama, 2017) and institutional patterns of prejudice (ECRI, 2016). The focus of the present work is specifically *Islamophobic hate speech on Twitter*. This is a domain of study which has just recently attracted sustained academic attention. Indeed, only back

in 2014, Awan had to explicitly make the point that, 'online Islamophobia must be given the same level of attention as street level Islamophobia' (Awan, 2014, p. 133).

The term *Islamophobic hate speech on Twitter* requires unpacking. The 'Twitter' part is the most straightforward. Twitter is a social media platform, and as such can be situated within broader definitions of these platforms, which tend to emphasize connectivity, collaboration and content production (Fuchs, 2017, pp. 38–39). For instance, Kietzmann et al. define social media as 'platforms via which individuals and communities share, co-create, discuss, and modify user-generated content' (Kietzmann, Hermkens, Mccarthy, & Silvestre, 2011, p. 241) and Kaplan and Haenlein, as 'a group of Internet-based applications that […] allow the creation and exchange of User Generated Content' (Kaplan & Haenlein, 2010, p. 61). Arguably, both these definitions over-emphasize the *active* nature of social media, failing to take into account the large number of 'lurkers' who consume content without producing it (Crawford, 2009). Accordingly, Twitter is better understood in line with broader accounts of social media which do not privilege content production over content consumption, such as van Dijck's; 'social media can be seen as online facilitators or enhancers of human networks – webs of people that promote connectedness as a social value' (van Dijck, 2013, p. 11).

The 'speech' part is open to discussion. In an online context, 'speech' can be understood as any content which is produced or shared, including text, videos, pictures, GIFs and emojis. This notion of speech is more in line with semiotics than linguistics. The specific type of speech studied in this thesis is text posts, specifically 'tweets' sent on Twitter. Text posts are one of the most common types of content shared online (Pew Research, 2016) and is a highly informative source of data for political studies (Grimmer & Stewart, 2013). On Twitter, it is also possible to share someone else's tweet through a 'retweet', which is a common activity. Retweets are also studied within this thesis.

I conceptualise sending a tweet as a kind of speech act – specifically, as a *behaviour* rather than simply as an *expression* or epiphenomenon of something which pre-exists, such as a belief or an opinion (Glynos, Howarth, Norval, & Speed, 2009). Focusing on behaviours is increasingly recognised as an important aspect in studies of prejudice, as Harris claims apropos South African inter-racial xenophobia, '"Xenophobia" as a term must be reframed to incorporate *practice*. It is not just an attitude: it is an activity.' (Harris, 2002). Harris' work highlights that behaviours are what harm victims and damage society and as such warrant the most attention – strictly considered, Islamophobic attitudes which are not articulated or expressed do not create harm. This is supported by arguments made in the linguistics tradition of pragmatics, in which the behavioural aspect of speech is highlighted. Pragmatics convincingly suggests that words do not only convey information but they *do* things by acting upon the world and changing the existing state of affairs (Austin, 1962; Searle, 1969). These insights, which were first developed to understand (offline) verbal communications, have increasingly been applied to understand behaviour on social media (Hodgkin, 2017).

Queer theorist Judith Butler has extended the idea of speech acts through her concept of 'performativity', which can be used to analyze the 'doing' dimension of speech and communication. She argues that words are not merely instrumental tools for performing (other) actions but that words are actions in-themselves (Butler 1988). This is in keeping with arguments put forward by other political theorists, such as Laclau and Mouffe who argue that speech is a discursive action which constructs, validates and regulates particular states of affairs (Laclau & Mouffe 1985, Laclau 2005, Butler et al. 2001). In the terms of speech act theory, Laclau and Mouffe's and Butler's key insight is that speech is inherently perlocutionary as well as locutionary; it not only describes the world but also has real consequences within it.

Butler's other work on the concept of 'excitable speech' also shows people's 'vulnerability' to language (Butler, 1997). She argues that people's experience of the world is mediated through language and that identity is interpellated only through linguistic and semiotic practices. This makes all people inseparable, in some sense, from the power of words. In making this argument, reflecting her prior work on performativity, Butler effectively collapses the distinction between representation of harm/hate (which for some is the sole function of language within prejudicial discourse) and the articulation of harm/hate. For her, language is itself capable of inflicting harm: 'Speech does not merely reflect a relation of social domination; speech enacts domination'. That is, language has an agential capacity to inflict harm on people, and need to be studied as a behaviour in its own right. This argument is also captured by Matsuda et al.'s point that it is possible for 'words to wound': they can be 'used as weapons to ambush, terrorize, wound, humiliate, and degrade' (Matsuda et al., 1993, p. 1).

These more theoretical arguments apropos the agential and 'doing' nature of speech are also supported by empirical research which show the huge impact that negative speech has on individuals' wellbeing (Tell Mama, 2017, 2018a, 2018b). Indeed, the harm caused by Islamophobic speech acts is why I use the word 'hate' in this thesis; Islamophobic speech acts should be described in terms which articulate their dangerous and unacceptable nature. This reflects previous work in the field of Internet studies. In an early study of virtual communities' experience of harmful behaviour online, Williams draws attention to forms of 'sociopathic' behaviour online, noting that in discussing linguistic injuries, people are often 'forced to draw [their] vocabulary from physical injury.' Nonetheless, the online harm caused by language is experienced as real and described in terms which reflect its serious impact. Whilst there is a tendency to ignore online crimes and harm as less serious than their offline counterparts (an issue which has

been somewhat addressed through the UK Government's 2019 discussion of an 'online harms' regulator (HM Government, 2019), they are still experienced as 'real' and can cause experiences of anxiety, impacting on individuals' wellbeing and engagement in the virtual space.

Finally, the power of language to inflict harm is empirically reflected in the broader historical context. Instances of genocidal violence have been committed alongside, and catalysed by, hateful language, such as the use of radio during the Rwandan genocide. Benesch develops the concept of 'dangerous speech' to describe speech which 'has a reasonable chance of catalysing or amplifying violence by one group against another' (Benesch, 2012, p. 2; Maynard & Benesch, 2016). 'Dangerous speech' reflects the role of discourse in both inflicting harm in its own right and also enabling other forms of violence to occur. This historical legacy, in turn, serves to further increase the harm caused by words alone as the fear of further action always exists.

The final part of the term *Islamophobic hate speech on Twitter* which needs unpacking is also the most difficult: 'Islamophobic'. Despite the plethora of research, considerable terminological confusion remains amongst academics and policy makers as to what 'Islamophobia' actually denotes (Sayyid, 2014). It can be best understood as what Gallie describes as an 'essentially contested concept' – numerous definitions and accounts proliferate in the literature, with little consensus on what the core features are (Gallie, 1956). Nonetheless, there is a pressing need to better define and stipulate Islamophobia. The House of Commons All-Party Parliamentary Group on British Muslims wrote in their 2018 report that only if we 'begin from the point of an agreed definition' can the negative effects of Islamophobia be 'reversed' (All Party Parliamentary Group on British Muslims, 2018), a position which is supported by the Muslim charity Muslim Engagement & Development (MEND) (Ingham-Barrow, 2018).

The existing legal frameworks around Islamophobia are largely viewed as inadequate even though, broadly put, Islamophobia is considered illegal under existing UK Law. There is growing consensus that the law provides only limited protection to victims of Islamophobia and is particularly poorly formulated with regard to online hate. Liberty, the UK-based civil liberties and human rights campaign group, argues that 'there needs to be a wholesale review of speech offences, particularly under the Public Order Act.' (Liberty, 2018) The Home Affairs select committee reached a similar conclusion in its wide-ranging 2017 review, recommending that, 'the Government should review the entire legislative framework' around online hate speech, harassment and extremism (HM Government, 2017a). As part of the evidence gathering process for this review, the Law Commission of England and Wales opined that the law lacked 'legal certainty' and would benefit from reform (Ibid.). In a wide-ranging review of legal responses to hate speech, Williams also draws attention to the role of the United Nations in combating online hate; the 2013 General Recommendation on Combating Racist Hate Speech explicitly includes hate speech sent electronically, including social media posts. Similarly, the European Union requires that member states criminalise speech that incites racist or xenophobic hatred (Williams, 2019). This reflects the strong International consensus against hate, although legal protections remain uneven globally.

The Crime and Disorder Act 1998 introduced racially and religiously aggravated offences – or 'hate crimes' – into UK law. These are offences (i) which are illegal in themselves and (ii) where the victim is targeted due to their identity. Actions specifically covered by the 1998 act include wounding, assault, criminal damage, harassment, stalking and threatening/abusive behaviour. To prosecute hate crimes both the 'basic' criminal offence must be proven, as well as the hateful racial/religious element. The 1998 Act also requires that when any crime is sentenced, any racial/religious element must be taken into account,

even if it is not charged specifically as a hate crime. Importantly, the definition of hate crime used in the Act, and elsewhere by the Police, is victim-oriented, describing hate crimes as 'any incident/crime which is *perceived by the victim, or any other person*, to be motivated by hostility or prejudice based on a person's race or religion or perceived race or religion.' (CPS, 2017; emphasis added).

Hate speech is regulated primarily through the Public Order Act 1986, which bans 'stirring up hatred' on the grounds of race or religion. 'Stirring up hatred' is defined as 'using threatening words or behaviours or displaying any threatening written material' (CPS, 2017, p. 3). It covers both verbal and non-verbal communications, including written leaflets, materials posted online, posters, broadcast media such as videos, and plays and comedic acts. Since 1994 it also covers any actions which are undertaken with an 'intent to cause a person harassment, alarm or distress'. Muslims have only been included within the protections of the Public Order Act 1986 since 2006 when the Racial and Religious Hatred Act 2006 was passed. Thus, Allen reports that in the early 2000s 'it was perfectly within the law to discriminate against someone on the basis of their being Muslim: a loophole that was exploited by far-right political groups following the attacks of 9/11.' (Allen, 2010) Williams also notes that the requirements to meet the bar of legally defined hate speech are high in order to protect freedom of expression, and go far beyond simply voicing an opinion or causing offence (Williams, 2019).

Online hate speech is particularly difficult to regulate, and the current legal response has received criticism for being confused and outdated. A patchwork of laws, including the Public Order Act 1986, the Communications Act 2003 and the Terrorism Act 2006, have been brought to bear on the issue. During the Home Affairs select committee 2017 hearing on online hate crime it was noted that 'we have not had a proper law passed since social media became in widespread use.' (HM Government, 2017a). The

Communications Act 2003 regulates communications which take place online and specifically bans 'malicious communications', which includes much hate speech. The Terrorism Act 2006 criminalises the 'encouragement of terrorism', which can include extreme forms of hate speech, such as when individuals call for Muslims to be attacked or systematically expunged. Overall, there is a need for a more joined up approach to the law around Islamophobic behaviour, especially online hate speech. Furthermore, due to its ambiguity and lack of clarity, the law is a poor basis for defining Islamophobia and alternative definitions must be identified.

The lack of conceptual consensus apropos Islamophobia within the academic literature is the product of at least three sources of ambiguity. First, is that Islamophobia has been studied in many disciplines – of which cultural studies and social psychology are the most prominent – which have competing methodological approaches and theoretical concerns. For instance, a recent trend in cultural and social psychology studies is focusing on less visible and more subtle types of Islamophobia (Haque et al., 2018; Nadal et al., 2012). However, this focus is not widespread in political science, leading to diverging works across these fields. In addition, Islamophobia 'never stands still' and its 'shape, size, contours, purpose, function' varies across different historical, geographical and political contexts (Taras, 2013, p. 422). For example, countries such as India have very different social dynamics and historical legacies, which means that accounts of Islamophobia developed in that context may be less relevant to studying Islamophobia in the UK (Anand, 2010).

Second, is that Islamophobia is not only an analytical concept but also a deeply divisive political concept with a contentious social history in the UK (Chris Allen, 2010b, 2017). Thus, many researchers approach Islamophobia not only as an empirical phenomenon to be studied, but also have a clear activist agenda. In many cases, researchers are deeply

invested in studying and countering Islamophobia, and bring their own political views to bear on the issue. As Bleich notes, 'what unites […] definitions, proto-definitions and underlying assumptions is a sense that Islamophobia is a social evil.' (Bleich, 2011, p. 1583). The normative nature of much research can lead to huge variations in what settings and manifestations of Islamophobia are studied.

Part of the challenge in grappling with Islamophobia in the UK is that the government has itself been accused of Islamophobia, and has come under considerable criticism for implementing policies which either disproportionately disadvantage Muslim groups or which create a culture of fear against Muslims, thereby 'fostering and furthering Islamophobia.' (Runnymede, 1997) In particular, the UK government's counter-terrorism strategy PREVENT has received criticism for being prejudiced and overly intrusive. For instance, the Joseph Rowntree Charitable Trust's racial equality group JUST claims that PREVENT is 'built on a foundation of Islamophobia and racism' (JUST, 2018) and the EU's European Commission against Racism and Intolerance (ECRI) similarly suggests that PREVENT 'may fuel discrimination against Muslims' (ECRI, 2016). Even the recent efforts to tackle hate crimes have been criticised for lacking funding, being uncoordinated and for being fundamentally reactive rather than proactive (Law Commission, 2014; Amnesty, 2017; The Guardian, 2017). These issues have made it difficult for Muslim voices to be heard and for advocates against Islamophobia to be listened to within policymaking circles.

Third, is that in many studies the term 'Islamophobia' is not defined at all, even when it is the main object of research. Such studies rely on the audiences' pre-existing, and as such potentially hugely divergent, understandings of what Islamophobia is. This can make interpreting the results of studies difficult and make it harder to identify commonalities across them and synthesize findings. This issue is particularly acute when

studying behaviours which are less explicitly Islamophobic as not all audiences and researchers may agree that they are prejudicial. This can only be resolved if Islamophobia is defined clearly at the start of research projects.

In response to the terminological confusion, several academics have called for more conceptual work to be undertaken; Bleich argues that Islamophobia should be refined so that it can become 'a more concrete and useable concept for social scientists' (Bleich, 2011, p. 1582) whilst Mondon and Winter argue researchers need to go beyond 'contextually specific' accounts (Mondon & Winter, 2017, p. 2152). There have been three main responses to the need for greater conceptual clarity. First, empirical works which either (i) describe the features of Islamophobia (Aguilera-Carnerero & Azeez, 2016; Amiri, Hashemi, & Rezaei, 2015; B. Lee, 2017; Mondon & Winter, 2017) or (ii) identify different types of Islamophobic individuals (Awan, 2014, 2016; Jacks & Adler, 2015). Such works provide more detail around how Islamophobia *manifests* but do not provide greater *conceptual* insight. Indeed, they may even cloud the debate by demonstrating an ever-increasing variety and complexity of manifestations of Islamophobia.

Second, various alternative – and generally more specific – terms have been put forward, such as 'anti-Muslimism' (Faliq, 2010; Halliday, 1999), 'anti-Muslim prejudice' (Malik, 2009), 'anti-Muslim hate' (Mondon & Winter, 2017), 'Islamoprejudice' (Imhoff & Recker, 2012), 'Muslimophobia' (Erdenir, 2010), 'Islamonausea' (Aguilera-Carnerero & Azeez, 2016) and 'miso-Islamia' (Hussain, 2012). At the same time, many have argued that the term 'Islamophobia' should be abandoned. Banton argues that Islamophobia is a 'folk concept' which is 'inadequate for sociological analysis because [it is] so marked by its political connotations and contexts' (Banton, 2015). Similarly, Bowen contends that because Islamophobia is 'used in an overly broad way and is highly polysemic […] using

it as an analytical term is a bit dicey.' (Bowen, 2005, p. 524). Springs provides a more politically oriented critique of the term, arguing that use of the '"phobia" moniker' can be counter-emancipatory as it 'isolates and opposes' Muslims in society (Springs, 2015). However, notwithstanding these critiques, as Bleich notes, 'Islamophobia has taken root in public, political, and academic discourse, and there is no putting the genie back in the bottle' (Bleich, 2011, p. 1584). It is a useful term precisely because it is so well-used and widely known; and deciding on a new 'label' does not resolve the problem of what that label should denote.

Third, researchers have made analogies with either anti-Semitism (Bunzl, 2005; Klug, 2014; Meer, 2013; Topolski, 2018) or racism – Islamophobia has variously been described as 'anti-Muslim racism' (Bayrakli & Hafez, 2016; Rana, 2007), 'cultural racism' (Saeed, 2007), 'multicultural racism' (Panayi, 2014) and 'differential racism' (Meer & Modood, 2009). There are merits to such approaches in that they render apparent the social significance and political implications of Islamophobia. This is important given the widespread view in the mid-20th century that supposedly 'cultural' forms of prejudice (such as Islamophobia) are less dangerous and harmful than biological forms, such as racism (Barker, 1981; Levi-Strauss, 1952). This position has been challenged by many subsequent thinkers, particularly in the work of Etienne Balibar (Balibar, 1991). It is also effectively countered when Islamophobia is directly linked with racism, which is viewed by most people as being far more socially unacceptable (Runnymede Trust, 2017). However, equating Islamophobia with any other form of prejudice (such as racism) is problematic and, crucially, does not obviate the need for conceptual work. It simply shifts the terminological co-ordinates of the issue; instead of having to define Islamophobia we are left with the similar challenge of having to define racism or anti-Semitism. It also potentially harms public discourse and the work of counter-Islamophobia activists by

conflating distinct notions; or, as Taras puts it, 'to categorize Islamophobes as racists is bad politics.' (Taras, 2013, p. 417).

Finally, it is worth considering how social media platforms have addressed the issue of Islamophobia. Twitter provides its own set of Rules which includes policies for 'Abuse and hateful content', which stops users from promoting 'violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' (Twitter, 2019). Islamophobia would fall under this definition as a form of religious affiliation. However, it is quite narrow as the use of the term 'direct' limits its scope to only overt forms of hate. It also focuses specifically on threats, attacks and the promotion of violence: these are also relatively overt forms of behaviour. This definition is useful as a way of signalling the platform's opposition to hatred (including Islamophobia) but is unlikely to be the basis of a conceptual definition. Facebook has moved to a three-tiered framework for dealing with hate speech (Facebook, 2019). Similar to Twitter they define hate speech as 'a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.' However, they distinguish between Tier 1 (primarily, threats of violence and dehumanization), Tier 2 (Statements of disgust, inferiority and contempt) and Tier 3 (calls for segregation and exclusion). This nuanced framework is useful for ensuring that appropriate content moderation processes are set in motion once content has been flagged, and relevant authorities are alerted where necessary.

Other platforms, such as Snapchat, Whatsapp and Instagram, have similar community guidelines to Facebook and Twitter in place. In all cases, these are useful at a high level for determining 'hate' but lack sufficient detail to be applied in a social science context.

Furthermore, given that platforms are concerned by (1) ease of implementation for the content moderators and (2) impinging on freedom of expression, these definitions and guidelines serve a somewhat different purpose to the research goals outlined here. As such, I do not explicitly draw on any platform community guidelines to address the conceptual question of what constitutes Islamophobia.

There is a pressing need for research which systematically investigates the conceptual basis of Islamophobia – rather than describing how it manifests, coining new terms or making analogies with other better-established forms of prejudice (such as racism). Providing a definition of Islamophobia is useful not only for the aims of this thesis but also for the study of Islamophobia more generally. Conceptual work undertaken here could be applied to other types of Islamophobic behaviour in other settings. In particular, the field of Islamophobic hate speech studies would benefit from more conceptual work. This is a rapidly expanding area of research which has so far been dominated by computer scientists rather than social scientists (as much research has focused on classification tasks), and as such lacks strong conceptual foundations. Accordingly, I respond to this gap in existing knowledge and the first research question addressed in this project is:

RQ 1: What is the conceptual basis of Islamophobia?

2.2 | Islamophobia within UK political parties

Islamophobia manifests across the UK party political spectrum, from parties which articulate and support explicitly Islamophobic policies (as with many far right organisations) to those which are more subtly imbricated in Islamophobic practices, such as (i) producing discourses which subtly demonise Muslims and normalise anti-Muslim fears, (ii) creating physical and communicative spaces which are hostile or unwelcoming to Muslims, and (iii) *implicitly* legitimizing Islamophobic attitudes and behaviours, thereby attracting support from people who engage in Islamophobic behaviour and hold Islamophobic attitudes (John, Margetts, Rowland, & Weir, 2004; Pogorelis, Maddens, Swenden, & Fabre, 2006; Richardson & Wodak, 2008; Wodak, 2016). All manifestations of Islamophobia harm Muslims and society, and as such should be challenged, whether they occur in niche far right settings or seemingly 'liberal' left-wing spaces – and, crucially, whether they occur offline or online.

Most research into political parties on social media has focused on their party representatives and candidates (Filimonov, Russmann, & Svensson, 2016; Gibson & Ward, 2009; Graham, Jackson, & Broersma, 2014; Jungherr, 2016; Wright et al., 2017). Remarkably little attention has been paid to their followers. This is surprising given that contemporary theoretical work on offline politics increasingly acknowledges the political party is a 'malleable concept' (Gauja, 2015, p.243) which includes not only leaders, representatives, members and voters but also affiliated supporters and party helpers (Fisher, Fieldhouse, & Cutts, 2014; Mair, Haid, Press, Borz, & Janda, 2008). Accordingly, and in line with much existing empirical research on online political behaviour, I contend that social media followers of political parties should be viewed as constitutive of the party. I bring the logic of this argument to bear on Twitter.

Following a party on social media can be best be understood as a micro-act which entails no financial costs and is logistically nearly costless (Margetts et al., 2015). However, just because following a party on social media is easy and 'cheap', this does not mean that it is insignificant. Indeed, interventions which situate social media within the 'attention economy', suggest that following an account is a meaningful act (Guo & Saxton, 2018; Tufekci, 2013; Zhang, Wells, Wang, & Rohe, 2017). In the attention economy, human attention is conceptualised as a scarce and valuable commodity – and in a digitized world saturated with lots of freely available high quality content, there is considerable competition for attention (Ciampaglia, Flammini, & Menczer, 2015). Following a party on social media means that their content will appear in a user's timeline or 'feed'. This is effectively 'spending' some attention. Thus, whilst following a party may be logistically and financially cheap it entails *attention* costs.

There are no direct analogies in the offline world for following a political party on social media. Some social media party followers will be involved with the party in the offline world as representatives, members and voters, whilst others are just 'digital foot soldiers', only expressing support for the party online (Vaccari & Valeriani, 2016, p. 306) – and yet others may be just interested observers, such as news reporters, or even opponents. Thus, given the variety of affiliations and relationships which are contained within the act of social media following, it can best be understood as an expression of affiliation and 'interest' rather than explicit 'support'. Nonetheless, just as not taking into account parties' loosely affiliated supporters and helpers creates an 'incomplete picture' (Fisher et al., 2014, p. 76), so too does not considering their social media followers. Accordingly, the main focus of this work is on the social media followers of political parties. This argument builds on Margett's prior conceptual work into the 'cyberparty' as a distinctive type of contemporary political party (Margetts, 2006).

In the remainder of this section I first discuss Islamophobia within the far right. I discuss the role of both the Internet and Islamophobia in the far right. I then explore the role of Islamophobia within mainstream parties, focusing on the Conservatives and Labour. I finish by considering Islamophobia within UKIP, the role of the party within contemporary politics and its relationship with other political parties. I outline two research questions in this section.

2.2.1 | Islamophobia in the far right

Many definitions of the far right exist in previous literature, and there is considerable 'terminological confusion' as to what the term itself means (Mudde, 2007a, p. 42). Carter draws attention to the need to distinguish between features which merely describe and those which *define* the far right (Carter, 2018). For her, authoritarianism, anti-democracy and exclusionary nationalism are definitional features whilst xenophobia, racism and populism are only 'accompanying characteristics' (Carter, 2018, p. 174). Similar distinctions are made elsewhere; Mudde claims the defining feature of the populist right 'is natural inequality or hierarchy, not nationalism' (Mudde, 2009, p. 331) and Rydgren argues that far right parties are 'embedded in a general socio-cultural authoritarianism' (Rydgren, 2010, p. 2). Broadly put, the defining features of the far right are: (i) excluding outgroups, often through violent and incendiary actions, rhetoric and policies, (ii) a Manichean view of society, in which both people and political actors are viewed explicitly as either 'good' or 'bad'), and (iii) authoritarianism, often expressed through fervent support for the rule of law (Biggs & Knauss, 2012; Goodwin, 2006; Goodwin, Ford, & Cutts, 2013; Ignazi, 2003; Macklin, 2013; Veugelers & Magnan, 2005).

2.2.1.1 | The far right and the Internet

Research indicates that the Internet has had a deep and lasting impact on far right parties, activists and organisations. For instance, in a review of the far right in Europe, Bayrakli and Hafez argue that, 'Islamophobic groups are especially active on the Internet. Often, the Internet is where right-wing groups emerge before materializing in "real life."' (Bayrakli & Hafez, 2018, p. 19) *How* the Internet is used has also changed considerably over the past decade. Initially, far right parties created static websites which served as 'broadcaster' one-way communication tools and information repositories – so-called 'Web 1.0' technologies (Atton, 2006; Margetts, 2006; Römmele, 2003). Since then, Internet use has shifted to interactive forms of communication ('Web 2.0' technologies). Far right supporters use dedicated forums, such as Stormfront[2], to build communities, share ideas and discuss politics with less fear of criticism (Bowman-Grieve, 2009; Caiani & Wagemann, 2009; Caren, Jowers, & Gaby, 2012; Froio, 2018; Meddaugh & Kay, 2009). More recently, far right actors have used multi-use social media platforms, such as Twitter, to (i) deepen the commitment of affiliates by creating ideologically closed reinforcing echo chambers (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015; Puschmann, Bastos, & Schmidt, 2017), (ii) publicise and share content, potentially amplifying the impact of their messages (Copsey, Dack, Littler, & Feldman, 2013; Lee, 2017; Williams & Burnap, 2016) and (iii) bypass traditional media 'gatekeepers' to recruit and engage with potential supporters (Alvares & Dahlgren, 2016; Caiani & Wagemann, 2009; Engesser, Ernst, Esser, & Büchel, 2017).

[2] The website Stormfront can be accessed at `https://www.stormfront.org/forum/` Last accessed on 4th January 2019.

Much research indicates that the Internet is not only a tool for far right actors but has also fundamentally altered their organization and activities. Caiani and Kroll argue that far right transnationalization is a product of 'the new virtual means of communication offered by the Internet […]' (Caiani & Kröll, 2015, p. 332) and that use of the Internet is associated with greater far right engagement in cross-country campaigns. A further effect of Internet usage is to create new organisational structures which combine elements of parties and social movements (Caiani, della Porta, & Wagemann, 2012; Caiani & Kröll, 2015). Gattinara and Pirro argue that, 'far-right networks often form and operate online […] facilitating the progressive integration of radical parties, extremist movements, and subcultural groups.' (Gattinara & Pirro, 2018, p. 3). This is supported by work by Froio and Ganesh on the interlinkages between far right parties, organisations and street movements on social media (Froio & Ganesh, 2018), as well as earlier social network analysis by Caiani and Wagemann on the connections between various far right websites (Caiani & Wagemann, 2009). Overall, previous research indicates that not only are new far right parties digital but even traditional far right parties have adopted digital practices (Atton, 2006).

In more recent times, following the presidential campaign and election of Donald Trump, the 'alt right' has emerged as a neologism for a wide range of loosely affiliated nationalist, sexist, racist and jingoistic online groups and activists, based primarily in the USA (Hope Not Hate, 2017). Alt right activists typically create and share interactive multimedia content on social media platforms (both mainstream ones, such as Twitter, and dedicated 'free speech' platforms such as Gab.ai) to spread hate speech, express support for Donald Trump's policies and troll opponents. Many users exploit the anonymity and, historically, looser regulations of the Internet to engage in behaviour which is either illegal or socially unacceptable in offline society (Futrell & Simi, 2017;

Hine et al., 2016). The rise of the alt-right points to the transformative impact of the Internet, and social media in particular, within the far right.

2.2.1.2 | The far right and Islamophobia

The term 'far right', both online and offline, refers to a complex and changing assemblage of parties and actors, and as such has been associated with many different types of prejudice, most noticeably racism, anti-Semitism, anti-Immigrant prejudice and Islamophobia (Ignazi, 2003; Mudde, 2002; Mudde & Kaltwasser, 2007). These can appear contradictory; anti-Semitism has long existed in the BNP whilst other far right parties in the UK, such as the English Defence League (EDL), support the (Jewish) state of Israel as part of their anti-Muslimism (Goodwin, 2013b; Zúquete, 2008). In the European context, Islamophobia has become a key feature of the far right (Bayrakli & Hafez, 2016, 2017, 2018). As Zúquete puts it, 'the threat that the Crescent will rise over the continent and the spectre of a Muslim Europe have become basic ideological features and themes of the European extreme right' (Zúquete, 2008, p. 322). This is evinced in the UK context by in-depth analysis of prominent far right organisations which are explicitly Islamophobic, including the BNP (Allen, 2010a; Goodwin, 2010; Kundnani, 2007) and the EDL (Allen, 2011; Goodwin, 2013b; Jackson & Feldman, 2011; Kassimeris & Jackson, 2015).

Numerous reports by the anti-Islamophobia monitoring charity *Tell MAMA* draw a link between the far right and Islamophobia, with one report finding that 69% of reported Islamophobic hate speech on social media had a link to the far right (Copsey et al., 2013; Feldman, 2015; Jackson & Feldman, 2011). Similarly, Awan argues that the EDL 'us[es] social networking sites like Twitter to post malicious statements […] promoting online hate' and elsewhere, with Zempi, that they 'exploit the virtual environment and [use] world- wide events to incite hatred towards Islam and Muslims' (Awan & Zempi, 2017,

p. 365). Demos also found that the EDL uses its following on Facebook and Twitter to raise agitation and promote antagonism against Muslims (Bartlett & Littler, 2011) and more recently, Froio shows that online far right groups create deeply affective anti-Muslim discursive frames (Froio, 2018). Overall, the evidence strongly demonstrates that the online far right is a locus of Islamophobic content – a view perhaps best captured by Awan's claim that far right groups have used social media 'to inflame religious and racial tensions' by creating 'walls of hate' (Awan, 2016, p. 17).

The far right is often depicted with a broad brush; party members, supporters and followers are viewed uniformly as extremists and Islamophobes. This is typified by the political and social 'cordon sanitaire' applied to far right parties, whereby mainstream political parties refuse to enter electoral pacts or coalitions with them (Akkerman & Rooduijn, 2015). However, the broad brush applied to the far right is surprising given offline research into the various causes of support for the far right, which include a desire to 'protest' or 'punish' mainstream parties (Akkerman & Rooduijn, 2015; Eatwell & Goodwin, 2010; Halikiopoulou & Vasilopoulou, 2014; van Der Brug, Fennema, & Tillie, 2000), economic disenfranchisement (Allen, 2017; Ellinas, 2013) and populist anti-elitism (Mudde, 2017; Vieten & Poynting, 2016). In some cases, far right support is driven not by out-group derogation (such as Islamophobia) but by its twin; in-group positivity (Brubaker, 2010; Green, Sarrasin, Fasel, & Staerkle, 2011). The wide range of motives for supporting the far right suggests that not all supporters are equally Islamophobic. Furthermore, the concerns about right wing radicalization discussed above demonstrate ideological *progression* within the far right, which implies that not all supporters are the same.

Part of the problem is that there is often a gap between the official public 'party line' and the views and behaviours of far right supporters in private settings. This has been

described in terms of the distinction between the 'esoteric' and 'exoteric' appeal of far right parties, where esoteric refers to what is discussed by 'converts or in closed circles' and exoteric to 'what is considered wise to say in public' (Eatwell, 1996). A similar distinction is between the 'backstage' and 'frontstage' of parties, where the frontstage refers to the public activities of parties (expressed through speeches, manifestos, leaflets and public interviews) and the backstage to their private activities (such as internal party newspapers, party conferences and closed online forums) (Fleck & Müller, 1998). The frontstage/backstage distinction is not only a conceptual problem but also methodological, pointing to the limitations of only studying the official far right ideology through party manifestos, interviewing party members and attending events. This issue exists for all parties, but is particularly problematic when studying far right parties as they tend to be more closed and more secretive – indeed, the problem is so severe that Mudde has queried, 'can the backstage be researched at all?' (Mudde, 2002). Research in social anthropology suggests that the idea of unmediated private behaviour is a myth (Miller & Horst, 2012). Nonetheless, the problem that behaviour which is studied in offline research is often highly mediated and censured can be somewhat addressed by studying the actual behaviour of parties' social media followers, recorded unobtrusively through the collection of digital traces, as is the case here.

Using a broad brush to characterize the far right is particularly ill-suited to studying social media followers as recent research into harmful and abusive behaviour online suggests it is highly likely that users are very heterogeneous. In a longitudinal study of users' behaviour in the comment section of a news website, Cheng et al. find that trolls are 'made not born' (Cheng, Bernstein, Danescu-niculescu-mizil, & Leskovec, 2017). They show that many users of a large online news platform have a propensity to engage in trolling behaviour in the right circumstances – but, fortunately, such circumstances occur

infrequently and so most users' trolling-like behaviour is only rarely activated. At the same time, there is a small committed cadre of more persistently troll-like users. The existence of both types of users means that any broad brush approach which characterises all users as the same will summarise the data poorly and is highly reductive. Furthermore, a large body of online research indicates that the distribution of most behaviours on social media on a per person is fat-tailed: a small number of individuals are responsible for the vast majority of any given behaviour (Bakshy, Hofman, Mason, & Watts, 2011). This suggests that Islamophobic behaviour within the far right is unequally distributed, with users exhibiting different patterns and trajectories of behaviour. Again, summarising user behaviour with a single value (such as an average measure of prejudice or aggression) will perform poorly at capturing the data.

There is a pressing need to go beyond the public oriented 'front stage' of the far right and to better understand the 'back stage'; to study the everyday actions and behaviours of party supporters. It is crucial to avoid assuming that they are all explicitly Islamophobic or, alternatively, to naively accept parties' self-descriptions – which can border on the ludicrous, as with the EDL's longstanding claim that it is a 'human rights organisation' (The EDL, 2018). In particular, given the rising use of the Internet by the far right, and its transformative effect on their organisational structure and activities, there is a need for more research into online far right actors, such as social media supporters of far right parties. There is also a need to better different pathways followed by far right followers on Twitter, including processes of radicalization and escalating extremism. Accordingly, this leads to the next research question:

> RQ 2: To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?

2.2.2 | Mainstream politics and Islamophobia

Islamophobia has long been associated with the far right, but research is also needed which addresses Islamophobia within mainstream UK political parties. Most mainstream parties and politicians have explicitly voiced their opposition to Islamophobia and, indeed, to any form of prejudice (Pogorelis, Maddens et al. 2005). Nonetheless, there is evidence that it can be observed – often in subtler but no less pernicious forms – in their discourses, practices and institutions. Williamson and Khiabany describe how the 'obsession' with the veil as a 'symbol of oppression' in UK politics, particularly within the Labour party, is often Islamophobic in nature (Williamson & Khiabany, 2010). Taking a longer timeframe, Allen also discusses how after the election of Margaret Thatcher, 'a shifting focus was identified in political discourse' as Islamophobia and anti-Immigration sentiment became normalized in politics (Allen, 2010b, p. 10). Along similar lines, the European Commission on Racism and Intolerance has drawn attention to the 'considerable intolerant political discourse' in the UK in the 2010s and raised concerns about 'the exploitation of [Islamophobia] in politics.' (ECRI, 2016, pp. 16–17). Sinno and Tatari also show how Islamophobia in UK politics goes beyond discursive actions to include institutional prejudice; Muslims, like other minority ethnic groups, are systematically under-represented at all levels of UK politics and suffer from considerable biases about their competency when being assessed for positions (Sinno & Tatari, 2009). Furthermore, initial evidence suggests that prejudicial behaviours have become far more widespread following the vote for Brexit in 2016 (The Guardian, 2018).

Criticisms about the treatment and representation of Muslims in UK politics has also been articulated by Muslim communities and prominent politicians. In 2018 Baroness Warsi criticised the Conservatives, her own party, for enabling Islamophobia, and joined the Muslim Council of Britain and the Conservative Muslim Forum in calling for an inquiry

into institutionalised Islamophobia in the party (The Independent, 2018). This reflects a longstanding view within Muslim communities that they are marginalised and derogated in UK politics. In 1990's the *Muslim Manifesto* Kalim Siddiqui, Director of the Muslim Institute and founder of the Muslim Parliament of Great Britain, argued that 'the Government, all political parties and the mass media in Britain are now engaged in a relentless campaign to reduce Muslim citizens of this country to the status of a disparaged and oppressed minority' (Siddiqui, 1992, p. 5).

The UK political system is a multi-party two-house representative system in which the executive part of government is selected from the legislature. In the lower house, the House of Commons, representatives are elected in the winner-takes-all 'first past the post' system, which is widely seen as a considerable barrier for niche parties seeking representation (Bischof, 2017). As such, the UK's political system is dominated by just two parties: Labour and the Conservatives. Since World War II they have received the most votes and seats in every election. Such is their dominance that the UK has long been described as a 'two-party' political system (Baldini, Bressanelli, & Massetti, 2018; Jones, Norton, & Daddow, 2018), despite challenges from the Liberal Democrats, Scottish National Party (SNP) and UKIP (Ford & Goodwin, 2014a). There is also anecdotal evidence of Islamophobia within both parties, particularly in the Conservatives (BBC, 2018a; The Independent, 2018). Given these factors, Labour and the Conservatives are the most important and relevant UK political parties to study in order to understand Islamophobia within mainstream UK party politics.

There is a clear tension within parties such as Labour and the Conservatives – between, on the one hand, voicing support for policies which challenge Islamophobia and support tolerance (often in a bid to attract votes from Muslims and ethnic minorities (Dancygier, 2017)) and, on the other hand, implicitly tolerating and even enabling Islamophobia. And

the real problem is that Islamophobia is hard to access; anecdotal evidence suggests that it is most prevalent outside of the scrutinised and media-curated 'frontstage' (The Independent, 2018). This would be in line with other forms of prejudice in UK party politics. There is a well-established gap between (i) parties' policies and stated outlook towards immigrants and (ii) the attitudes of their voters and members (broadly, parties are far more favourable towards immigrants than their supporters, and citizens in general) (Ford, Jennings, & Somerville, 2017; Hainmueller & Hopkins, 2014; Statham & Geddes, 2017). Potentially, a similar phenomenon can be seen apropos Islamophobia whereby there could be an *Islamophobia gap* between parties' official policies and discourse and how their supporters act and think.

The notion that Islamophobia is not only a problem of the far right but also exists within mainstream politics has far reaching consequences for how far right parties and their relationship with the mainstream is understood (Bale, 2018; Canovan, 2002; Laclau, 2005a). Mudde argues that whilst many view far right parties as a 'normal pathology' of Western liberal democracy (i.e. an irrational aberration which only gains support during periods of crisis and instability), they are better understood as a 'pathological normalcy'. For Mudde, pathological normalcy refers to the fact that far right ideology – which primarily concerns (i) the teleological importance of the nation, (ii) opposition to so-called 'non-native' elements, and (iii) distrust of institutions and the elite – differs from the mainstream only in degree rather than in kind. Thus, far right ideology represents 'a radicalization of mainstream views' (Mudde, 2008, p. 9). It is not identical with the mainstream (i.e. there is still a meaningful qualitative distinction to be made between the mainstream and the 'far' right) but, crucially, it is not entirely separate. The far right is simply the furthest point of an ideological continuum which includes the mainstream middle (Mudde, 2014). This theoretical argument has also been corroborated empirically.

John and Margetts find that in the UK many voters from across the political spectrum are *potentially* willing to vote for the BNP, indicating substantial latent support (Margetts, John, & Weir, 2004), a finding supported by Ford's analyses of attitudinal affinity between mainstream parties and the far right (Ford, 2010).

The 'pathological normalcy' thesis suggests that the study of Islamophobia within UK politics should not focus exclusively on the far right but should also recentre to include the mainstream. Indeed, reassessments of mainstream politics are already underway. In a study of Australian political discourse, Poynting and Briskman, utilising the language of the USA Presidential election campaigns, describe the spread of Islamophobia 'from deplorables to respectables' whereby 'liberal political leaders and press leader-writers who formerly espoused cultural pluralism now routinely hold up as inimical the Muslim folk devil' (Poynting & Briskman, 2018, p. 1). In a similar vein, Bayrakli and Hafez argue that, 'From Sweden to Greece, from Poland to the Netherlands, the rise of far-right parties is a vital threat to democratic order in Europe. What is more dangerous is the mainstreaming and normalization of the far-right policies within mainstream politics' (Bayrakli & Hafez, 2018, p. 13). The lack of attention paid to mainstream Islamophobia is itself a product of mainstream political discourse; Allen argues that the 2010-2015 Coalition sought to make Islamophobia a 'rather more exceptional and extraordinary' political phenomenon – thus ignoring the extent to which it is a distressingly everyday occurrence (Chris Allen, 2013, p. 7).

2.2.3 | UKIP and Islamophobia

The rise of UKIP during the late 2010s demonstrates the limitations of seeing Islamophobia as solely a problem of the far right. There is little academic consensus as to how it should be defined and ideologically situated – a problem heightened by the

rapidly changing support for the party. UKIP won the 2014 European elections, returning 24 MEPS, and received ~4 million votes at the 2015 general election, but then only ~600,000 votes at the 2017 general election and has suffered from many resignations since. UKIP has also made faltering attempts at finding a political voice following the Brexit vote in 2016 and has had a succession of new leaders (in the two years since Nigel Farage stepped down as leader in 2016, there have been four leaders). Two competing positions on the party exist in the literature: those who view UKIP as broadly part of the mainstream and those who think it is on the fringes of the far right.

Links between UKIP and both Labour and the Conservatives have been well-studied (Webb & Bale, 2014); Ford et al. report that much support for UKIP comes from 'strategic' voters who 'votes instrumentally for UKIP in European elections to express hostility to the EU but retain positive feelings towards the political mainstream, and returns to the Conservative Party at general elections' (Ford, Goodwin, & Cutts, 2012). Bale also argues that UKIP has close ideological ties with the Conservatives, arguing that 'the relationship between the radical right and its mainstream, centre-right counterpart is more reciprocal, and even symbiotic, than is commonly imagined' (Bale, 2018, p. 1). Mellon and Evans, in a lively academic exchange with Ford and Goodwin, contend that UKIP and the Conservatives appeal to a similar electorate, and that UKIP poses the greatest electoral threat to the Conservatives (Evans & Mellon, 2016; Mellon & Evans, 2016) – a position which Ford and Goodwin refute, contending that UKIP voters have more in common with Labour voters and, due to the geographical distribution of their supporters and the UK's first past the post electoral system, pose a greater threat to the Labour party (Ford & Goodwin, 2014b).

These arguments, and the empirical evidence they are based on, position UKIP as a mainstream party which has much in common with Labour and the Conservatives. UKIP

may have initiated a 'revolt on the right' (Ford & Goodwin, 2014a), but it is precisely a revolt within the *mainstream* of the right. This point is also made by Mudde, who explicitly argues that UKIP is not part of the 'populist radical right', a relatively new 'party family' which sits in between 'extreme right' parties and mainstream right wing parties (Ennser, 2012). Mudde argues that its only radical right policy is stridently rejecting the EU (Mudde & Kaltwasser, 2007). This position is supported by the party's dominant figure and former leader, Nigel Farage, who has repeatedly claimed that 'UKIP is not a racist party' (BBC, 2014) and that it is only a small number of bigots – who he terms 'idiots' – who mean the party is associated with the far right (The Guardian, 2014).

Nonetheless, many academics have long questioned whether UKIP has more in common with the far right than the mainstream. John and Margetts highlight links between UKIP's ideology and policies and those of the BNP (John et al., 2004; Margetts et al., 2004). Others also demonstrate that, particularly prior to their pre-Brexit surge in support, strong affiliations and ideological similarities exist between supporters of UKIP and those of the BNP (Bowyer, 2008; Hayton, 2010). In particular, Lynch et al. show that UKIP electoral candidates view the party as ideologically very similar to the BNP (on a 10-point scale of left to right, candidates placed UKIP at ~6.5 and the BNP at 7) (Lynch, Whitaker, & Loomes, 2011). Furthermore, whilst historically UKIP has sought to distance itself from extreme right politics, banning members of the EDL and the BNP from joining and rejecting an electoral pact with either the BNP or Britain First, during 2018 they have arguably moved far closer to the far right. In November 2018 Leader Gerard Batten claimed that Islamophobia was a 'made up word' and appointed ex-leader of the EDL Tommy Robinson as his special advisor (BBC, 2018d). Noticeably, several of the parties' elected representatives, including Nigel Farage, have left the party during 2018 due to a perception that it is has radicalized. Overall, the lack of consensus shows that UKIP is a

party which, more than any other in UK politics, straddles the boundary between mainstream and extreme.

Thus far, insufficient attention has been paid to Islamophobia within mainstream parties as Islamophobia has been viewed solely as a problem of the far right (bracketed as a dangerous 'pathology'). Anecdotal evidence suggests that the Islamophobia expressed in mainstream parties is qualitatively different (i.e. subtler and more nuanced rather than direct and aggressive) but this requires further investigation – and it is important to note that weaker varieties of Islamophobia are still pernicious and harmful. Studying the Islamophobia of social media followers of UKIP would also contribute to scholarship on the ideological nature and position of the party, specifically whether it is more akin to the mainstream or the far right. This, in turn, would provide insight into the wider UK political landscape and the normality of pathological behaviours such as Islamophobic tweeting. Ultimately, evidence that traits associated with the far right (such as Islamophobia) are widespread amongst mainstream parties could put into question long-held views that the UK is a case of 'far right failure' (Ignazi, 2003) and that the far right is 'forever a false dawn' (Goodwin, 2013a). Thus, building on the existing preliminary evidence, what is now needed is robust large-scale research into Islamophobic behaviour across mainstream parties in order to quantify the extent to which it exists.

This leads to the next research question:

> RQ 3: To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?

2.3 | Theories of Islamophobia

Understanding the drivers of Islamophobic behaviour is a growing focus of academic work. Theories have been put forward across the social sciences. Some approaches emphasize individuals' traits, focusing on either their personalities or socio-demographics; others emphasize group-level interactions or the role of events. In many cases, these theories examine slightly different (although largely overlapping) aspects of Islamophobia; studying what drives it is different to studying how it can be reduced, and Islamophobic behaviours are different to Islamophobic attitudes. Largely, the study of drivers of Islamophobia has been separated from the study of political parties, mainly because, as discussed above, (i) Islamophobia is bracketed off as a problem of only the far right and thus is not studied in relation to other political parties, and (ii) the far right is viewed, with a broad brush, as perpetually Islamophobic. An exemplar of this is Biggs and Knauss' detailed study of the geographical location of BNP members, in which they use BNP membership as a way of understanding 'why people denigrate or dislike minorities defined by ethnicity, race, religion, or foreign birth' (Biggs & Knauss, 2012, p. 633). This may be a reasonable assumption given their research goals, but nonetheless typifies the lack of attention paid to drivers of Islamophobia *within* the far right.

In the remainder of this section I consider five of the most prominent and applicable theories of Islamophobia available: (1) contact and conflict, (2) economics, (3) individual psychology, (4) social interactions and (5) cumulative extremism. I identify key limitations with the first four and focus on the fifth – the theory of cumulative extremism – as the most suitable for the current work. I put forward several proposals to enhance this theory, applying it in the context of both party politics and social media. I also link it to the large body of empirical research into the role of Islamist terrorist attacks in driving Islamophobic behaviour.

2.3.1 | Contact and conflict

Contact and conflict theory are two of the most pervasive explanations of Islamophobia and originate from social psychology. Contact theory suggests that bringing different groups (e.g. white Britons and migrant Muslims) into contact reduces prejudice by increasing individuals' knowledge of the other group and reducing feelings of anxiety about their perceived 'alienness' – thus a lack of contact can be used to explain intergroup prejudice (Hewstone, Rubin, & Willis, 2002; Pettigrew & Tropp, 2008; Pettigrew, Tropp, Wagner, & Christ, 2011). In contrast, conflict theory suggests that groups enter into conflict when they reside near each other as they enter competition for symbolic, political and economic resources. Both theories emphasize the importance of interactions between different groups and have been applied to online contexts. In particular, many are optimistic about the potential in online settings to enable low-cost multimedia interactions between members of different groups or 'e-contact' (White, Harvey, & Abu-rayya, 2015). However, in a recent paper Kim and Wojcieszak argue that the evidence on online contact is 'still limited' (Kim & Wojcieszak, 2018), as do both Schumann et al. and Walther et al., which suggests that, overall, research remains at a nascent stage (Schumann, Klein, Douglas, & Hewstone, 2017; Walther, Hoter, Ganayem, & Shonfeld, 2015).

There are two main reasons which suggest contact and conflict theories may not be useful for explaining prejudices (and their *extent*) which emerge in online settings. First, many individuals seek to create – or inadvertently create through algorithmic reinforcement – closed 'echo chambers' and 'filter bubbles' (Parisier, 2012; Sunstein, 2001). Thus, whilst opportunities for intergroup contact abound, they may not be pursued, which would make it difficult to test these theories. Second, on social media individuals can disassociate their social media personas from their offline identity (Suler, 2004). Although some

contend this could free individuals from 'the restraints imposed by group membership', such as normative expectations to behave prejudicially (Postmes, Spears, & Lea, 2002, p. 1074), a large body of research suggests that the opposite also happens – whereby individuals are empowered by anonymity, 'context collapse' and the absence of social cues to act in troll-like and prejudicial ways (Lowry, Roberts, Romano, Cheney, & Hightower, 2006; Marwick & boyd, 2011; White et al., 2015). More broadly, an issue with contact and conflict theories is that they focus on how individuals develop Islamophobic *attitudes* – rather than what drives them to engage in Islamophobic *behaviours*. For these reasons, contact and conflict theories are less relevant for the present work.

2.3.2 | Economics

Other theories suggest that economic circumstances, measured through income and unemployment status, are positively associated with hate crime prevalence (Catalano, Novaco, & McConnell, 2002; Curthoys, 2013; Kutneski, 2009). This association is explained primarily via the 'frustration-aggression' thesis, which suggests that individuals who have suffered economically experience frustration and then express it through aggression (Ryan & Leeson, 2011). There are parallels between this theory and 'modernisation theory' of far right support, which suggests that individuals who are (self-defined) losers from processes of modernization and economic growth are attracted to the far right in response to their changed circumstances (Ford & Goodwin, 2010; Ignazi, 1997; Mudde & Kaltwasser, 2007). Empirical support for the 'economic circumstances' thesis is at best partial. In a rigorous study using both historical and then-contemporary datasets, Green et al. found no evidence of a robust relationship (Green, Glaser, & Rich, 1998). And, equally, modernisation theory may provide insight (in some cases) into why individuals are attracted to the far right but not, in turn, what drives them to be

Islamophobic – or why individuals who support other parties engage in Islamophobia. More broadly, a key problem is that in online settings traditional marks of certainty, such as socio-demographics or economic status, do not necessarily have the same causal impact and as such might be less explanative (Margetts et al., 2015). Accordingly, I do not focus on the role of economic circumstances in driving Islamophobia.

2.3.3 | Individual psychology

The role of individuals' psychology has been extensively studied in relation to prejudice. Much research stems from Adorno's classic 1950 text *The Authoritarian Personality* (Adorno, 1950) and Allport's *The nature of prejudice* (Allport, 1954), especially his hypothesis that 'people who reject one out-group tend to reject other out-groups' (Allport 1954). This suggests that certain individuals are highly prejudiced *in general*, holding negative views about many different outgroups rather than just any one. This has motivated research which suggests that such individuals have an underlying propensity to be prejudiced, primarily because of their innate personality rather than socioeconomic traits (such as income) or contextual determinants (such as outgroup contact). In a meta-analysis of extant research Sibley and Duckitt report strong evidence in support of various personality-driven theories to understand prejudice, including Right Wing Authoritarianism (Altemeyer, 1998), Social Dominance Orientation (Pratto, Sidanius, Stallworth, & Malle, 1994) and the Big Five personality traits (Sibley & Duckitt, 2008).

There are considerable ethical concerns around using personality to understand individuals' online behaviour; capturing information about such traits is problematic following the widely publicised Cambridge Analytica personality-prediction scandal (BBC, 2018c). Furthermore, most of the available algorithmic methods for predicting personality rely on language traits within tweets which are likely the same traits which

can be used to identify Islamophobia within tweets; this introduces a clear risk of confounding (Agarwal, 2014). Finally, most research into personality traits has focused on understanding prejudicial *attitudes* rather than prejudicial *behaviour* and as such may not be appropriate to the present work. For these reasons, I do not focus on the role of personality type in driving Islamophobic behaviour.

2.3.4 | Social interactions

Research in social movement studies (Porta & Diani, 2006), radicalization studies (McCauley & Moskalenko, 2008) and political participation (Djupe & Sokhey, 2011) suggest that intra-group dynamics in the form of *social effects* can be used to explain individual behaviours, such as Islamophobic tweeting. The basic intuition behind what I term the 'social effects thesis' is captured by Della Porta and Diani, who write that 'individuals do not make decisions in isolation but in the context of what other people do, hence the importance of network connections' (Porta & Diani, 2006, p. 119). Broadly put, previous research indicates that there are two types of effect on individuals' political and social behaviour: (i) the social environment in which they are embedded in and *observe* and (ii) the social environment that they *interact* with, such as other people (Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2008; McClurg, 2006). In turn, this is due to two main mechanisms: (i) informational mechanisms, whereby individuals are provided with information and opinions which shape their preferences and attitudes and (ii) normative mechanisms, whereby individuals are encouraged to act in a certain way to meet group approval and avoid criticism and personal attacks (Heiss, Schmuck, & Matthes, 2018; Wojcieszak, 2008; Wojcieszak & Mutz, 2009).

The social effects thesis is particularly well-suited to understanding Islamophobic *behaviour*, rather than Islamophobic *attitudes*. Behaviour, as an outward-orientated

activity, is more susceptible to group influence, particularly normative mechanisms of influence, than internal belief formation. For instance, in a study of the white supremacist forum *Stormfront*, Wojciesak suggested that through 'meeting like-minded extremists online' individuals might 'activate their prejudices' (Wojcieszak, 2010, p. 643). Here, forum membership enabled and encouraged certain types of prejudicial behaviour, even if it potentially did not influence individuals' actual beliefs and values. This is akin to the study discussed earlier by Cheng et al.; in the right settings almost any user can be 'activated' to act as a troll online (Cheng et al., 2017).

Notwithstanding the power of the 'social effects thesis', it has several limitations for the current work. First, long but weak ties proliferate on social media (Centola & Macy, 2007; Ramasco, Moro, Pujol, Eguiluz, & Grabowicz, 2012), and as such there is considerable debate as to whether interconnected users on multi-use social media platforms constitute groups/communities (Cheung, Chiu, & Lee, 2011; Dunbar, 2016). The requirements of most group definitions are typically not met, such as the stipulation that group members must share a common sense of belonging or know most of the others in the group (Hogg, Abrams, Otten, & Hinkle, 2004; Hogg & Tindale, 2001). In practice, this means that social effects are less likely to operate on Twitter, even though they may function within small closed forums or offline amongst groups of friends and peers (Christakis & Fowler, 2007, 2008; Fowler & Christakis, 2008). For instance, the now-infamous 'emotional contagion' study on Facebook, in which the timelines of 689,003 users was manipulated to contain posts with more or less negative/positive words, found a significant but incredibly small effect (Kramer, Guillory, & Hancock, 2014). A similar large scale (61 million person) study by Bond et al. into how the provision of social information on Facebook drives political participation also found a relatively small effect (Bond et al., 2012). These problems are likely to be exacerbated on Twitter, which is

primarily used to *broadcast* opinions, and has 'no technical requirement for reciprocity, and often, no social expectation of such' (Marwick & boyd, 2011, p. 117). Thus, it is highly unlikely that the *interactions* part of the social effects thesis would be an effective explanation.

In addition, there are several methodological and research design issues which make the social effects thesis less favourable. First, with non-experimental data it is difficult to identify the direction of causality. For instance, it would be unclear whether individuals act Islamophobically *because* of their interactions with, and connections to, Islamophobic users or whether the opposite is true: they choose to interact with and connect to Islamophobic users because their behaviours are similar. Second, it is likely that interacting with Islamophobes could have countervailing effects on individuals. Some individuals might become more Islamophobic whilst other individuals would be pushed to become less, possibly as a response to observing and interacting with very extreme and explicit haters (Munger, 2017; Rudas, Surányi, Yasseri, & Török, 2017; Wojcieszak, 2010). Disentangling these countervailing processes could be either difficult or impossible, potentially invalidating any findings. Third, without any interventions, individuals' social connections and interactions may not change considerably during the period studied; this means that there would be insufficient 'action' to measure how – at an individual level – changes in context and interactions drive changes in behaviour. Fourth, changes in both (i) social context and interactions and (ii) levels of Islamophobia could be the product of a third variable, such as a political event. This event (or any other third variable) would be the true cause of changes in the prevalence of Islamophobia but if this is not explicitly modelled then it would appear as though social interactions were the driver. This is a very problematic source of confounding. Fifth, studying what users *observe* is impractical as it would require a data-sharing agreement with Twitter (or any

other social media platform). Because of these methodological and theoretical reasons, I do not explore the 'social effects' thesis in this work.

2.3.5 | Cumulative extremism

The theory of cumulative extremism was developed by Eatwell in 2006 to explain how 'one form of extremism can feed off and magnify other forms' (Eatwell, 2006, p. 205). It was initially proposed as a way of understanding the actions and support base of the far right (specifically, the BNP) in response to Islamist terrorism in the context of UK politics but has since been generalised to other settings and ideologies, such as Northern Ireland separatists (Carter, 2017) and far left movements in Greece (Gerodimos, 2015). The main insight of cumulative extremism is that extremisms are symbiotic. Bartlett and Littler summarise this concisely with their account of EDL street demonstrations against Islamist terrorism; 'EDL marches encourage radicalisation in Muslim groups, which in turn reinforces the EDL's casus belli' (Bartlett & Littler, 2011, p. 13). Cumulative extremism, which emphasizes both causal, organizational and attitudinal similarities between different types of extremists, has been widely adopted within academia and policy-making circles as a way of understanding the complex connections between extremisms, as well as addressing extremist behaviours and attitudes (Eatwell & Goodwin, 2010; Feldman, 2015; Ranstorp, 2010).

Busher and Macklin offer several proposals for augmenting and enhancing the theory of cumulative extremism's analytical power (Busher & Macklin, 2015). First, they recommend distinguishing between actions and narratives, detailing the need for researchers to explain *what* extremist actions, discourses or ideas leads to what *other* types of extremism actions, discourses or ideas. They emphasize the need for cumulative extremism to be used for social scientific 'process tracing' and identifying social

processes, such as through delineating causal mechanisms of cumulative extremism. This is particularly important given considerable work on the gap between actions and beliefs, and the complex pathways which lead individuals from one to the other (McCauley & Moskalenko, 2008, 2014). Extremism takes many forms, and researchers should be clear about which particular type (whether that is a behaviour, action or identity) they are investigating. Busher and Macklin's argument is well-aligned to the approach adopted in this thesis, where I am specifically focusing on speech acts (construing Islamophobic tweeting as a type of behaviour) rather than attitudes or ideology.

Busher and Macklin also propose that researchers should 'describe in detail the ebb and flow of interactions between the opposing "extremist" groups' (Busher & Macklin, 2015, p. 890). This is much needed given the short time spans in which extremists can radicalise and dangerous behaviours can emerge, particularly in online settings. This is also effectively a methodological recommendation, as increasing the level of detail within analyses favours using quantitative tools and large, granular datasets. At the same time, Busher and Macklin also call for greater *depth* in research; but this likely comes with a trade-off against *breadth*. For instance, Carter provides a broad multi-year review of processes of cumulative extremism within Northern Ireland from 1967 to 1972 but does so by focusing at a high level on groups rather than individuals (Carter, 2017). She does not provide detailed insight into how individual confrontations and acts of violence unfolded. More granular research designs – in which cumulative extremism processes could be measured within a time span of days if not hours or seconds – cannot provide the same coverage as Carter's work but can offer far more detailed insight.

Given the analytical and methodological strengths of the theory of cumulative extremism, I use this as the theoretical basis of the current work. I make two contributions to develop the theory further. First, I apply it in an online context. In previous research, the theory

has largely been applied offline – even though it is highly suited to understanding online political behaviour and discourse, which is often aggressive and polarised (Gruzd, 2014; Parsell, 2008; Wright et al., 2017). I anticipate that on social media processes of cumulative extremism are intense and short lasting as individuals are continually exposed to large volumes of content, as well as information about other individuals' social responses to that content, and have the ability to themselves respond very quickly (Hale, John, Margetts, & Yasseri, 2018). Most likely, extremist behaviour follows a 'flash in the pan' dynamic where it quickly erupts but then fades away. If this is the case, it would point to key differences between online and offline extremist political behaviour, and potentially also give insight into the nature of social media platforms themselves.

Examining the impact of extremist events online also raises questions about what 'cumulative' means in this context; if effects are only short-term and do not lead to either (i) a sustained increase in extremism, such as individuals sending more Islamophobic tweets several months after the attack or (ii) increasing frequency of periods of rapidly escalating hate (such as more outbursts of Islamophobia occurring within each time period), then it may be that 'cumulative' is an inappropriate term. Perhaps, the dynamics of extremism are better understood as *reactive* rather than cumulative. Reactive extremism would capture the idea that extremisms rapidly feed off each other but then do not continue to escalate or set in motion future extremisms; the impact of extremist events might be considerable but is also contained. This would offer an important refinement of the theory of cumulative extremism, and open a new analytic. It should be noted that lack of evidence for sustained increases in extremism may also just reflect that the dynamics of cumulative extremism are more complex and circuitous, manifesting in.a diffuse way. This would need further research to verify.

Second, I extend the focus of cumulative extremism theory. As currently articulated, it focuses solely on community and group level dynamics (i.e. how far right groups emerge in response to action by Islamist terrorist groups, which then motivates more individuals to join such terrorist groups and so on). This is a problem as the connection between individuals and groups is not always clear – as, indeed, Busher and Macklin question, 'whose actions represent the core processes of cumulative extremism? Social movements are very rarely homogeneous organizations' (Busher & Macklin, 2015, p. 889). They propose that the theory of cumulative extremism should be expanded to include different 'groups, sub-groups, factions, or cliques'. I argue it should go further to include processes of cumulative extremism *within the mainstream* – whereby the individuals at risk of cumulative extremism are not only individuals who are (or may become) supporters of explicitly far right groups, but also those with non-extremist, and seemingly non-prejudicial, political affiliations. This could manifest in one of at least two ways; individuals engage in more extremist behaviour but stay affiliated with the mainstream (potentially subtly changing the nature of the mainstream party to become more hateful) or they change affiliations from the mainstream to a niche extremist group. The latter is broadly in line with existing theorisations of cumulative extremism; different extremisms motivate both *more* extremist behaviour from already affiliated extremists but also enables extremist groups to recruit new members (Bartlett & Littler, 2011). However, if I find evidence of the former it suggests that extremism is not an easily demarcated niche issue (i.e. all extremists flock together) but something which can manifest, perhaps only briefly, in non-extremist settings. This extends the theory of cumulative extremism by showing that extremism is more complex and widespread than previously thought, and its dynamics more diffuse.

This analysis, in turn, raises the question of what the term 'extremism' means, and whether there is a risk of 'conceptual stretching' in this line of reasoning, specifically with regard to viewing all Islamophobic acts as extremist (Sartori, 1970). As discussed in the literature review, prejudice and hate is a key part of far right extremism and their ideology (Mudde, 2002). The intimate connection between prejudice and hate is also reflected in Government and policymaker understandings of extremism. Ed Balls, the head of the Department for Education in 2008 (then the Department for Children, Schools and Families) stated that, 'Extremists of every persuasion tend to paint the world as black and white, accentuating division and difference, and exploiting fears based on ignorance or prejudice.' (The Guardian, 2008). Islamophobia can therefore be seen as a form of extremism, particularly the most overt, directed and strong varieties. Nonetheless, due caution should be taken when interpreting the results of any analysis which equates hate speech (especially legally permissible hate speech) with extremism. This should be evaluated based on the particular context in which the research is conducted. Given the social sensitivity of Islamophobia, and its prominence within far right politics, this link is justified within this thesis.

Most studies of cumulative extremism focus on two sides of social antagonisms. In the context of this thesis' research, it suggests that Islamist terrorism and Islamophobia should not be viewed as separate phenomenon – whereby one straightforwardly causes each other – but are imbricated in a mutually reinforcing loop of causation. This implies that, at a theoretical level, it is misplaced to search for a single origin of extremist behaviour in society. At the same time, for the purposes of conducting empirical research it is often necessary to narrow research to just one side of any symbiotic cumulative extremism process. To operationalize the theory of cumulative extremism, I focus on one step in the cycle of extremism; how one extremist action (an Islam*ist* terrorist attack)

drives another (Islam*ophobic* tweeting). In doing so, I connect the theory of cumulative extremism with a large body of empirical work on the role of terrorist attacks in contemporary politics which, whilst not explicitly situated in relation to this theory, can be seen to corroborate its basic tenets.

Several studies show that the volume of hate crime spikes following terrorist attacks, emphasizing the importance of *temporal* as well as *geospatial* dynamics in understanding prejudicial behaviours (Borell, 2015; Miro-Llinares & Rodriguez-Sala, 2016). Byers et al. estimate the impact of the 9/11 terrorist attack on offline hate crime in the USA, showing that there was a significant increase in anti-Islamic hate crimes in its aftermath and that this lasted for approximately eight days – which suggests the impact is short and sharp (Byers & Jones, 2007). King and Sutton use a lagged negative binomial model to measure the impact of same-sex marriage laws, terrorist attacks and the Rodney King and O.J. Simpson verdicts on various types of offline hate crime (King & Sutton, 2013). They also find that there is a short sharp increase in hate crime, which lasts only a short period. Hanes and Machin use a fixed effect model to measure the impact of both the 9/11 and 7/7 attacks on offline hate crime in the UK (Hanes & Machin, 2014). Interestingly, they do not find that the level of hate crime increases rapidly and then declines in the immediate aftermath to its starting level. Rather, they show that the level of hate stays at a higher level even one year after. They argue this points to a fundamental change in attitudes towards Muslims following attacks. It also provides evidence of a genuinely *cumulative* process of extremism rather than just a *reactive* extremism, as discussed earlier.

Burnap and Williams model cyberhate in response to the Woolwich Islamist terrorist attack (Williams & Burnap, 2016). They fit a zero-inflated negative binomial model to measure both the size of information propagations following the terrorist attack and the

lifetime of each information flow. They identify racial and religious hate speech using a supervised machine learning classifier (Burnap & Williams, 2015), and test several hypotheses. Their study shows that cyberhate manifests rapidly following a terrorist attack, peaking in the first 24 hours – they describe this as an 'impact stage' in which the social reaction is amplified. However, cyberhate has a 'half-life' and there is a 'rapid de-escalation post-impact' (Williams & Burnap, 2016, pp. 232, 234). Their results show at a very granular level that cyber hate bubbles up fast but then dies away and that prejudicial narratives do not gain widespread acceptance even during such emotive and difficult periods. This study, as well as their prior research into the broader social media response to the Woolwich attack (Burnap et al., 2014) is the most relevant recent work informing this thesis.

Borell summarizes the key finding of research into the role of terrorist attacks as 'terrorist attacks instil a sense of uncertainty and risk and Islamophobia and hate crimes are to a large extent event-driven and reactive, and tend to flare up on the heels of dramatic events' (Borell, 2015, p. 409). He notes that one risk of emphasizing the role of terrorist attacks is to ignore the deep roots of Islamophobia in the UK, and the West more generally (Borell, 2015). However, the key insight of research in this area is not that Islamophobic *attitudes* are formed in response to Islamist terrorism – a point which is somewhat addressed by Hanes and Machin but requires considerably greater empirical investigation through multi-year or multi-decade panel data. Rather, it is that Islamophobic *behaviours* emerge in response. This reflects evidence that behaviours can change far more rapidly and greatly than attitudes (Fishbein & Ajzen, 2010; Maio, Haddock, & Verplanken, 2019; Vogel & Wanke, 2016). Potentially, as Cheng et al. argue apropos trolling on social media (Cheng et al., 2017), any individual can be driven to

engage in Islamophobic behaviours, even though this might not reflect adoption of an Islamophobic mindset.

A key limitation of existing research is that none of the studies are methodologically individual as they rely on aggregate datasets, such as statistics provided by the police or based on hashtag datasets from social media platforms. This makes it difficult to disentangle crucial questions, such as whether terrorist attacks drive new people to act Islamophobically, or whether the same people simply act *more* Islamophobically, and whether the *same* individuals are recurrently Islamophobic across different attacks. A further limitation is that offline studies – such as the research by Byers et al., Hanes and Machin, and King and Sutton – are not very granular, with data only recorded for each day. This could miss important dynamics at the level of hours or minutes. Finally, offline studies only focus on legally defined hate crimes, which is a high bar to meet. It is plausible that legal 'everyday', but still prejudicial, behaviours are even more powerfully affected by terrorist attacks as the barrier to engaging in legal prejudicial attacks is far lower. There are several actions that can be taken to respond to these issues and extend existing empirical research into how Islamist terrorist attacks drive Islamophobic behaviour. First, focus on a specific set of users for a defined period of time (such as one year), thereby enabling analysis of how individuals' behaviour changes. Second, using very granular digital trace social media data, which records the timestamp of tweets to the second, to increase the level of detail. Third, focusing on a broader category of hate, such as Islamophobic hate speech, which includes both legal and illegal behaviours. Fourth, studying several Islamist terror attacks (or other relevant political events) rather than just one.

Through this work I will bring into dialogue the theory of cumulative extremism and the large body of empirical research into the role of Islamist terrorist attacks. This could open

new avenues of future research, as well as enhancing current understanding of how these events impact Islamophobia within the UK. Accordingly, this leads to the next research question:

> RQ 4: To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?

This research is situated within the context of UK party politics on social media, considering both far right and mainstream political parties. If I find evidence that cumulative extremism affects followers of different political parties then it would indicate that (i) the theory should be extended beyond extremist groups to also include extremist individuals within the mainstream and (ii) that there are fundamental similarities in the behavioural dynamics of followers of different parties. As discussed, exactly what form of cumulative extremism operates will also be investigated. This analysis can be linked to RQ 3, outlined above, regarding differences in Islamophobic behaviour across followers of different parties. As such the final research question is:

> RQ 5: Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?

2.4 | Studying Islamophobia on social media

Studying Islamophobia on social media poses considerable methodological challenges due to the messy and unstructured nature of the data. This has several methodological implications for identifying hateful and harmful content. First, the huge volume of social media content makes it difficult to *find* the hateful or undesirable content. Second, there are many different types of harmful online behaviours, including hate speech (such as racist and sexist content), abusive content and harassment (commonly known as trolling and 'doxing') and content that enables illegal, harmful and intrusive behaviour (such as sharing images of a sexual or paedophilic nature). This makes it difficult to allocate resources and prioritize responses. Third, monitoring online spaces through content classification is a deeply contentious issue as malign online content is illegal or in contravention of platforms' codes of conduct in only very few cases. This is reflected in the media response to the University of Indiana's project studying information sharing on social media (known as "Truthy"), which was covered critically by the American media and strongly attacked by Fox News (Columbia Journalism Review 2017).

Social media platforms each have specific socio-technical affordances, which can introduce additional challenges for classifying content. Twitter limits the length of posts to just 280 characters (increased from 140 characters in 2017) – a feature of the platform which shapes the way in which users engage with each other and produce content. Much political science research indicates that the design of Twitter is not conducive to deliberative democratic dialogue but, rather, aggressive antagonistic interactions (Alorainy, Burnap, Liu, & Williams, 2018; Conover, Ratkiewicz, & Francisco, 2011). The constraint on post length also affects the type of content that users produce. Because only a few characters are allowed, users often link to media from outside Twitter. They also often split statements across multiple tweets into a single 'thread'. These platform-

specific behaviours make it challenging to study tweets, heightening the need for a robust method.

In this research I opt to use a supervised machine learning classifier to measure Islamophobic hate speech, which will be informed by the conceptual work undertaken in order to answer RQ 1 ("What is the conceptual basis of Islamophobia?"). The classification of Islamophobia hate speech is discussed in detail in the Methods chapter.

Accordingly, I stipulate an additional research goal:

> RG: To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia

2.5 | Conclusion

In this literature review, I have expanded the thesis's research aim:

> *To understand the nature and dynamics of Islamophobia amongst followers of UK political parties on Twitter*

To identify five research questions and an additional research goal:

- RQ 1: What is the conceptual basis of Islamophobia?
- RQ 2: To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?
- RQ 3: To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?
- RQ 4: To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?
- RQ 5: Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?
- RG: To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia

In the next chapter (Chapter 3), I provide an overview of the methods and research design. I then answer RQ 1 in Chapter 4, conceptualizing Islamophobia in terms of negativity and generality. I then fulfil the additional research goal in Chapter 5 by using these conceptual arguments to create a supervised multi-class machine learning classifier. This is used in Chapter 6 to answer RQ 2; existing work in political science suggests that the far right is constantly and extremely Islamophobic but, drawing on work in internet studies, I anticipate that users will exhibit considerable heterogeneity. In Chapter 7 I first

answer RQ 3; I anticipate that followers of different parties will exhibit different strength and prevalence of Islamophobic hate speech and that Islamophobia will not be confined solely to the far right. I then answer RQs 4 and 5; I anticipate that Islamist terrorist attacks will drive a considerable increase in Islamophobic hate speech, and that this will affect followers of all parties – this is because I anticipate that Islamist terrorist attacks will motivate many otherwise non-Islamophobic users to engage in Islamophobic hate speech.

Answering these research goals will enable me to make conceptual, methodological and theoretical contributions to our understanding of Islamophobic hate speech amongst followers of UK political parties on Twitter. It is hoped that the findings generated here will also contribute to broader debates in social scientific research, such as the nature of social media and the far right, as discussed in Chapter 8.

Chapter 3 | Research approach, methods, data and ethics

In this Chapter I define and justify the research approach (complementary computational social science). I also explain the data collection process and discuss relevant ethical issues. I provide only a brief overview of the methods because these are very different across the thesis and are discussed in detail in each of the empirical chapters.

In Section 3.1, I start by critically reviewing literatures on big data and computational social science. Much previous work in this area has tended to either focus on the social aspects of computation (McCosker & Wilken, 2014; Seaver, 2018) or has been purely computational – here, I argue for an approach in which the social and computational are tightly integrated. Specifically, drawing on the work of Blok and Pederson I advocate for a 'complementary' computational social science in which computational social science is both *social* and *computational* - and where both aspects are recognised as being of equal importance (Blok & Pedersen, 2014). I then consider how this approach relates to inductive and deductive scientific logics, and observational and experimental research designs. I argue that a complementary computational social science straddles both of these divisions, bridging traditional divides in the social sciences.

In Section 3.2, I provide an overview of the methods used in this work. Drawing on the previous arguments, I select methods based on both their suitability to the task at hand and also how they fit into the wider research design. This ensures that the complementary computational research approach is realised in practice. I provide considerable detail around the decision to use a supervised machine learning classifier to measure Islamophobia.

In Section 3.3, I discuss the choice of data source (Twitter) and describe the data collection process. I outline the results of a small pilot study, evaluating the impact of using different data collection frequencies. I find that weekly data collection is less onerous than daily collection and sufficient for the goals of this project. I also report on the results of a separate pilot study, which the nature of party followership on Twitter. I show that followers of political parties are overwhelmingly positive or neutral towards the party, providing an initial validation check of the theoretical argument (presented in the literature review in the previous chapter) that social media followers are constitutive elements of contemporary parties.

In Section 3.4, I consider different ethical issues which pertain to the present work, including consent, anonymization, harm, and the distinction between public and private data. To ensure the ethical integrity of this work, I opt to not present any information which could lead to individuals being identified, present results in aggregate where possible, and amend, partially redact and synthesize any individual tweets which are presented.

3.1 | Research approach

3.1.1 | *Computational* social science

Contemporary research is marked by the increasing use of large scale datasets, statistical analyses and algorithms. This 'data deluge' (The Economist, 2010) or 'data revolution' (Mayer-Schönberger & Cukier, 2013) has generated much discussion within the social scientific community as to whether, and in what ways, so-called 'big data' can be used to advance knowledge (Blok & Pedersen, 2014; boyd & Crawford, 2012, 2015; Housley et al., 2014; Kitchin, 2014; Lazer, Kennedy, King, & Vespignani, 2014; Lazer et al., 2009; Lazer & Radford, 2017; Margetts, 2017b; Watts, 2007). Computational social science has been closely linked with the emergence of big data. This is not surprising given that the 'computational turn' was largely necessitated by the widespread emergence of datasets too large for human individuals, or even research teams, to analyse (Berry, 2011). Or as, Boyd and Crawford put it, 'Big Data not only refers to very large data sets […] but also to a computational turn' (boyd & Crawford, 2012, p. 665).

Many researchers, particularly those situated at the intersection of social science and computer science, have welcomed the 'computational turn' and the 'promise of big data' (Labrinidis & Jagadish, 2012). For instance, Housley et al. discuss how 'big and broad datasets', which are often publicly available, enable researchers to pose and answer new types of questions and respond more rapidly to emerging social issues (Housley et al., 2014). Lazer et al. argue that computational social science leads to a newfound level of granularity and breadth, with little trade-off between them (Lazer et al., 2009). Watts argues that big data can provide insight into age old questions by 'mak[ing] visible social processes that are much more difficult to study in conventional organizational settings' (Watts, 2007, p. 489) and in political science Grimmer and Stewart argue that 'automated

content methods can make possible the previously impossible' (Grimmer & Stewart, 2013). The real promise of big digitized datasets is that social science can have its cake and eat it too. As Latour et al. put it, it can be rendered 'quantitative without losing [its] necessary focus on the particulars' (Latour, Jensen, Venturini, Grauwin, & Boullier, 2012, p. 613). However, there is also considerable disagreement as to what big data entails and how it differs (if at all) from the use of traditional datasets and methods (Akoka, Comyn-wattiau, & Laou, 2017; De Mauro, Greco, & Grimaldi, 2016).

The view that 'big data' and computation simply refers to the use of statistics and machine learning methods is embodied in the most widely used definitions, which characterise big data and computation in terms of either the technical infrastructure required (e.g. the use of cloud servers) (Wu, Zhu, Wu, & Ding, 2014) or the features of the data, such as the so-called Five Vs (usually, velocity, veracity, variety, value and volume – but there can be others (Uprichard, 2013)). Other approaches define big data not in terms of quantity but *quality*. Savage and Burrows characterise big data in terms of 'social transaction data' (Burrows & Savage, 2014; Savage & Burrows, 2007) – which is also called digital trace data (Howison, Wiggins, & Crowston, 2011). Transaction data can be defined as 'evidence of human and human-like activity that is logged and stored digitally' (Freelon, 2014, p. 59). This data is created as a by-product of individuals engaging in everyday behaviour, such as shopping online, making phone calls or using social media. It includes hyperlinks, shared online content, user-generated text and online social connections, and often has very granular metadata, such as timestamps accurate to the minute or second.

Studying socially generated data is useful for social scientific research as it avoids the 'Hawthorn effect', which often negatively affects the validity of research conducted using traditional methods, such as surveys and experiments. This states that participants in research projects do not act as they would normally *because* they are being researched –

'the consequent awareness of being studied […] [has a] possible impact on behaviour' (McCambridge, Witton, & Elbourne, 2014, p. 267). This is particularly like to be an issue when studying Islamophobia as it is a highly contentious and sensitive subject. Studying social media behaviour on Twitter should provide better insight into how individuals actually engage in Islamophobia rather than a version of sanitized Islamophobia which has been distorted by the presence of an intrusive researcher (Margetts et al., 2015). In recent times, socially generated data has become widely available due to the low cost and technical ease of corporate record keeping and the proliferation of Application Programming Interfaces (APIs). This data is particularly valuable to social scientists because it is not mediated by traditional research artefacts, such as surveys and questionnaires, which can introduce considerable biases (Lazer et al., 2009; Margetts, 2017a; Watts, 2007). Contrasted with purely technical definitions of big data (such as those which emphasize size), Savage and Burrows' definition is useful because it highlights the *social* aspect of big data and how it is produced.

A major issue with computational analyses, particularly those which are algorithmic, is that they can reproduce, and potentially even create, social biases and implicit forms of discrimination and oppression – despite often giving the appearance of objectivity (Golbeck, 2018; Noble, 2018). This issue relates to other problems with computational analyses, namely that they pose considerable ethical problems and can enable new forms of social control. This is evinced by a recent paper from researchers at Stanford which used facial images to classify individuals' sexual orientation (Wang & Kosinski, 2017). The study has received much condemnation in the media (The Economist, 2017), partly because of its ethical limitations, but also because it demonstrates how seemingly neutral technologies (such as machine learning) can be used for problematic or nefarious goals. A further, related, problem is that computational tools are often very hard to interpret.

Most computational methods are 'black boxes' to outsiders, and in some cases even for those who have used them (Wachter et al., 2017). Accordingly, whilst recognising the immense opportunities computational analyses afford, researchers should be attuned to the social and ethical implications of such work and the fact that technical artefacts which seek to capture social phenomena are themselves also social entities (Orlikowski & Scott, 2008).

3.1.2 | Computational *social* science

Big data computational social science is often contrasted with in-depth ethnography and qualitative approaches, or so-called 'small data' analyses (Kitchin, 2014; Kitchin & McArdle, 2016; Wang, 2013). A widely held view is that big data is useful for generating wide, precise but shallow knowledge, produced at a distance from the object of study, whilst small data is useful for generating rich, deep and thick knowledge, produced up close to what is being studied (Geertz, 1973). Manovich makes this point in relation to the benefits of ethnography, arguing that 'algorithms used by computer scientists […] will never arrive at the same insights and understanding of people and dynamics in the community' (Manovich, 2011, p. 8). boyd and Crawford similarly make the point that some stories and processes cannot be uncovered just 'by farming millions of Facebook or Twitter accounts' but, instead, require in-depth qualitative and ethnographic work (boyd & Crawford, 2012, p. 670). This dichotomy is useful insofar as it highlights the importance of not simply adopting a quantitative approach to all research tasks – but, at the same time, it risks creating even bigger issues regarding how researchers become bifurcated into separate methodological camps. Part of the problem is that qualitative social scientists and ethnographers mythologize big data and computation, treating them as technical artefacts of research which is essentially incomprehensible.

The dichotomy between computational and qualitative research (and big- and small- data (Onwuegbuzie & Collins, 2007)) is theoretically problematic as computational analyses are never *purely* computational; as Seaver puts it, 'If you cannot see a human in the loop, you just need to look for a bigger loop' (Seaver, 2018, p. 378). Similarly, some contend the inverse: that *all* research, whether computational or not, takes place in a big data context. For instance, Savage and Burrows argue that there has been a 'crisis' in empirical sociology whereby big data has challenged traditional sociology's 'claim to jurisdiction over knowledge of the social.' (Burrows & Savage 2014).— In contemporary computational research the *computational* aspect is often more heavily foregrounded than the *social*; as Cihon and Yasseri note, 'despite its name, [computational social science] has drawn from computer scientists, mathematicians and physicists far more than social scientists' (Cihon & Yasseri, 2016, p. 7). The outsized emphasis on the *computational* aspect of computational social science has led to its misplaced association with what van Dijck calls 'dataism'; the 'belief in the objective quantification and potential tracking of all kinds of human behaviour and sociality through online media technologies' (van Dijck, 2017, p. 198). This is a reductive approach to computational social science – it conceptualizes it as merely the application of computation to social research.

The view that computational social science needs to pay more attention to its social dimension is increasingly well accepted (Cowls & Schroeder, 2015). Grimmer argues that the promise and power of big data necessitates *more* rather than *less* social scientific rigour as the era of big data 'is as much about social science as it is about computer science' (Grimmer, 2014, p. 80). Elsewhere, Kitchin suggests that using larger datasets and more advanced methods does not obviate but actually enhances the role of the researcher; more complexity means more interpretation is needed as 'patterns found within a data set are not inherently meaningful' (Kitchin, 2014, p. 4). Snijders et al.

similarly argue that social science theories need to act as 'guidance' for big data analyses (Snijders, Matzat, & Reips, 2012, p. 2) Snijders et al.'s argument is supported by considerable empirical evidence apropos the limitations of using data mining methodologies for social analysis; Calude and Longo find that 'very large databases have to contain arbitrary correlations […] too much information tends to behave like very little information.' (Calude & Longo, 2017, p. 595) Similarly, a large body of research shows how, without due interpretative work, computational analyses are at risk of manipulation and abuse through p-hacking and results phishing (Cohen, 1994; Gelman & Loken, 2013; Vidgen & Yasseri, 2016).

Blok and Pederson, drawing on Latour et al.'s work on digital datasets (Latour et al., 2012), argue that to produce research which is quantified but nuanced, the mutual interdependence of the social and the computational needs to be recognised (Blok & Pedersen, 2014). That is, the emergence of big data should not mark the 'end of theory' (Anderson, 2008) and well-specified research designs but the development of a 'complementary social science' in which qualitative and quantitative methodologies are used in tandem as part of a unified research design (Blok & Pedersen, 2014). Hamann and Suckert make a similar point regarding the need for a 'quantified qualitative approach' to social science (Hamann & Suckert, 2018). Indeed, integrating social and computational analyses is crucial for realising the 'promise' of big data and computation, which are supposed to provide 'unprecedented breadth and depth and scale' (Lazer et al., 2009, p. 722) and 'a *deeper*, clearer understanding of our world' (Lazer et al., 2014, p. 1205). If computational social science is not also *complementary,* then it is unlikely that this 'depth' will be realized. Arguably, this is not fully recognised in the field. For instance, in an article about big data and social science, Grimmer and Stewart argue that 'we are all social scientists now' (Grimmer & Stewart, 2013)– a claim which is arguably

inaccurate descriptively, even though it works well as a normative injunction: everyone working with big data *should* be a social scientist.

Complementary computational social science might, at first sight, seem similar to multi-methodological research design, an increasingly popular way of conducting research (Onwuegbuzie & Collins, 2007). However, there are important ontological differences between the two. In multi-methodological research, separate methods are 'combined' in order to compensate for the others' weaknesses (Steckler, McLeroy, Goodman, Bird, & McCormick, 1992; Wang, 2013). This is well embodied by 'triangulation', a widely used way of increasing the validity and reliability of findings by using multiple methods, research sites and datasets (Humble, 2009; Mathison, 1988). The goal of triangulation is to find a 'singular proposition about the phenomenon being studied' (Mathison, 1988, p. 13) – paradigmatically, the separate methods are used simply to reinforce a single result by accumulating more (and varied) evidence. Complementary social science differs fundamentally from mixed methods approaches in that it offers 'a reconfiguration of traditional splits between quantitative and qualitative research methods' (Blok & Pedersen, 2014, p. 3). It rejects the idea that qualitative and computational methods are inherently separate and should be merely combined in an additive manner. It suggests, instead, that computational works must be fully integrated with the social (at least, when computation is used in the social domain). The present work fits within the complementary computational social science research philosophy.

3.1.3 | Logics of scientific inquiry

Different types of research rely on different scientific logics, of which two particularly salient: deductive logics and inductive logics (Glynos & Howarth, 2007). Inductive methodologies start from the data and then identify relevant categories through a process

of exploration and critical reasoning. In contrast, deductive methodologies start from pre-defined categories and search for them within the data (Elo & Kyngäs, 2008; Hsieh & Shannon, 2005). Another key difference is that inductive reasoning starts from specific instances and aims to establish generalisations, whilst deductive reasoning starts from generalisations and then tests whether these apply to specific instances (Hyde, 2000). Traditional approaches suggest that research must predominantly fit into one category of scientific logic; as Morse puts it, 'All projects have either an inductive or a deductive drive; they can neither be neutral nor informed equally be deductive and inductive studies' (Morse, 2003, p. 197). However, this view has been challenged in recent times.

Combining inductive and deductive research logics is increasingly commonplace in some disciplines, such as the field of text analysis. For instance, Fereday and Muir-Cochrane discuss how a 'hybrid' approach, incorporating both inductive and deductive logics, should be adopted for thematic coding (Fereday & Muir-Cochrane, 2006). Integrating inductive and deductive logics is even part of 'grounded theory', a supposedly inductive approach to text analysis. For instance, Backman and Kyngas describe how grounded theory researchers engage in 'a complex process of inductive and deductive thinking' (Backman & Kyngäs, 1999, p. 250). Noticeably, computational social science can incorporate both inductive and deductive logics. Indeed, Conte et al. argue that computational methods can enable 'an escape from the deductive/inductive dichotomy' (Conte et al., 2012, p. 340). There are good reasons to think that both logics *should* be incorporated within single projects; inductive qualitative research can be used to identify initial trends and relationships, which can then be systematized and measured through the use of computation. Accordingly, the present work draws on both inductive and deductive logics of inquiry, according to the requirements of the different parts of the research.

3.1.4 | Experiment and observation

A key distinction in research is whether the design is experimental or observational (Imai, Keele, Tingley, & Yamamoto, 2011). Both designs can be used for identifying and testing causal relationships, although randomized controlled trials are typically viewed as the 'gold standard' of evidence (Cartwright, 2007; Christ, 2014) as observational studies can suffer from confounding and it can be difficult to identify the direction of causality (Grimes & Schulz, 2002; Shalizi & Thomas, 2011). Observational research designs are more flexible, and can also be used to answer descriptive, interpretive, critical, exploratory and conceptual, research questions. In the present work data is collected from Twitter (discussed in detail below), which can be understood as a form of digital trace 'big' data, as discussed above. Paradigmatically, the use of digital trace data is associated with observational research designs – in the context of health research, Khoury and Ioannidis make this association explicit, claiming that 'big data *are* observational in nature' (Khoury & Ioannidis, 2014, p. 1054). However, in social science, researchers have used big data in both experimental and observational research designs, depending upon their goals.

Experimental big data research on social media takes one of two forms. First, researchers manipulate a platform through a randomized controlled trial. This involves developing a relationship with the platform provider and is typified by the infamous 'emotional contagion' study by Kramer et al.. The content of Facebook users' News Feeds was manipulated, and its impact on how much the users used positive words was measured (Kramer et al., 2014). Similarly, Bond et al. studied how manipulating users' feeds to contain more social information pertaining to the 2010 US congressional elections influenced their propensity to vote (Bond et al., 2012). Both these studies posed

considerable ethical issues in that the researchers directly interfered with the social reality they sought to understand. Second, researchers exploit natural change in the platform design, such as when social media companies introduce new user features. This type of experimental design poses far fewer ethical issues and can be equally robust. An example is provided by Hale et al., who test how changes to the design of a digital platform, namely a petition website, influenced political participation. The changes were implemented by the platform owner, which unintentionally 'creat[ed] the conditions for a natural experiment' because visitors experience the change 'as if it were random' (Hale et al., 2018, p. 3). Experimental studies should be seen as exemplars of big data research as the data is transactional and large scale.

The current project is, broadly, observational in nature – the Chapter 4 inductively examines the philosophical basis of Islamophobia, Chapter 5 is methodological, Chapter 6 is observational, and Chapter 7 is observational but uses naturally occurring events. In this sense, it is a mix of observational and experimental research designs.

3.2 | Methods overview

Different observational methods are used for each chapter. The choice of method is based upon the chapter's function within the overall research design and the nature of the research task at hand, as advised by Shapiro (Shapiro, 2002).

The goal of Chapter 4, the first empirical chapter, is to understand and conceptualize the nature of Islamophobia on Twitter. This is the first part of the complementary research design. The research is primarily inductive in nature as it is inherently explorative, seeking to work from the ground up to understand Islamophobia as it is actually expressed. Both quantitative and qualitative methodologies for text analysis could be used in Chapter 4. Baumer et al. discuss how unsupervised machine learning methods, such as topic modelling, can be used in exploratory applications to derive insights inductively (Baumer, Mimno, Guha, Quan, & Gay, 2017). In one intervention, Mimno et al. use topic models to investigate far right discourse within Swedish parliamentary debates (Mimno, Magnusson, Barrling, & Ohrvall, 2017). Nonetheless, I opt to use a qualitative methodology in Chapter 4. This is in-line with nearly all previous conceptual and thematic investigations of Islamophobia on social media (Awan, 2014, 2016; Awan & Zempi, 2016; Ekman, 2015; Jacks & Adler, 2015; Larsson, 2007; Lee, 2017). Qualitative text-analysis is particularly well-suited because the task at hand involves not only analysing the content of tweets thematically but also critically examining them philosophically. In this regard, this chapter is akin to previous qualitative text analyses undertaken by Taras, Iqbal and Saeed, each of which investigate conceptual and empirical manifestations of Islamophobia (Iqbal, 2010; Saeed, 2007; Taras, 2013).

The goal of Chapter 5 is to create a machine learning classifier to detect Islamophobic hate speech. I use computational methods here, drawing on the findings from the

preceding chapter (Chapter 4), in which I thematically and philosophically investigate the conceptual basis of Islamophobia. This is discussed in more detail below (Section 3.2.1).

Chapters 6 and 7 both use the machine learning classifier from Chapter 5 to statistically analyse Islamophobic behaviour. In Chapter 6 I fit a latent Markov chain model and in Chapter 7 I fit several longitudinal regression models, including fixed effects regression, negative binomial regression, and segmented variants of each. These statistical analyses are the final point in the complementary computational social science research approach, quantifying and providing precise insight into the dynamics of Islamophobia amongst followers of UK political parties on Twitter.

3.2.1 | Measuring Islamophobia in Language

Both qualitative and quantitative approaches can be used to detect and measure Islamophobia. To realise the Research Goal in this project in a scalable and time-efficient manner, I opt to use a quantitative approach. I develop a new theoretically-informed and contextually-specific machine learning classifier, which is a form of Natural Language Processing (NLP).

Many NLP methods exist for text classification, from semi-automated keyword searches to statistical network analysis of word relationships to machine learning. Machine learning is arguably the most promising computational method for classifying content. It can be defined as the process of enabling computers to learn without being explicitly programmed (Samuel, 1959). There are two main types of machine learning: supervised and unsupervised. Supervised machine learning presents the computer with example data in which the correct output has been labelled and relevant input features identified (called ‘training data’). The computer learns from the training data how to best classify each

instance into the outputs, and then uses this to classify new unlabelled data. In contrast, unsupervised machine learning presents the computer with unlabelled data and then clusters it into groups.

Supervised machine learning has been used in many political science applications, including detecting hate speech on social media (Williams & Burnap, 2016), coding policy issues (Burscher, Vliegenthart, & De Vreese, 2015), and monitoring the aggressiveness of states' foreign policy discourse (Stewart & Zhukov, 2009). Supervised machine learning has enabled researchers to measure and quantify processes and behaviours which could previously only be studied qualitatively. Such methods enable more precision in that error and bias can be reliably tested and evaluated. The methods are also more consistent as they can be implemented with less subjective judgement. Computational approaches are also highly scalable, which can increase the overall robustness of research by analysing larger quantities of data. Some have questioned whether machine learning can truly replicate the nuance and insight of a human annotator in complex tasks, studies show that for reasonably simple tasks – where, for example, only two to five categories are annotated. However, the performance of classifiers can be tuned to closely approximate human annotators, provided that sufficient training data is used (Collingwood & Wilkerson, 2011).

Most previous research on hate speech detection has focused on hate speech detection in general, rather than the more specific and nuanced task of Islamophobic hate speech detection (Schmidt & Wiegand, 2017). Nonetheless, many prior studies are still relevant to the present discussion, not least because most of them use Twitter data. The main focus of prior research has also been binary rather than multi-class classification tasks (Schmidt & Wiegand, 2017). Multi-class classifiers enable far more nuanced insight into online hate speech, and as such are likely to be more appropriate to the present work. However,

performance is typically far lower for multi-class rather than binary tasks. As Salminen et al. noted only in 2018, 'existing works using multi-label classification for online hate speech are extremely rare, and we could not locate prior work that had achieved good results.' (Salminen et al., 2018, p. 331) The few multi-class hate speech studies can be divided into one of two groups. First, are studies which aim to identify different targets of hate. These often trade off nuance against breadth, typically focusing on only the more 'extreme' forms of hate within each targeted group.

Burnap and Williams train a classifier to distinguish between racism, homophobia and anti-disability prejudice (Burnap & Williams, 2016). They use a 'blended model' which enables them to not only study each type of prejudice but to also study them in combination, enabling analyses of intersectional hate. This is an important step forward in the study of hate, given prior research on intersectional prejudice (Zempi & Chakraborti, 2015) and also the extent of the challenge. As the authors note, 'individual models of cyber hate do not generalise well across different protected characteristics' (Burnap & Williams, 2016, p. 11). For the blended model they achieve precision of 0.85 and recall of 0.54. Park and Fung test both a 'one-step' and a 'two-step' approach to classify racism and sexism in tweets (Park & Fung, 2017). In the two-step approach tweets are first classified for whether or not they are offensive, and then they are sub-categorised into racist and sexist classes. Performance of both approaches is similar, but Park and Fung recommend the two-step approach for more complex datasets. They report an F1-score of 0.827.

Salminen et al. train a multi-class classifier on 21 distinct categories and sub-categories of hateful targets in Facebook and YouTube comments, achieving an F1-score of 0.79 (Salminen et al., 2018). This high performance is possible because the targets vary considerably, including religion, the army, the media and financial power. Saleem et al.

use a community-based lexicon approach to identify hate targeted against Black people, overweight people and women on Reddit, Voat and various web forums. Their method consists of three separate classifiers rather than a single multi-class classifier and as such is less relevant here (Saleem, Dillon, Benesch, & Ruths, 2017). Silva et al. study different targets of 'serious' (i.e. overt and aggressive) hate speech on Twitter and Whisper across nine categories, including Disability, Race and Sexual orientation (Silva, Mondal, Correa, Benevenuto, & Weber, 2016). Their goal is not to classify social media posts at scale in 'the wild' but to describe the content of posts and such is also less relevant to the current work.

Second, are studies which distinguish between different strengths of prejudice within one domain. Focusing on just one domain is important for enabling more nuanced forms of analysis, for as Saleem et al note, 'hateful speech classification systems require target-relevant training' (Saleem et al., 2017, p. 7). This task is harder than the first one as there is less variation between classes when classifying based on strength rather than target. Burnap and Williams train a classifier to distinguish between different levels of cyberhate (divided into 'moderate' and 'extreme' classes) targeted against Black Minority Ethnic and religious groups on Twitter (Williams & Burnap, 2016), achieving precision of 0.77. Malmasi and Zampieri distinguish between 'Hate' speech, 'Offensive' speech and 'Ok' speech. They achieve 78% accuracy but on an unevenly weighted training/testing dataset – over half of their corpus is 'OK'. Their model struggles to distinguish between non-OK content; of 2,399 'Hateful' instances in their dataset, 1,050 are categorised correctly, 1,113 are miscategorised as 'Offensive' and 236 as 'OK'. They also do not test their model on unseen data, only reporting the results of cross-validation (Malmasi & Zampieri, 2017). Jha and Mahmidi distinguish between 'benevolent' and 'hostile' sexism. They use Waseem and Hovy's dataset of 16,000 tweets as well as ~7,000 newly

collected ones (Waseem & Hovy, 2016). Using SVM they report an F1 score of 0.80 for Benevolent tweets, 0.48 for Hostile and 0.89 for Others. As their data is imbalanced towards Others, overall performance is strong.

Kumar et al. (Kumar et al., 2018) distinguish between overtly aggressive, covertly aggressive and non-aggressive tweets using a dataset of 15,000 Facebook posts. In a competition entered by 130 teams (of which 20 completed it and provided the technical details of their model), the highest performing obtained a weighted F-score of 0.64. As the authors note, 'the results […] depict how challenging the task is' (Kumar et al., 2018, p. 1). Davidson et al. train a model to distinguish between hate speech and offensive speech, and non-offensive speech in tweets. They report impressive results, with precision of 0.91, recall of 0.90 and an F1 score of 0.90. Their work demonstrates the potential for multiclass classification, makes an important theoretical argument apropos the need to separate different types of content, and introduces the use of 'Ease of Reading' metrics as an input feature. However, as they note, their model performs poorly with hate speech, of which almost 40% is misclassified. The high F1 score is largely due to the fact that their classes are very uneven (76% of the data is in the 'offensive speech' category). They also train and test their classifier on a single dataset, which could risk overfitting.

Previous research to classify content suffers from three considerable limitations. First, most studies are not focused specifically on Islamophobia but focus on other, overlapping, forms of hate (Schmidt & Weigand, 2017). Second, accuracy and precision in machine learning classification is typically quite low, which limits how useful they are for empirical research – as a heuristic, van Rijsbergen recommends a minimum precision of 0.7 for classifiers used in empirical studies (van Rijsbergen, 1979). Third, most studies pay insufficient attention to social scientific insights about Islamophobia.

Overall, measuring Islamophobic hate speech is a considerable challenge. A newly created single classifier is appropriate given the scale of the data collected, the need for a robust and replicable method, and the fact that the data is reasonably specific in terms of the platform studied (only Twitter), the time period over which the data was collected (limited to approximately one year) and the narrowness of the context (only followers of UK political parties). This classifier will be useful not only for the present work but also future studies.

3.3 | Data

The empirical focus of this thesis is Twitter. Twitter is a microblogging social media platform which allows users to instantaneously post messages (called 'tweets') and to interact with other users and their tweets. Users can subscribe to 'follow' other users in order to see their tweets in a live feed. Following is one-directional and does not have to be reciprocated. The accounts that a user follows are known as their 'friends'. A key part of Twitter's appeal is that users can receive tweets from a wide variety of sources, including celebrities, newspapers and politicians, because following other users is very easy. Twitter is an important platform to study due to its high-profile role in contemporary politics (Cihon & Yasseri, 2016; Gerbaudo, 2012; Margetts, 2017a), evidence that many of its users engage in Islamophobic behaviour (Awan & Zempi, 2016; Burnap & Williams, 2015; DEMOS, 2017), and its large overall user base as one of the top five most well-used social media platforms (Forbes 2017). Previous studies also indicate that Twitter plays an important 'disseminator' role, spreading niche extremist content from less well-used platforms such as Reddit and 4chan (Ludemann, 2018; Zannettou et al., 2017).

Twitter has recently been heavily scrutinised for its use by extremists, including far right ideologues. In a report on online hate crime, the Home Affairs Select Committee criticised Twitter for not removing and banning harmful Islamophobic tweets and users with explicit neo-Nazi sympathies (Home Affairs Select Committee, 2017). Twitter has also been held responsible for offline Islamophobic actions; in sentencing the perpetrator of the Finsbury Park mosque terrorist attack, the Judge stated, 'Your use of Twitter exposed you to racists and anti-Islamic ideology […] you allowed your mind to be poisoned[.]' (BBC, 2018b). Concerns that Twitter is a breeding ground for Islamophobes and extremists are supported by evidence that many white supremacist, neo-Nazi and far

right groups have considerable influence on Twitter (Berger, Strathearn, & Meleagrou-Hitchens, 2013; Hope Not Hate, 2017). Overall, Twitter is a highly suitable source of data for realising the research aims of this thesis.

3.3.1 | Data collection process

Data is collected and wrangled using scripts written in R and Python. Python is used primarily for the data collection because it scales easily, is well-supported on Linux servers and has many dedicated libraries for online data collection, such as 'requests', 'urllib' and 'auth'. R is used for the data wrangling because, whilst not as efficient as Python, it has many well-supported packages for data cleaning and analysis, such as 'data.table', 'dplyr' and 'ggplot2'.

Data is collected through both Representational State Transfer (REST) and Stream Application Programming Interfaces (APIs) provided by Twitter. Both APIs follow standard HTTP Request protocol. With REST researchers send requests to the API which are acknowledged and either fulfilled (i.e. the data is returned) or rejected. There are restrictions on the amount of data provided in each request and the frequency of requests. For instance, researchers can only collect up to 3,200 historical tweets from each users' timeline, calls are split into batches of 200 tweets per time and calls are rate limited per hour. The REST API is the most effective way of collecting tweets for specific users and for collecting a list of their followers and friends. The Stream differs from REST in that requests are not sent in batches but, instead, a constant connection is opened with the API. Any results which match the search parameters are automatically returned. This is the most effective way of collecting tweets in real-time which match certain keyword criteria, such as user names and hashtags. For more information, see Twitter's documentation (Twitter, 2018). In this thesis, I use the REST API.

The data collection and wrangling pipeline for both network and timelines data is shown in Figure 1. First, data is collected from a Linux server using a Python script based on Hale's previous work (Hale, 2014). The data is then transferred to a MacBook Pro where it is converted from a JSON file into a comma separated file. The data is then cleaned using a script in R and stored efficiently as an RData file. Preliminary analysis is then undertaken in R to extract the most recent ID values. The REST API for collecting users' tweets uses 'pagination', which allows researchers to only collect new tweets from Twitter, avoiding duplicates. Pagination only works if the most recent ID values have been extracted from each users' tweets. Scripts are available online at, `https://github.com/bvidgen`. All data is stored securely on two encrypted external hard drives, which ensures there are at least two records of every data file.

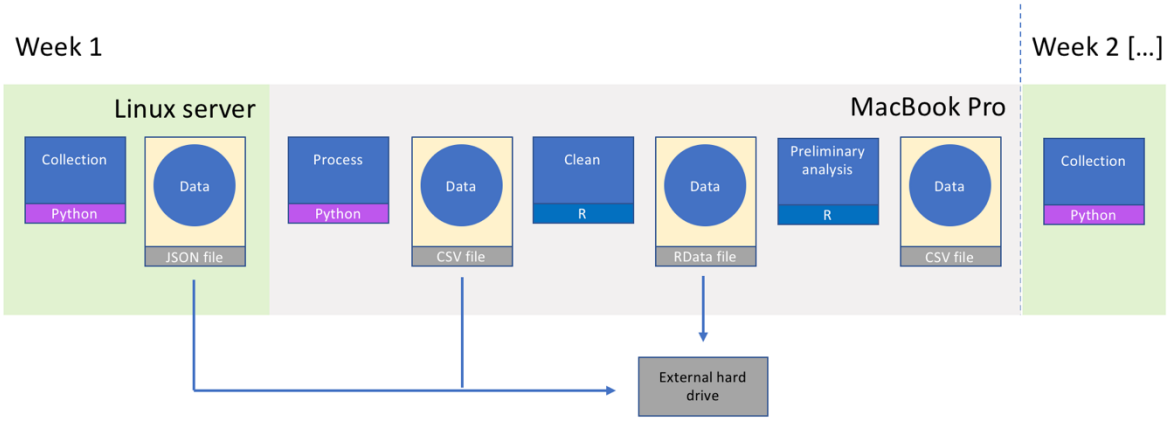


Figure 1, Data collection and wrangling pipeline

3.3.2 | Data collection frequency

Collecting data frequently poses considerable logistical and financial costs, and as such there are benefits of collecting data infrequently. Most users tweet rarely or not at all; research by data monitoring firm 'Twopcharts' found that 44% of Twitter users had never tweeted and that 30% had sent fewer than 10 tweets (Wall Street Journal, 2014). Accordingly, for most users, it would be sufficient to collect all of their tweets just once

per year (as the REST API allows for the last 3,200 tweets to be collected). However, whilst this would capture most *users* it would lead to very poor coverage of the total *volume* of tweets. Alongside the large number of 'lurking' users, there is also small number of 'power users' who tweet frequently and are responsible for a large proportion of tweets – such as the bot account @VENETHIS which had sent over 37 million tweets by March 2016 (Five Thirty Eight, 2016). The large discrepancy in how much users tweet, combined with the large number of users in this study, poses considerable problems for data collection as whilst most users require their data to be collected only infrequent (such as once per month or even just once per year), a small number of users need very regular data collection (potentially, daily).

To ensure that the data collection method balances pragmatic feasibility with coverage, I run a test study in which different frequencies of tweet collection are evaluated. For a two-week period, from Thursday 9th August to Thursday 23rd August 2018, I collect tweets for 1,000 users on both a daily and weekly basis, in both cases using Twitter's pagination function. The users are sampled randomly from the followers of the BNP's account (@bnp) on Wednesday 8th August 2018. Out of 1,000 sampled users 657 users tweet during the two-week period (the remaining 343 either do not tweet or have their accounts set to private). All users appear in both datasets. However, there is a discrepancy of 1,828 tweets (4.95%) between the two datasets; the daily collection method collects 38,739 tweets whilst the weekly collection method collects only 36,911. For 85.7% of users both data collection methods return the same number of tweets, for a further 12.6% the discrepancy is less than 5% and for just 1.7% is the discrepancy greater than 5%. Whilst the weekly data collection method collects tweets for all users – and in the vast majority of cases, collects all of their tweets – for some high-volume users the weekly method collects considerably fewer tweets.

To further validate the relative merits of the data collection methods (weekly and daily), I check their impact on (i) the daily volume of tweets and (ii) the prevalence of bots (which are automated or semi-automated accounts which could bias the results (Kollanyi, Howard, & Woolley, 2016) – they are discussed in more detail in subsequent chapters). The results indicate that the daily data collection frequency consistently outperforms the weekly frequency. However, crucially, the overall trend of the volume of tweets is the same for both methods. Noticeably, users who are under-recorded using the weekly method as opposed to daily method are more likely to be bots. This suggests that the under-recording of the weekly method is more likely to affect bots than regular users. These results show that daily data collection outperforms weekly. However, the differences are small and are unlikely to materially affect the analysis. As such, given the logistical and financial constraints of the present work, I opt for weekly data collection. Full details of the testing, including further analysis and visualizations, are reported in Appendix 3.1.

A further challenge in collecting data from Twitter on Islamophobia, or any other form of hateful and harmful behaviour, is that Twitter's policies prohibit certain forms of hate speech (as mentioned in the Literature review). Depending on the efficacy of their reporting mechanisms, which are currently required by an EU directive to investigate extremist content within 24 hours, this might bias data collection as strong Islamophobic content is likely to be removed. There have been several noticeable policy changes in how Twitter moderates hateful content. In December 2015 Twitter introduced a ban against hateful content for the first time, establishing rules on what users cannot post which ensured that flagged hateful posts would be deleted. In August 2016, they introduced a 'Quality Filter' which is turned on by default for all users. It covers hate speech as well as spam and automated content and is not applied to content from people

who users follow or interact with. In effect, it removes notifications about low quality content and removes this content from users' timelines. In February 2017, they introduced a feature to 'Hide Sensitive Content' which both allows users to flag their own profile as 'sensitive' (and has been widely used by accounts sharing pornographic and other sensitive material) and allows them to avoid such content from their feed (Twitter, 2019). If it does appear, it comes with a sensitive content warning. This is implemented at a user level, which means that even non-sensitive content produced by accounts marked sensitive have the warning. Fortunately, this change was introduced prior to the start of data collection in this thesis (March 2017) and as such has not introduced an additional bias during the collection period.

These content moderation processes can have a considerable impact on the type of content which is permitted on Twitter. In particular, it is likely to affect the prevalence of the most serious types of overt and targeted Islamophobic hate. Some of the algorithmic effects of these moderation strategies, such as the 'Hide Sensitive Content' feature, are hard to quantify as they may have uneven and undisclosed impact on users' media environments. An additional challenge here is that several Twitter accounts (such as those using the #FarRightWatch hashtag) report Islamophobic content, as do several dedicated charities, such as Tell Mama. To account for how Twitter's content moderation policies and activist user activity may affect data availability, and in particular its impact on strong Islamophobic content, I analyse the prevalence of Islamophobia in both the daily and weekly collection methods. The initial results show that the prevalence of Islamophobia is very similar across both of the methods; approximately, 90% of values are none, 8% are weak and 2% are strong. I also investigate whether the collection of tweets with different strengths of Islamophobia is affected by the timing of the weekly data collection method (i.e. whether the fact that it usually occurs on a Sunday impacts data collection

coverage over the subsequent days). The results show that the recording of Islamophobia on each day is very close across the two methods, with no systematic deviations in which days have the best coverage. This further validates the use of a weekly collection method. This is reported in more detail in Appendix 3.1.

3.3.3 | Social media followers

In the literature review (Chapter 2), I argued that social media followers should be seen as a constituent part of contemporary political parties. To validate this argument, I conduct a small test study on the followers of the BNP, UKIP, Conservatives and Labour. The findings are relevant for all of the empirical chapters, which all involve studying the behaviour of followers of political parties.

For each party, I sample 200 followers and collect all of the tweets they produce from 1st April 2017 to 1st April 2018 (n = 694,456). In cases where users produced more than 100 tweets, I randomly sample just 100 to minimize the impact of high volume tweeters and bots. This reduces the dataset to 52,673 tweets. I then use a keyword search to identify tweets which contain references to the four main UK political parties (Conservatives, Labour, Liberal Democrats and UKIP) as well as the BNP. The keywords consist of the name of the party, any highly relevant abbreviations and the parties' top politicians. This reduces the dataset to 2,541 tweets. This is shown in Table 1.

Party	Number of tweets produced by the sample of 200 followers	Number of tweets after random sampling	Number of tweets after filtering
Conservatives	220,106	12,937	641
Labour	104,071	14,070	850
UKIP	104,544	12,256	598
The BNP	265,735	13,410	452
TOTAL	694,456	52,673	2,541

Table 1, Number of tweets for followers of each party after filtering

I manually annotate the dataset of 2,541 tweets for whether they contain expressions of support or opposition to any major UK political party (defined as any party with representation at any level in the UK). Only literal expressions of support or opposition

for each party, including their flagship policies and activities, are considered. This reasonably strict annotation criteria ensure that the analysis remains focused on parties, rather than political ideologies (or any *a priori* view about the beliefs that followers of certain parties have), and as such minimizes the number of false negatives.

This study produces three main preliminary findings, which provide preliminary insight into the role of social media followership within political parties – although due care should be taken when interpreting the results given the small size of the study and the simple methodology. First, most of the followers do not engage in party political talk. Of the 800 users in this study, 566 are not recorded as expressing explicit negativity or positivity towards any political parties. Second, when all of the different points of discussion are considered across all of the tweets, most political party talk is negative rather than positive. Of the 2,541 tweets I annotated, 1,731 tweets express either positivity or negativity towards a political party (68%). The reason why so many tweets express negativity/positivity but so few users do so is that users who express negativity/positivity tend to tweet in far larger volumes. Of the 1,731 tweets which express negativity/positivity, 1,178 express negativity (69%).

Third, 96.25% of users do not send any tweets which express negativity against the party they follow. This means that even though negativity is far more prevalent that positivity, it is nearly always directed against another party rather than the one which the user follows. Initial qualitative analysis of the tweets also suggests that a small amount of negativity directed towards a party does not necessarily indicate that the user is opposed to it. Particularly when situated in the context of other statements of support, such criticisms may merely be indicative of an agonistic and reflexive relationship with the party. This is reasonably consistent across all four parties studied, and the breakdown for each party is reported in Table 2. Note that of the users who send at least one tweet which

expresses some negativity against the party they follow (3.75% of users), some of the negativity is quite extreme and expressed in several tweets.

These preliminary results in a small-scale study provide initial evidence that social media followership of a political party can be viewed as an act of interest and possible affiliation. Social media followership likely indicates a degree of affinity with a party, even if it is of a passive nature, as shown by the lack of negative comments towards that party and prevalence of positive comments. At the same time, a small number of followers of a party should be considered active opponents. This reflects the ambiguity of social media followership and the difficulty of understanding what precisely a 'follow' means; it lacks the certainty and reliability of traditional offline acts of political participation, such as voting for a party of joining a party as a member. Nonetheless, overall, these results suggest that social media followers are usually not opposed to the party they follow and more often than not are supporters. This is a form of confirmatory falsification: the evidence does not challenge the idea that social media followers a constitutive part of modern political parties, as argued in the literature review, and so can be viewed as initial evidence. Further research is required to validate this more robustly, such as in-depth qualitative interviews. In particular, the political importance and organisational function of social media followership needs investigating to ascertain the significance of these actors within party organisation. For the purposes of this thesis, these results justify the research design and, as such, I study the behaviour of political party followers in the empirical chapters.

Party	Percentage of followers who do not express negativity against the party
Conservatives	96%
Labour	96.5%
UKIP	97%
The BNP	95.5%
TOTAL	96.25%

Table 2, Percentage of users who do not express negativity against the party they follow

3.4 | Ethics

Ethical approval has been given for this thesis by the Oxford Internet Institute's Departmental Research Ethics Committee. The reference number is SSH_OII_C1A_16_073.

Ethics are a key concern for researchers studying online spaces such as social media. Recently, high-profile new stories, such as the infamous emotional contagion study on Facebook (Jouhki, Lauk, Penttinen, Sormanen, & Uskali, 2016) and the Cambridge Analytica election scandal (BBC, 2018c), have increased public concern about how social media data is used and the power of algorithmic analyses. This has increased the onus on researchers to engage in ethically-aware research which engenders public goodwill. As Zimmer argues, whilst using social media data has created new opportunities for social research, 'it is our responsibility to ensure our research methods and processes remain rooted in long-standing ethical practices.' (Zimmer, 2010, p. 324) Others argue that consideration of ethical issues has not kept pace with the explosive growth in social media research (Ahmed, Bath, & Demartini, 2017). Williams et al. note that even though Twitter has become part of the 'sociologists data diet' substantial ethical mistakes are often made, such as releasing highly sensitive content without obtaining users' consent or inadequately anonymizing the data (Williams & Burnap, 2017). This is supported by empirical analyses. In a review of 382 papers from 2012 which used Twitter data Zimmer and Proferes found that only 4% made any mention of ethical issues (Zimmer & Proferes, 2014).

At present, there is a lack of consensus as to what ethical norms and practices should operate within the field of social media research (Fiesler & Proferes, 2018; Vitak, Shilton, & Ashktorab, 2016; Williams & Burnap, 2017). Ethical guidelines have been established by relevant bodies, such as the Association of Internet Researchers' 'Ethical

Recommendations' (Markham & Buchanan, 2012) and the British Sociological Association's 'Ethics Guidelines for Digital Research' (British Sociological Association, 2017). However, these provide only very general principles for Internet based research rather than recommendations for the practicalities of studying social media specifically. Legal, institutional and platform-specific ethical codes are another source of guidance, yet these typically constitute only a 'minimal' set of ethical obligations and may not be appropriate for all types of research. In the remainder of this section, in line with the comprehensive ethics literature review undertaken by Townsend and Wallace (Townsend & Wallace, 2016), I explore four key issues: (1) public/private datasets, (2) consent, (3) anonymization and privacy, and (4) harm. It should be noted that, ultimately, maintaining ethical integrity requires a researcher that is ethic*al*, and that any policies and processes are implemented with due care and consideration.

3.4.1 | Public and private data

Twitter is often seen as a public resource by researchers (Gagliardone, Gal, Alves, & Martinez, 2015, p. 14), and as such users' explicit consent is not required for their data to be used. This view is justified primarily on the grounds that data sharing with third parties is stated explicitly in Twitter's Privacy Policy: that users 'consent' when they sign up to the platform, and that users have the option to set their account to 'private'. The privacy setting makes all of a user's personal metadata, tweets and friends/follower lists unavailable to third parties, including researchers (Twitter, 2018). However, previous research indicates that most Twitter users do not fully understand the extent to which their behaviour on the platform is recorded, monitored and monetized. Evans et al. show that most users are unaware that Twitter makes their data available to academics and developers, and that the platform generates revenue by selling access to users' content (Evans, Ginnis, & Bartlett, 2015; Proferes, 2017). Prior research points to other problems

with how users view social media spaces. On Twitter, not all users are aware of how their tweets can spread across and beyond the Internet and misconceive both the character and extent of their true audience (Marwick & boyd, 2011, 2014). These problems are more acute in relation to 'everyday' users rather than prominent public figures, who have a heightened awareness of the scrutiny they will likely be under.

Categorising Twitter as either public or private creates considerable theoretical tensions. Although users' posts are usually publicly available, most treat Twitter as though it were a more exclusive and private space. Boyd and Crawford note that 'just because content is publicly accessible does not mean that it was meant to be consumed by just anyone.' (boyd & Crawford, 2012, p. 672) This reflects a longstanding issue in studies of public discourse, as articulated by Hammersley and Atkinson in 1995; 'What is public and what is private is rarely clear-cut. Is the talk among people in a public bar public or private? Does it make a difference if it is loud or *sotto voce*?' (from (Hammersley & Atkinson, 1995) quoted in (Williams & Burnap, 2017)). Thus, the public/private dichotomy may itself be unhelpful. Twitter is best viewed as a hybrid space, in which offline ethical norms are relevant but cannot be applied straightforwardly. For this reason, sharing Twitter datasets is ethically problematic as the data is 'private' as much as it is 'public'. None of the datasets from Twitter used in the current work will be made publicly available.[3]

3.4.2 | Consent

Consent is arguably the ethical linchpin of research involving human subjects, as set out in the Belmont Report (National Commission, 1979). Ascertaining informed consent

[3] Twitter's Terms of Service currently allow researchers to share the ID strings of up to 10,000 tweets from each day.

from participants is important because it provides a formal mechanism to realise 'respect for persons'; participants should only be included in a study if they *choose* to do so. However, consent can take many forms, which require different levels of engagement, information and consideration – and not all users of technology have the same expectations regarding consent (Martin & Shilton, 2016). Specifically, research shows that Twitter users do not have a unitary view regarding the importance of consent. In a study of users' attitudes towards consent on Twitter, Fiesler and Proferes found that users are less concerned by research which uses computational methods of analysis, is undertaken by academics and uses large quantities of data (Fiesler & Proferes, 2018). Other research also shows that users are less concerned to give consent if their data is fully anonymized and only used in aggregate analyses (Evans et al., 2015). These findings align well with this thesis' large-scale and computational research design.

Most ethical guidelines adopt an individualist ontology, viewing each study participant separately. This has been challenged by recent ethical work developed in response to the advent of big data. Buchanan argues that many analyses which involve large numbers of participants treat those participants as a *collective* data subject rather than as a series of individual participants (Buchanan, 2017). As such, individual consent for a *particular* study is not an ethical requirement, provided that users have given consent *in general* for their data to be used by third parties, such as academic researchers. Thus, in-line with existing research practices, explicit consent is not sought from Twitter users whose data is used in the present work.

3.4.3 | Anonymization and privacy

Anonymization and privacy are important considerations for protecting users whose data is used in studies of social media, particularly when explicit consent is not sought from

each individual. These issues are particularly pressing in the current work, given its contentious subject matter (Islamophobia). Skopek distinguishes between anonymization and privacy on the basis that, 'under the condition of privacy, we have knowledge of a person's identity, but not of an associated personal fact, whereas under the condition of anonymity, we have knowledge of a personal fact, but not of the associated person's identity. In this sense, privacy and anonymity are flip sides of each other.' (Skopek, 2014, p. 1755) Skopek's work, as well as others who have built on his anonymity/privacy distinction (Daries et al., 2014), suggests that researchers must opt to foreground either anonymization or privacy when seeking to balance protecting participants' wellbeing with conducting robust science. For the present work, anonymization is the most important consideration given that (1) the analysis focuses on users' content, and therefore requires 'knowledge of a personal fact' (making it very hard to maintain privacy – although this is still maximised wherever possible) and (2) given the subject matter, there is a risk that non-anonymized users could be targeted (or 'trolled') by so-called 'anti-Fascist' activists (Coles & West, 2016).

Standards of anonymization vary, and in some cases supposedly anonymized datasets have been hacked and the users identified. In a now infamous 2008 study, researchers shared a supposedly anonymized longitudinal dataset of students' Facebook relationships at a USA university (Lewis et al., 2008). Within days, third party researchers had identified the university and, more worryingly, some of the individual participants. This was possible because participants' nationality was recorded in the dataset, even though for some nationalities there was only one person. This study – as well as other subsequent research (Daries et al., 2014) – shows how third parties can use anecdotal information, identifying traits (such as nationality or network position) and even additional datasets to identify participants (Zimmer, 2010). Thus, anonymization, whilst an important tool in

protecting user privacy, has to be implemented with due care to achieve its purpose. In the present work, to ensure that user's anonymity is maximally protected, user-level information is not just anonymized but also presented in aggregate. No individual users' data (including network data or metadata) is presented. These choices also have the added benefit that users' privacy is maximised too.

A further issue relating to anonymity is how the content of users' specific tweets should be reported. This is a crucial artefact of the research, particularly for the qualitative analyses (primarily in Chapter 4). However, this is ethically problematic as the tweets could be used to identify users, jeopardizing their anonymity. Fiesler and Proferes advise 'not quoting tweets verbatim without reason' (Fiesler & Proferes, 2018, p. 10), which they base on Bruckman's argument regarding the need to 'disguise' content produced by online users in published research (Bruckman, 2002). Williams et al. suggest that when tweets are directly quoted, explicit consent should be sought from participants unless the content is not sensitive and opt-out consent has been sought within a reasonable time frame (Williams & Burnap, 2017). Given the sensitive nature of the present research, it is highly unlikely that consent would be forthcoming. Another option is to present synthetic tweets, which typically are constructed by amalgamating and synthesizing several real tweets (Townsend & Wallace, 2016).

Much of the analysis in the present work is computational. In these parts of the thesis (namely, Chapters 5, 6 and 7), real tweets are not reported. Instead, only the results of analysing the tweets are shown, such as summaries about the overall prevalence of certain types of Islamophobia. Chapter 4 consists of detailed qualitative analysis of tweets. In this chapter, it is crucial to the nature of the work that the tweets are discussed in detail. As recommended by Fiesler and Proferes, they are not quoted verbatim but, instead, are (i) cleaned extensively and (ii) amended before being reported. All punctuation, links,

emojis, @ mentions of accounts and unusual spellings are removed, spelling and grammar corrected, and any excessively offensive language (including swear words and insults) has been either removed or redacted where possible. The original meaning of the tweets is retained but the distinctive features are removed, thereby making it difficult for other researchers to subsequently identify them (and, by extension, their authors). In cases where their content is particularly sensitive or easily identified, tweets are amalgamated – akin to the synthetic approach to anonymization, outlined above. The choices outlined here should ensure that no users are identifiable from the analysis and presentation of results.

3.4.4 | Harm

Preventing or limiting harm is a key ethical obligation of researchers. It has been described as 'the Golden Rule' of both online and offline ethical research (Christina Allen, 1996). Participants can experience various forms of harm, including not only physical but also emotional and social forms. A key point at which participants experience harm is when the research is implemented. This is a real concern with research in this area – for instance, a recent study by Munger on how bots can be used to reduce prejudice on Twitter directly led to some users producing racist tweets (Munger, 2017). In contrast, the research design of this thesis is primarily observational, and no facets of the Twitter environment or users' experience are manipulated. Accordingly, the likelihood that the data collection process will itself directly cause users harm is very low. Similarly, because all results are presented in aggregate it is also unlikely that the data reporting will cause participants harm.

Harm can also be inflicted at a societal level. This is likely to manifest in two ways. First, is that the study participants – or people ideologically akin to them – may use the findings

of the research to become more effective Islamophobes. This is a worrying concern but one that is unlikely to occur as Islamophobes are unlikely to find and make use of the present research. Second, is that the research may be used by the government, large tech firms or civil society bodies to monitor or manipulate users on Twitter who are engaging in behaviours that might be undesirable but are nonetheless legal. Concerns about the uses of behavioural research have been considerable since Facebook's infamous 'emotional contagion' experiment, discussed above (Kramer et al., 2014, p. 8788) Concerns about social manipulation are important but in this case are outweighed by (1) the greater harms inflicted by Islamophobic behaviour and need for it to be both better understood and challenged (issues which arguably place an onus on academics to conduct research in this area), and (2) the fact that social interventions are not a direct outcome of the present work.

There is a risk of harm being imposed on the author of the present work through exposure to harmful and dangerous content. This is somewhat mitigated by the fact that the author is not a primary target of Islamophobic hate. Furthermore, the overall risk is low because a large portion of the work is computational. For the qualitative analyses, appropriate support and guidance has been sought from the University of Oxford's Social Sciences Division. A further issue is how the author should respond to any Islamophobic hate speech that is observed. Although the researcher is not responsible for this content, one could be considered complicit in its production and dispersion, which could inflict harm on individuals outside of this research. Accordingly, the author has engaged with the anti-Islamophobia reporting charity, *Tell MAMA* throughout the implementation of the present work – this is discussed in further detail in the conclusion.

3.5 | Conclusion

This Chapter discussed in detail the key methodological approaches and choices that have been undertaken in completing the present work. In arguing for a complementary form of computational social science, I have pointed to the multi-methodological nature of the research, and the need for computational analyses to be informed by qualitative work. Noticeably, the choice of methods is driven by the nature of the task at hand the function of each chapter in the overall thesis design; different quali-quantitative methods are used to construct a cohesive thesis rather than to validate or verify a single finding, as with triangulation. It is hoped that this approach will ensure that the computational work maintains both breadth and rigour alongside nuance and attention to detail. The discussion of ethics in Section 3.4 points to the complexity of studying phenomena on social media, the relative newness of this area, and the need for researchers to be attuned to the fast-changing ethical environment. In balancing ethical requirements with users' right to privacy and anonymity, I place greatest emphasis on maximizing anonymity. It is not anticipated that the research will cause harm to the participants, society as a whole or the researcher.

This Chapter highlights the central role of the research*er* when conducting research. Noticeably, the creation of a fully integrated complementary computational social science project depends upon the researcher having adequate experience of not only several different methods but also different approaches to research design and problem identification. This is unlikely to be an issue for large multi-member research projects where different researchers can contribute different skills but could be a considerable barrier for many researchers embarking on single-authored projects. In this regard, the present work utilizes a somewhat unorthodox research design for conducting thesis research – which, whilst it should lead to more robust results overall, is also more complex and time consuming. Similarly, the discussion of ethical considerations

highlights the centrality of the researcher. Ethics cannot be reduced to a set of simplistic injunctions but requires the researcher to constantly, and iteratively, negotiate between the research goals, the specific research site, consideration of societal impact and relevant ethical principles. This requires rectitude and trustworthiness as much as it requires a comprehensive literature review. With regards to both these aspects of the present work (i.e. the research design and ethical integrity), Darie et al. summarize the key point succinctly; 'If we want to have high-quality social science research […] we must eventually have trust in researchers' (Daries et al., 2014, p. 63).

Chapter 4 | What is Islamophobia? An investigation into Islamophobic hate speech on Twitter

In recent times, Islamophobia has received considerable attention from academics, government bodies, civil society groups, large tech companies and the media (Chris Allen, 2017; BBC, 2017; HM Government, 2012; Ingham-Barrow, 2018; Tell Mama, 2016, 2017). Indeed, reflecting on the growing volume of research produced each year as well as its theoretical sophistication, Klug suggests that the study of Islamophobia has finally 'come of age' (Klug, 2012). Nonetheless, despite the advances that have been made, numerous disagreements remain. One of the most striking aspects is the continued terminological confusion apropos the conceptual basis of Islamophobia. A multitude of competing – and often quite radically different – definitions of Islamophobia are used in existing research. As such, in this Chapter I answer the first research question from the literature review:

RQ 1: What is the conceptual basis of Islamophobia?

Two big risks are posed for empirical analyses by a poorly specified definition of Islamophobia: either (i) certain forms of Islamophobia, such as 'small' or 'micro' actions, will be missed or (ii) non-prejudicial behaviours will be mistakenly included through 'conceptual stretching'. This is where concepts are distorted out of shape to fit empirical phenomena, even if they are not well-suited to capturing them (Sartori, 1970). This is problematic as it is only with 'stable concepts and a shared understanding of categories' that we can produce robust and generalisable research and develop theoretical knowledge (Collier & Mahon, 1993). Thus, without a robust definition of Islamophobia, there is a risk of an abundance of either false positives or false negatives when it is used in

empirical research, either of which could invalidate findings. As Goodhart argues apropos racism, 'we need a nuanced language […] when almost everyone is racist, no one is.' (Goodhart, 2014, p. 251). Defining Islamophobia also serves an important political function in enabling Muslims to challenge and counter Islamophobia; a report by the charity MEND describes it as 'an act of recognition […] It officially validates [Muslim's] experiences as undeniable facts in need of address.' (Ingham-Barrow, 2018, p. 10)

As discussed in the literature review, UK Law is disjointed and it does not provide an adequate conceptual starting point for understanding Islamophobia. Legal protections are balanced against the need for freedom of expression, which means that inevitably the bar for Islamophobia is set high (Williams 2019). More broadly, the need for the law to be translated into a set of guidelines which can lead to prosecutions means that it highly detailed, and focused towards understanding and labelling very specific behaviours. It is not, therefore, necessarily suitable for understanding the conceptual basis of Islamophobia; this is not necessarily important for its stated purpose of defining what behaviours are and are not permissible. Furthermore, the common law tradition means that considerable attention is paid to translating the (relatively newly recognised) issue of Islamophobia into existing legal frameworks rather than considering its core conceptual meaning. Thus, whilst the Law has been a useful starting point for this chapter, and some of the publications reviewed are from Legal research, it does not resolve the overarching problem of *what is Islamophobia?*

The goal of this Chapter is to investigate the conceptual basis of Islamophobia. This is achieved through recursively examining two datasets. First, a corpus of academic articles pertaining to Islamophobia. Second, a corpus of 12,000 tweets produced by far right Twitter accounts. As a result, the findings are constrained to the three main contexts of

this data: first, social media (specifically, Twitter). Second, the far right. Third, Islamophobic *language*. Nonetheless, the arguments are relevant for understanding other forms of Islamophobia, and in other contexts, particularly Islamophobic speech offline. As such, the discussion is intentionally posed at a general level.

Both of the data corpuses are analysed using 'close reading'. This is a qualitative approach which involves manually reading the data, critically reflecting on its content and identifying the most important themes (Jänicke, Franzini, Cheema, & Scheuermann, 2015; Wesley, 2010). Close reading is an inherently subjective endeavour and as such is well-suited for gaining nuanced deep insight into data – rather than reproducible measurable findings (Grimmer & Stewart, 2013). As such, it is well-suited to the present work, the goal of which is to conceptualise and clarify what is meant by the term 'Islamophobia', rather than to measure prevalence. The two corpuses are analysed in tandem; the corpus of tweets is used to probe the conceptualisations of Islamophobia in the articles, which are, in turn, used to better understand how Islamophobia manifests in the tweets. The work is also informed by three years of in-depth study of far right users on Twitter. None of the analysis is computational.

In this Chapter, I make four contributions. First, I show that in the data hate speech is directed against both Islam and Muslims. I use this to argue that, conceptually, both should be considered forms of Islamophobia. Second, from the corpus of academic articles, I construct a typology of six distinct conceptualisations of Islamophobia, namely: 'fear', 'threat', 'stereotypes', 'difference', 'dominance' and 'negativity'. Third, I critically investigate these conceptualizations by examining Islamophobic hate speech in the corpus of tweets. I find that none of the conceptualisations are, on their own, sufficient for understanding the full nature of Islamophobia. I also identify new insights into the empirical nature of Islamophobic hate speech. Noticeably, I find that the literal

interpretation of Islamophobia (that it pertains to *fear of* Muslims) is misplaced, as few tweets express fear of Muslims – although many can be considered fear-*inducing*. I then use these analyses to put forward a definition of Islamophobia based on two dimensions: (1) negativity and (2) generality. These two dimensions provide a robust way of capturing what is at stake with Islamophobia conceptually rather than merely *describing* how it manifests. I tie this to Islamophobic speech, providing a definition of Islamophobic hate speech which can be used in the empirical parts of this thesis. Fourth, I use these two axes to identify qualitatively different varieties of Islamophobic hate speech, namely 'weak' and 'strong'. This final contribution marks an important step forward in the study of Islamophobic hate speech as few previous studies have explicitly sought to systematize its different modalities.

4.1 | Data

Two datasets are used in this Chapter. First, a corpus of 100 academic articles. These were identified using keyword searches in Scopus, Web of Science and Google Scholar, implemented in June 2018. These three bibliometric databases were used as together they have the greatest coverage of articles, conference papers and books, and have been widely used in previous research (Franceschini & Maisano, 2016; Harzing & Alakangas, 2016; Mongeon, 2016). Articles were identified by using the Boolean search query, 'Islamophobia OR "anti-Muslim" OR "anti-Islam"' (using wildcard variations) for keywords, publications and titles in Scopus and Web of Science, and for keywords alone in Google Scholar. The search criteria are narrow which ensures that most articles in the corpus are highly relevant. Scopus and Web of Science returns 1,550 unique articles and, using the software, 'Publish or Perish', Google Scholar returns approximately 52,000 (Harzing & Alakangas, 2016; Harzing & Adams, 2009). However, many of the Google Scholar articles are highly irrelevant and, as such only the top 500 are included in the corpus (of which 328 are not duplicates). The size of the remaining corpus (1,878 articles) is too large to be analysed by a single researcher and, accordingly, is sampled to a feasible number (n = 100). Sampling academic literature so that the most relevant, up-to-date and impactful articles are retained is a considerable challenge (Mortenson & Vidgen, 2016; Tsvetkova et al., 2015, p. i). I use a qualitative methodology, and retain articles based on their impact (the number of citations), relevance (keywords), authorship (the number of publications) and newness (the publication date).

The second dataset consists of a corpus of 12,000 tweets from far right users. The far right is chosen for sampling because previous research indicates that they are likely to engage in Islamophobic behaviour (Allen, 2011; Awan & Zempi, 2016). 10,000 tweets are sampled from a set of 109,488 tweets and retweets sent by a set of 45 far right Twitter

accounts, which were identified by the charity Hope Not Hate in their 2015 and 2017 annual reports (Hope Not Hate, 2015, 2017). The tweets are stratified before sampling so that a similar number of tweets are collected from each user and from each month. An additional 2,000 tweets are sampled from a dataset of 5.5 million tweets sent by followers of the BNP from April 2017 to April 2018 (n = 5,510,893). I sample these tweets using the Boolean search query: 'Islam OR Muslims' (with wildcard variations) to ensure that many of the tweets pertain directly to Islamophobia (Waseem, 2016). I limit the amount of tweets taken from any individual account to just 20.

As discussed in the ethics section of the previous chapter (Chapter 3), user information is not provided and tweets are not quoted verbatim. Before being reported, tweets are cleaned extensively and amended. In cases where their content is particularly sensitive or the tweets (and authors) could be easily identified, they are amalgamated.

4.2 | The target of Islamophobia

A key debate in conceptual discussions of Islamophobia is whether it pertains to anti-Islamism (which can be understood as opposition against Islam *qua* religion/institution) or anti-Muslimism (which can be understood as opposition against Muslims *qua* social group) – or, indeed, both together. Researchers in social psychology often emphasize that prejudice refers solely to the treatment of individuals/groups and not to institutions or ideologies (Brown, 2010; Pettigrew & Tropp, 2008; Pettigrew et al., 2011). Some researchers in the field of Islamophobia make a similar argument; in a discussion of hate speech Rosenfeld explicitly states that 'disparaging religion cannot […] be equated with disparaging the religious.' (Rosenfeld, 2012, p. 277) whilst Halliday argues that 'the enemy image, then the enemy is not a faith or a culture, but a people' (Halliday, 1999, p. 898). However, many others who study Islamophobia, and in particular those who work closely with victims, argue that it should include both anti-Islamism and anti-Muslimism as both Islam and Muslims are targeted (Awan & Zempi, 2016; Chakraborti & Garland, 2015). Bleich argues that, 'Islam and Muslims are often inextricably intertwined in individual and public perceptions' (Bleich, 2011, p. 1587) and Allen similarly finds that Islamophobes often criticize Islam as a proxy for criticising Muslims (Allen, 2010b).

One way of excavating this issue is to situate it in relation to the 'non-identity' problem of moral philosophy. This suggests that actions are only morally important if they are 'person-affecting' in that they relate to the treatment of real living individuals. Or, as Parfit puts it, '"bad" acts must be "bad for" someone' (Parfitt, 1987, p. 363). Although the current investigation is analytical in nature rather than moral, this notion is a useful tool for understanding – and problematizing – the Muslim versus Islam dichotomy. Attacks against Muslims are clearly person-affecting in that they directly affect anyone who self-identifies as a Muslim. The link between individuals and their group identity is

well-established in cultural studies of religious aggression. For instance, Matsuda et al. explicitly argue that Islamophobic behaviours are an 'injury to a group. To privatize it ignored the greatest part of the injury.' (Matsuda et al., 1993, p. 8) When Muslim identity is attacked it is Muslims individuals – even if no single Muslim person is directly targeted – who feel this 'injury'. It is therefore incontrovertible that anti-Muslimism should be considered Islamophobic. The more contentious issue is whether anti-Islamism can also be considered person-affecting, and as such should also be considered Islamophobic.

In the dataset both 'Islam' and 'Muslims' are referenced frequently. Many references to Islam do not present it as an ideology, institution or doctrine but as a living breathing entity with intention, character traits, speech and moral responsibility. Intention is one of the most frequent tropes, as in tweets which state 'Islam wants to take over' or 'Islam is trying to change Britain'. Human character traits are also common, such as when Islam is described as 'sneaky', 'vicious', 'violent' or 'deadly'. In other cases, tweets exhibit prosopopoeia, attributing to Islam the ability to speak; 'Islam has called for this violence' and 'Islam begs forgiveness' – in many instances, such tweets provide considerable insight into how users perceive the intentions of Muslims. Yet other tweets suggest that Islam has the moral status of an individual with responsibility for its actions; 'Islam is responsible for this!' and 'Islam is to blame'. This anthropomorphism is widespread and indicates that Islam is often used as a proxy for discussing Muslim communities and the intentions, responsibilities and character traits ascribed to them. Insofar as this is true, then anthropomorphic references to Islam can be considered 'person affecting' because the true target of discussion is Muslims and not an abstract concept.

Not all discussions of Islam are anthropomorphic. In some tweets Islam is identified specifically as an ideology or doctrine. For instance, the nature of Islam is repeatedly described as 'fundamentalist' or 'totalitarian' – words which are far better suited to

describing an ideology or doctrine than a person or group of peoples. The term 'Sharia' also appears frequently, indicating that users are aware that Islam is an institution (i.e. that it can be implemented in a system of Law). This raises a new question: *can content which targets Islam be considered Islamophobic even if it is not anthropomorphic?*

The nature of the attacks made against Islam in the dataset indicate it is likely that they are experienced by individuals as attacks against Muslims *qua* group. For instance, several tweets attack Halal meat, describing Islamic practices of meat preparation 'barbaric', 'revolting' and 'unjust'. On the one hand, the target of these discussions is Islamic doctrine. But, on the other, because this doctrine is *materialized* and *actualized* by Muslims who prepare and consume such meat it is understandable that they experience it as a personal attack. Furthermore, even doctrinal attacks against Islam (for instance, as already noted, the claim that Islam is fundamentalist or totalitarian) imply something about Muslims that they would either willingly submit to such a doctrine or could be indoctrinated into accepting it. In few cases in the data is Islam discussed at a purely doctrinal level, such as by reviewing the content of scriptures or Islamic teachings. Previous research indicates that Muslims *experience* such attacks against Islam as personal attacks because their sense of self is so closely entwined with the religion – which is not to reduce Muslims solely to their religious identity (itself a trope of Islamophobic discourse) but, rather, to highlight its importance within the broader assemblage of identities which Muslims hold (Birt, 2009; Sadek, 2017; Wright, 2014).

Treating discourses about Muslims and Islam as separate phenomena constructs a false dichotomy. Muslims are the bearers of Islamic identity and as such an attack against Muslims *qua* Muslims entails – however implicitly – an attack against Islam. When Muslims are attacked for being violent or aggressive, it necessarily casts a shadow on the religion which they are associated with. At the same time, because Islam is the basis of

Muslims identity then any attack against Islam entails – again, however implicitly – an attack against those individuals which willingly bear that identity. Thus, whilst it may seem theoretically robust to associate Islamophobia only with Muslims rather than Islam (given that the main motivation for studying Islamophobia is how individuals – and not abstract concepts – are treated), this ignores the intimate connections between the two. Anti-Islamism is not the same as anti-Muslimism, but both should be considered constitutive aspects of Islamophobia. The target of Islamophobia is best conceptualised as a continuum, running from Islam at one end to Muslims at the other. Some tweets are targeted more against Islam and others more against Muslims, but all necessarily contain a *mix* of the two dimensions.

Two caveats should be noted apropos the imbrication of anti-Muslimism and anti-Islamism. First, with any bit of content which targets Muslims or Islam it is worth considering how much weight is placed on the 'Islam' end of the Muslim-Islam continuum. Some content discusses Islamic doctrine and practices in such a way that only minimal emphasis is placed on the Muslim dimension, and as such the content can only be considered very weakly 'person-affecting'. In practice, it may be worth excluding such content from the purview of Islamophobia. This can only be decided on a case-by-case basis, taking into account the context. It is unlikely to be necessary with social media because, as already discussed, most references to Islam in the data contain highly charged overtones about Muslims. Second, all that is at stake here is to determine the *target* of tweets. Just because the target of a tweet is person-affecting and pertains to Muslims, this does not mean that it is necessarily Islamophobic – this depends upon the *nature* of the tweets, as is discussed in the next section.

4.3 | Ideal types of Islamophobia

A key limitation of much existing work is that Islamophobia is not *defined* conceptually so much as *described* empirically. Indeed, in some studies, Islamophobia is not defined at all, leaving the audience to rely on their own (and potentially wildly divergent) understanding of what it is. In contrast, the present work is informed by Bridgman's seminal work on 'operationalization', and the need to disambiguate conceptual and empirical components of research (Bridgman, 1938; Crandall & Sherman, 2016). Specifically, Bridgman advises distinguishing between underlying concepts (such as 'prejudice' and 'harassment'), how they manifest (such as through physical attacks or verbal slurs) and how they are measured (such as by duration, intensity and target) (Bridgman, 1938). Thus, in empirical research one should distinguish between the core features of a concept and what merely describes the manifestations of that concept. Within the field of Islamophobia studies, Allen makes a similar argument, whereby he distinguishes between actions which are intentionally Islamophobic (Islamophobia-as-process) and actions which lead to Muslims experiencing harm (Islamophobia-as-product) (Allen, 2010b). In the present work, the focus is on conceptualising Islamophobia as a *process*.

The key problem with most definitions of Islamophobia is that they do not distinguish between conceptual and empirical aspects. As such, they are too broad, covering many non-conceptual aspects of Islamophobia, including how it manifests, actors, settings, effects, causes and even why Islamophobia is worth studying. This can be illustrated by analysing an existing definition from the literature, namely that offered by Bahdi and Kanji (for the Canadian context) in a 2018 paper titled, 'What is Islamophobia?'. They define Islamophobia as:

> ‘Perpetrated by private actors and the state for the purposes or with the effect of creating fear and hostility towards Muslim communities, Islamophobia is the belief that Muslims are different from the rest of Canadian society, and that Canada needs to be protected from Muslims because they are inherently violent, patriarchal, alien, and inassimilable. Islamophobia includes the explicit and motivated targeting of Muslims, as well as legislative, policy, and adjudicative silences that implicitly perpetuate long-standing, negative stereotypes of Muslims. Private and public forms of Islamophobia exist in a mutually reinforcing dialectic relationship.’ (Bahdi & Kanji, 2018, p. 345)

This definition includes several different facets. First, the *actors* who engage in Islamophobia (‘private actors and the state’). Second, the *effect* or result (‘creating fear and hostility towards Muslim communities’ and ‘implicitly perpetuate long-standing, negative stereotypes of Muslims’). Third, the supposed *cause* (‘because they [Muslims] are inherently violent, patriarchal, alien, and inassimilable.’) Fourth, the *manifestations* of the behaviour (‘explicit and motivated targeting of Muslims, as well as legislative, policy and adjudicative silences’). Fifth, the *dynamics* of Islamophobia (‘[…] exist in a mutually reinforcing dialectic relationship.’). These additional focuses are hugely important areas of study – but they are not, fundamentally, what defines Islamophobia. Indeed, the only part of the definition that is *conceptual* is: ‘Islamophobia is the belief that Muslims are different from the rest of Canadian society, and that Canada needs to be protected from Muslims.’ The underlying problem is that the authors of this definition – like many others in the field – are really trying to *describe* Islamophobia rather than define it.

From the corpus of academic literature, I identify a typology consisting of six distinct conceptualisations of Islamophobia: Fear and anxiety, threat, stereotypes, difference,

dominance and negativity. Each of these conceptualizations is broad enough that they could plausibly serve as the basis of Islamophobia – although, as discussed in the sections below, this would be highly problematic. These six conceptualisations are thematically separate accounts of Islamophobia, which are mutually exclusive (in that they are distinct) and, as far as I am aware, collectively exhaustive (in that they cover all of the main positions) of the surveyed academic literature.

The six conceptualisations can be understood as 'ideal types', a widely used typological tool for describing and categorizing various phenomena. Weber describes the ideal type as a 'one-sided accentuation […] [T]he synthesis of a great many diffuse, discrete, more or less present and occasionally absent concrete individual phenomena.' (Weber, 2017, p. 90) Ideal types are descriptive abstractions which enable researchers to identify the most salient features of social entities (such as ideologies, institutions and behaviours) and then use these features to categorize them into groups. Note that, despite its literal meaning, the term 'ideal type' is not a normative concept but an analytical one – 'ideal' refers to 'ideal*ized*' rather than something which is desirable. Because they are abstractions, the ideal types may not accord exactly with any of the definitions in the corpus. Many of the actual definitions are hybrids and will cut across the ideal types or will only *partially* embody them (Mudde, 2007a, pp. 12–15). Nonetheless, ideal types are a useful tool because they systematize and render explicit how conceptualisations of Islamophobia differ, making clear their most important and unique aspects. In the remainder of this section, the six ideal types of Islamophobia identified from the literature are investigated by analysing them with, and through, the Islamophobic hate speech observed in the corpus of tweets. This serves two purposes: (1) to better understand the content of the tweets and (2) to better understand the full implications of each conceptualisation.

4.3.1 | Fear and Anxiety

The first ideal type draws on the literal interpretation of Islamophobia, associating it exclusively or predominantly with *fear* of Muslims and the Islamic faith (Hervik, 2015). This ideal type can be defined as:

> 'Islamophobia consists of fear or anxiety towards Muslims or Islam'.

Examples of this ideal type are widespread in the literature, particularly with through intergroup threat theory. For instance, Kunst et al. advocate for a 'fear-based' conception of Islamophobia, which 'is about explicitly focusing on the fear response towards Muslims and their religion'(Kunst, Sam, & Ulleberg, 2013, p. 226). Similarly, Lee et al. define Islamophobia as 'fear of Muslims and the Islamic faith' (Lee, Gibbons, Thompson, & Timani, 2009, p. 93), Abbas as 'fear or dread of Islam or Muslims' (Abbas, 2004, p. 28) and Zúquete as 'a widespread mindset and fear-laden discourse in which people make blanket judgments of Islam as the enemy, as the "other"' (Zúquete, 2008, p. 323). The fear-based approach to Islamophobia is also ubiquitous amongst government institutions; as noted above, the Council of Europe defines Islamophobia as 'the fear of or prejudiced viewpoint towards Islam, Muslims and matters pertaining to them' (European Monitoring Centre on Racism and Xenophobia, 2006, p. 61).The United Nation's Human Rights Council offers a somewhat similar, if more expansive, definition of Islamophobia as 'a baseless hostility and *fear* vis-à-vis Islam, and as a result, a *fear* of and aversion towards all Muslims or the majority of them.' (The UN, 2007)

Related to the association of Islamophobia with fear is the association of Islamophobia with *anxiety*. Anxiety can be understood as 'a diffuse, unpleasant, vague sense of apprehension' (Sadock, Sadock, & Ruiz, 2014). Fear and anxiety are closely linked; the main difference is that whereas fear relates to the association of Muslims with particular

negative traits and characteristics – traits which are fear-*inducing* – anxiety relates more to a general uncomfortableness with Muslims even when there is no explicit provocation. Gottschalk and Greenberg put anxiety at the centre of their widely-cited definition of Islamophobia, arguing that Islamophobia 'reflects a *social* anxiety towards Islam and Muslims cultures […] this anxiety relies on a sense of otherness.' (Gottschalk & Greenberg, 2008, p. 5) Saeed similarly refers to how Muslims 'are the subject of public anxiety' (Saeed, 2007, p. 443) and Taras identifies contemporary responses to Muslims within a wider 'persisting European anxiety about Orientalism' (Taras, 2012, p. 112).

The traditional emphasis on 'fear' in Islamophobia suggests that fear and anxiety should be the most prevalent responses to Muslims and Islam. However, the data shows remarkably little evidence that users fear or are anxious about Muslims. In almost no instances do users explicitly state that they 'fear' Muslims or Islam, and tweets which express similar emotions are also infrequent – only a few tweets contain content such as: 'So concerned about Muslims' or 'Worried about Islam'. Interestingly, some tweets explicitly reject the notion of 'fear' whereby users claim that they are 'Islamorealists' rather than Islamophobes. Islamorealism is the idea that one cannot fear Islam because it is 'rational' rather than emotive to hate and oppose Islam. This does not in-itself prove that tweets do not express fear – and members of the far right are not known for being good arbiters of social scientific terminology – but it does suggest that linking Islamophobia to 'fear' may be misplaced given that many users explicitly do not see themselves as expressing fear or anxiety.

At a conceptual level, the association of Islamophobia with fear is deeply problematic. For something to be feared, presupposes the existence of a reason for that fear (however ill-founded or biased or emotive that reason is). Interestingly, this is embodied in the description of Islamophobia in the Runnymede Trust's landmark 1997 report. They write,

'Islamophobia is a useful shorthand way of referring to dread or hatred or Islam – and, therefore to fear or dislike of all or most Muslims' (Runnymede Trust, 1997, p. 1). Their use of the word 'therefore' constructs a causal and temporal link between hating Islam (step one) and then fearing or disliking all Muslims (step two). This definition itself indicates that fear is not an appropriate conceptual basis of Islamophobia; fearing Muslims supervenes on, and is therefore subsequent to, some prior outlook on Islam.

The data shows that fear has other connections with Islamophobia. Many tweets can be considered *scaremongering* in that they are likely to produce fear of Muslim in others. For instance, tweets include statements such as 'More mega mosques planned for the UK!' and 'Muslim men escaping justice, allowed to keep raping Brit girls'. These tweets do not express any fear but are likely to incite it in others. This suggests that the 'phobia' in Islamophobia refers not to fear contained in the Islamophobic behaviour but to the fear that it elicits in others. This is problematic for two reasons. First, is that many things may elicit fear – even including the actions of Muslims. Thus, taking this argument to its logical conclusion, it suggests that actions undertaken by Muslims could be considered Islamophobic (insofar as they contribute to making other members of society scared), which is a highly dubious – if not outright harmful – way of approaching Islamophobia. For instance, a Muslim wearing a burqa may incite 'fear' simply by walking past a (bigoted) person on the street. Claiming that the act of walking down the street is itself Islamophobic (because it elicits fear in others) is evidently ludicrous.

Second, is that this approach conflates Islamophobia with its effects which, as argued in the previous section, need to be kept separate. Another interpretation is that fear is important because it is what motivates users to express anger, hatred or aggression towards Muslims. In this regard, fear is a latent variable which *precedes* the expression of Islamophobia. For instance, tweets such as 'Muslims are scum' could be a product of

the user fearing Muslims or Islam. This is plausible empirically but does not make fear a good conceptual basis of Islamophobia. Indeed, it changes the question of 'what is the conceptual basis of Islamophobia?' to 'what creates Islamophobia?' – an important but nonetheless separate issue.

These three analyses show that fear and anxiety are complexly imbricated in the dynamics of how Islamophobia manifests – which may well be why the term 'phobia' has become so well-established in this field of research. However, they should not be viewed as the conceptual bases of Islamophobia as they are inherently separate to it.

4.3.2 | Threat: 'dangerous idiots'

The second ideal type can be defined as:

> 'Islamophobia consists of viewing Muslims or Islam as a threat'.

This is closely linked with the first ideal type as fears often emerge in response to threats. However, this link is in no way necessary; individuals may fear Muslims without explicitly viewing them as threatening and may view them as threatening without experiencing fear. Furthermore, these two ideal types emphasize different aspects of Islamophobia. The 'fear' ideal type places greater emphasis on how Muslims are *responded to*, whilst the 'threat' ideal type places greater emphasis on the *perceived role* of Muslims.

The 'threat' ideal type is commonplace in academic discourse. For instance, Bravo López defines Islamophobia as the view that Muslims are 'an internal and external enemy that threaten[s] both social cohesion and national security.' (Bravo López, 2017, p. 141) Part of Bahdi and Kanji's definition of Islamophobia is '[…] [the Country] needs to be protected from Muslims' (Bahdi & Kanji, 2018, p. 345) and in a study of Islamophobia in the Sweden Democrats, Mulinari and Neergaard argue that the hegemonic form of anti-

Muslim hatred is the discursive construction of 'the Muslim Threat' (Mulinari & Neergaard, 2011). Much academic research highlights how many UK far right groups, such as the EDL, portray Muslims as a threat to 'British values' (Allen, 2011; Jackson & Feldman, 2011; Treadwell & Garland, 2011).

The notion of threat is particularly prominent in Intergroup threat theory. This is closely linked to contact and conflict theory, discussed in the literature review, and places specific emphasis on how intergroup conflict emerges from the perception that another group poses a threat. Stephan et al. discusses how ethnic conflict in America is driven by both real and symbolic threats, reflecting how previously dominant social groups fear their economic, cultural and political power may be at risk (Stephan, Ybarra and Bachman, 1999). The perception of intergroup threat has been closely linked with support for far right parties, and empirical evidence provides greatest support to the idea that symbolic threats (including religious differences) are the most powerful motivators of conflict rather than economic ones, such as competition for scarce resources or employment opportunities (Lucassen and Lubbers 2012, Aichholzer and Zandonella 2016). Threat theory is likely applicable in this context, given the highly symbolic nature of Islamophobia within the UK context, perceived differences between it and the host nominally Christian culture, and the importance of Islamic identity to adherents, which enables bifurcation of areas into competing Muslim and non-Muslim groups (Jetten et al., 2004). Intergroup threat theory is also highly relevant for understanding online contexts, which Croucher proposes is an important site in the process of group acculturation and intergroup conflict and mediation (Croucher 2011).

Threats can take many different forms, including real threats (such as military, political and economic) and symbolic (including religious and cultural). Interestingly, only a few tweets in the data express the view that Muslims pose a religious threat to other faiths in

the UK, such as Christianity. What is more prevalent is the idea that the religious threat posed by Muslims is against the Christian *character* of the UK, although this too is only expressed infrequently. This is somewhat unexpected given the strong Christian nature of many far right parties (Bayrakli & Hafez, 2016; Wood & Finlay, 2008). No tweets express the view that Muslims pose an economic threat – which is surprising given that previous research indicates far right groups view Muslims as a 'drain' on the welfare state (John et al., 2004). Muslims are viewed as military threats only in relation to Islamic terrorism in the international sphere, such as with the rise of the Islamic State in Iraq and Syria (ISIS) in the Middle East. For instance, no tweets express the view that Muslims, or Muslim countries, are attempting to launch an army-based military attack against the UK. The two areas in which Muslims are perceived to pose the greatest threat is security and culture. Many tweets suggest that Muslims pose a security threat as they seek to commit atrocities in the West, such as terrorist acts. In particular, tweets often reference historical and contemporaneous Islamist terrorist attacks (whether in the UK or abroad). Tweets also often suggest that Muslims are engaged in 'cultural conflict' with the West, a longstanding idea amongst the far right (Eatwell, 2006; Stockemer & Barisione, 2017). For instance, several tweets describe how Muslims want to 'take over' the UK or are constitute an 'invasion'. The trope of 'creeping Sharia' is widely articulated, as is the notion of a 'militant' Islam – which, interestingly, despite its literal referent, is always situated in relation to terrorism and culture rather than the actual military.

Overall, the idea that Muslims are a threat is quite widespread in data, which may explain its prominence as a conceptual basis of Islamophobia, as well as its empirical strengths in studies of intergroup contact. Nonetheless, associating Islamophobia with threat poses two considerable conceptual problems. First, the notion of a 'threat' entails some 'thing' which is threat*ened*. In the data, the 'thing' which is threatened is primarily British values

or British society. However, what is most striking is that the discursive focus of most tweet is entirely on the Muslim opponent. In only a few tweets is any 'thing' explicitly referenced and it is often difficult to identify even implicit references. This severely dilutes the utility of a threat-based conceptualisation of Islamophobia. This finding is in line with previous research, which indicates that it might be generalisable to other contexts. For instance, in studies of the European populist radical right, Mudde finds they pay far more attention to what they oppose (such as Immigrants, Muslims or the EU) rather than what they stand for (Mudde, 2014; Mudde & Kaltwasser, 2007).

Second, the notion of 'threat' itself has certain implications; to view something as a threat is to view it as capable of inflicting some form of harm. This ascribes to Muslims a certain level of power, intention and competence. This is in tension with other depictions of Muslims found in the data, including statements which claim Muslims are lazy, simple-minded, disinterested or incapable of contributing to UK society. However, this tension is easily explained; Muslims are not revered or shown deference because they have the power to threaten. Instead, as one user put it, Muslims are 'dangerous idiots'; they are a threat not because of any competence or talent or initiative but because of their violence, idiocy and duplicitousness. This points to a fundamental problem with conceptualising Islamophobia in terms of threat; the threat posed by Muslims is entirely borne out of their supposed negative traits. This suggests that 'threat' is secondary to ascribing negative traits to Muslims – which, as such, would make a better conceptual basis of Islamophobia. This also reflects the basis of intergroup threat theory; threats are an excellent way of explaining *why* individuals express negativity against Muslims and Islam (Croucher 2011). However, this does not mean that threat and Islamophobia should be viewed as coterminous but, rather, the very opposite: threat motivates, and as such precedes, Islamophobia.

4.3.3 | Stereotypes

Many researchers link Islamophobia with the construction of stereotypes about Muslims – that is, fixed and oversimplified representations of Muslim identities, practices and beliefs. The third ideal type can be defined as:

> 'Islamophobia consists of creating and reproducing stereotypes about Muslims or Islam'.

For instance, Moosavi states that 'Islamophobia refers to stereotypical generalizations about Islam and/or Muslims' (Moosavi, 2015, p. 41), Johns and Saeed describe how in the West 'Islam […] is widely viewed through stereotypical lenses' (Johns & Saeed, 2002, p. 209) and Marranci defines Islamophobia in terms of 'the misrepresentation of the Muslim world' (Marranci, 2004, p. 107). Halliday similarly discusses Islamophobia in terms of stereotypes (Halliday, 1999) and Nadal et al. provide a taxonomy of Islamophobic micro-aggressions which includes 'endorsing religious stereotypes' (Nadal et al., 2012). Stereotypes are viewed as problematic because they grossly simplify the nuances of Islamic doctrine and Muslim practices and do not consider Muslims' internal heterogeneity (Nacos & Torres-Reyna, 2007). That is, stereotypes essentialize Muslims by presenting them as a *one* rather than a *multiplicity*. Stereotypical statements may be true for a small minority (although this is not necessarily the case) but are in no way representative of every Muslim or, indeed, the majority of Muslims.

The data shows that stereotypes about Muslims are widespread. In general, they can be identified by use of determiners such as 'every' and 'all', and the plural 'are' – but they can also be identified even when a single pronoun is used, such as 'I bet he'll be up to no good when he's older'. Common stereotypes in the data include different types of 'threat', such as the idea that Muslims pose security or cultural threats (as discussed in

the previous ideal type). Other stereotypes relate to how Muslims are patriarchal, homophobic, and 'foreign'. Interestingly, not all of the stereotypes found in the data are explicitly malign. For instance, in a small number of cases, Muslims are portrayed as inclusive, studious and abstentious – attributes which, in most settings, could be considered neutral if not benign. An example of this is use of the word 'staunch' to describe Muslims and references to the fact that many do not drink alcohol.

Even though stereotypes about Muslims are commonplace, equating this with Islamophobia is unsatisfying given how much stereotypes vary in their nature and tenor. The chief issue here is whether stereotypes *qua* stereotypes are inherently prejudicial – an argument which requires showing that even benign stereotypes about Muslims can be considered Islamophobic. This is a difficult position to hold given research in social psychology which suggests that stereotypes are innate to any process of social identity cognition. Social categorization theory suggests that when most people 'construe persons' (i.e. meet someone) they contextualise them in terms of how the group to which they are thought to belong is perceived (Kawakami, Amodio, & Hugenberg, 2017). Freemen and Ambady describe how 'the perception of other people is accomplished by a dynamic system involving continuous interaction between categories, *stereotypes*, high-level cognitive states, and the low-level processing of facial, vocal, and bodily cues.' (Freeman & Ambady, 2011, p. 247) Insofar as stereotypes (considered in a general sense) are a form of social categorization, and social categorization is intrinsically harmless then stereotypes are, innately, neither harmful nor hateful. Whether or not the stereotype is prejudicial depends upon the *nature* of the categorization.

Taken to its extreme, the stereotypes definition suggests that to simply talk of Muslims *qua* group is Islamophobic as this ignores their internal variety as persons. The problem here is that not only does the categoriz*ation* matter but also *who* does the categoriz*ing*

(Benesch, 2012). Categorizations which are put forward by Muslims, or by Muslim groups such as the Runnymede Trust, are inherently different to categorizations put forward by non-Muslims, such as far right groups or those with considerable social privilege. This argument – which is widely established in both cultural and linguistic analyses of language, in particular amongst speech act theorists (Hodgkin, 2017; Nemer, 2016), shows the limitations of using stereotypes/categorisation as the basis of Islamophobia. What matters is not the substantive nature of the stereotype but who constructs it and for what purpose. In turn, the 'who' is defined in relation to their subject position and social dominance; which suggests that stereotypes are not the axiological basis of Islamophobia but a secondary articulation which supervenes on relations of social dominance (as discussed and criticised in Section 3.5).

As with fear/anxiety, the dimension of stereotyping supervenes on a more primal form of Islamophobia (i.e. negativity directed against Muslims). Thus, whilst stereotypes *can* be and often are Islamophobic, this is only when the stereotypes are negative – it is not necessarily entailed by just the fact that they are stereotypes. Three important caveats should be noted here. First, is that in many contexts the line between positive and negative is blurred, and this distinction is more analytical than empirical. For instance, research on anti-Semitism shows that often seemingly positive statements (such as, 'Jews are so smart!') can bleed into negative statements (such as '… *and* cunning!') (Schiffer & Wagner, 2011). Second, and relatedly, is that it is plausible that many individuals who express even positive stereotypes about Muslims also express – possibly in other settings or to other interlocutors – hate against Muslims. That is, even positive stereotypes might be the epiphenomenon of other underlying prejudices and hateful attitudes. This still does not make positive stereotypes intrinsically Islamophobic, but points to a more nuanced relationship with Islamophobia. Second, even positive stereotypes may lead to bad

outcomes for Muslims. This is akin to Allen's point (discussed earlier) about how Islamophobic *results* can emerge from non-Islamophobic *processes* – and vice versa (Chris Allen, 2010b). Again, this does not make positive stereotypes Islamophobic, but raises important questions about how they function in society.

4.3.4 | Difference

Many accounts of Islamophobia hinge upon the idea that Muslims are radically different from others in mainstream society. Thus, the fourth ideal type can be defined as:

> 'Islamophobia consists of constructing and accentuating differences between (i) Muslims or Islam and (ii) the rest of society'.

This is closely related to the ideal type of stereotypes as many stereotypes hinge upon a cartoonish representation of difference between groups. Nonetheless, the ideal types are separable as difference can be discussed in many ways, not all of which involve the use of stereotypes. In the academic literature, both Hopkins and Gale and Modood situate Islamophobia in relation to the 'politics of difference' (Hopkins & Gale, 2009; Modood, 2003), Bahdi and Kanji define Islamophobia as 'the belief that Muslims are different from the rest of […] Society' (Bahdi & Kanji, 2018, p. 345) and Gottschalk and Greenberg claim that Islamophobia pertains to 'perpetuating notions of radical difference' (Gottschalk & Greenberg, 2008, p. 144). In a detailed conceptual study of the role of difference in Islamophobia, Werbner arrives at the conclusion that, 'Islamophobia is like other phobias and racisms, an incapacity to cope not only with difference but with resemblance' (Werbner, 2005, p. 8). A similar argument is also made by Meer, who describes the development of a 'language of difference' in Western literary and religious traditions, targeting Muslim minorities and Islam more generally (Meer, 2013). Others argue that discussing difference is often inherently prejudicial. This is nicely captured by

Bonnefoy who describes how discussions of difference are often processes of 'stigmatisation by distinction' (Bonnefoy, 2004).

The notion of difference is also common in theories of social psychology and political science which emphasize the simultaneously divisive and constructive role of the 'us/them' distinction. Mouffe argues that the construction of an 'us' and a 'them' is the fundamental task of all political movements, whether they be authoritarian and exclusionary in character or emancipatory and progressive: for her, the main challenge in political discourses is to decide *who* should fill the role of the 'them' (as it is a role that must be filled) and, as such, the role of academic researchers is to uncover the complex ways in which the us/them distinction manifests (Mouffe, 2005). Further, she argues that the 'us' can only emerge in response to identification of a 'them'; the construction of an in-group depends on its demarcation from an out-group, and as such establishment of difference is inherently normative (Mouffe, 2009).

'Us' and 'Them' are also widely used in social psychology to understand how ingroups are separated from outgroups, and social differences are constructed. In particular, the 'minimum group' paradigm suggests that simply by creating a social group a degree of in-group favouritism is established and perpetuated (Postmes, Spears & Lea, 1999, 2002). Simply the act of labelling a group of people *as a group* can lead to the emergence of ingroup bias as well as recognition of ingroup diversity. Even in seemingly random and arbitrary group settings (such as in experimental conditions), individuals are shown to favour, and respect the variety of, their ingroup. Existence of ingroup bias does not necessarily entail outgroup discrimination, but it raises important questions about the divisive role of establishing social differences. This also reflects other work in the social sciences regarding how constructing a group is a form of categorization, and as such impacts social relations, power and individuals' life experiences. All of these theoretical

perspectives suggest that, whilst it may seem neutral and 'descriptive' to identify difference, it is very rarely just difference but comes burdened with normative values and assumptions.

In one sense, the data shows considerable evidence of the difference ideal type. Several tweets make a direct comparison between Muslims and either (i) society in general or (ii) members of other groups, such as atheists. Many more tweets also indirectly point to the role of difference, implicitly suggesting that Muslims are separate by using pronouns such as 'you', 'they', and 'them'. However, in nearly all cases, 'difference' is not articulated on its own. Instead, it is accompanied by a negative value judgement, in which difference is framed as a problem. This is particularly apparent with tweets which describe Muslims as 'aliens' or 'incompatible' – in such cases, difference is used as a discursive tool to suggest that Muslims will cause harm to individuals (by, for instance, behaving misogynistically towards women because of their 'different' cultural values).

There are two key points here. First, at a discursive level, difference is not necessarily a marker of Islamophobia. Properly construed, difference is simply a fact of society and social identity; it is how difference is constructed and referred to which makes a tweet Islamophobic or not. And it should be noted that difference need not only be articulated negatively. Many political actors celebrate difference as the basis of multiculturalism and mutual intergroup respect (Archer, 2009). Indeed, countering Islamophobia and providing support to Muslims often requires recognising difference. As Cockbain notes apropos racism, 'refusing to talk about race at all risks fuelling racialised stereotypes and racist discourses.' (Cockbain, 2013, p. 30). Second, difference is rarely (if ever) just difference but is nearly always articulated with some normative aspect. This is demonstrated clearly by looking at the retweets in the dataset, many of which are from mainstream and liberal sources. Some of the original tweets contain either (i) positive

endorsements of the difference between Islam and other groups – for instance by celebrating the varied cultural, theological and religious contributions of Muslims – or (ii) support for the multicultural and multi-religious nature of the UK. In both cases, these are retweeted just so that they can be mocked and attacked. The partisan nature of the responses to difference demonstrates its inseparability from normativity.

The importance of difference when studying Islamophobia is somewhat paradoxical: difference is a constitutive part of Islamophobia – but it is also constitutive of any intergroup relation and this is precisely the reason why it is not a solid basis for conceptualising Islamophobia. Taken to its extreme, viewing difference as the conceptual basis of Islamophobia suggests that to simply talk of Muslims *qua* group is Islamophobic – as one is positing some sort of difference between Muslims and other members of society. This is an untenable position and may even be detrimental for groups seeking to raise awareness of Islamophobia and challenge it (Runnymede 2017, Ingham-Barrow 2018). Rather, the data shows that it is the *negative value judgement* associated with the identification of difference which distinguishes Islamophobic tweets from those which are non-Islamophobic. This is perhaps best summarised by MEND's work on the assumptions which underlie Islamophobia. They note that Islamophobia involves the view that 'Muslims are not only different, but this difference also makes them *inferior*.' (Ingham-Barrow, 2018, p. 10)

4.3.5 | Dominance

The fifth ideal type situates Islamophobia in relation to dominance. It can be defined as:

> 'Islamophobia consists of the systematic domination and exclusion of Muslims and Islam'

This is the only conceptualization of Islamophobia which explicitly considers the broader institutional, social and political structures in which Islamophobia takes place. It is also arguably the most abstract and hard to identify empirically. Bayrakli and Hafez offer a paradigmatic example of this thematic definition; Islamophobia is 'a dominant group of people aiming at seizing, stabilizing and widening their power by means of defining a scapegoat – real or invented – and excluding this scapegoat from the resources/rights/definition of a constructed "we".' (Bayrakli & Hafez, 2016) The dominance ideal type is internally heterogeneous but similar examples can be identified across the literature; Breen-Smyth defines Islamophobia as a 'form of subordination' (Breen-smyth, 2014, p. 223), Quinn as 'discriminatory oppression engrained within […] hierarchical divisions' (Quinn, 2018, p. 109) and Jackson as, 'a form of Eurocentric spatial dominance, in which those identified as Western receive a better social, economic and political "racial contract", and seek to defend these privileges against real and imagined Muslim demands' (Jackson, 2018).

Definitions which are rooted in dominance are highly varied, reflecting that the term itself is deeply contested (Howarth, 2016); the examples above focus on economic, political, cultural and social domination. This thematic grouping is best suited to understanding Islamophobia at the societal level, and is difficult to apply to understand individual behaviours, such as hate speech. There is little evidence in the tweets that Muslims are explicitly victims of other users' dominance. This is not because dominance does not exist. It is because the exclusion of Muslims from these communicative spaces is so complete that their dominance is rendered partially invisible (as it is in other, offline settings). Users are segregated such that Muslims are completely absent from the data and there are no dialogic exchanges with them. Thus, the dominance definition is hard to identify empirically – paradoxically, not because it does not function but because it

functions so effectively. This makes it difficult to study empirically, and as such means it is inappropriate for the present work. The role of dominance is also further complexified by the fact that some users express feeling that they are the dominated rather than dominators – that is, in some cases, white self-identified 'natives' suggest they are suffering at the hands of Muslims. This is reflected in tweets which claim that Muslim get preferential treatment and unfair advantages.

A key problem is that many conceptualisations of dominance situate the perpetrators of Islamophobia in larger fixed social structures (Giroux 1994; Butler, Laclau et al. 2000). By linking the question of whether a tweet is Islamophobic to the users' identity, these approaches evaluate individuals' behaviour in relation to who they are and not what they do. Taken to its extreme, this renders it difficult (if not impossible) for individuals to either act *contra* their identity or for their behaviour to change. The focus on individuals' identity rather than their behaviour is problematic as some tweets in the data express nuanced positions and uncertainty about the role of Islam and Muslims in society. For instance, several tweets actively and critically debate Islam's relationship with extremist behaviour, such as terrorism, indicating that the users' views are not fixed. Definitions which fix individuals in terms of their social identity effectively preclude the possibility of behavioural or ideological change such as this.

A related problem is that the dominance definition largely rejects the possibility that dominated groups can themselves be prejudicial. That is, these definitions focus on how dominant groups treat Muslims – but what if another dominated group (such as Immigrants) expresses anti-Muslim sentiment or discriminate against Muslims? It is not possible to identify the identity of all users in the dataset, and some may be from other oppressed and marginalised groups, such as the Gay community. Under Bayrakli and Hafez's definition, this could not be considered Islamophobic as the perpetrators are from

a group which is not in a position of social dominance. That said, an intersectional approach to Islamophobia could help to ameliorate this issue by examining cross-cutting forms of privilege and domination (Bilge, 2010; Mirza, 2013), but this is outside the scope of the present work.

4.3.6 | Negativity

The sixth and final ideal type explored here is negativity against Muslims. This can be defined as:

> 'Islamophobia consists of negativity directed against Muslims or Islam'.

This is one of the most widely used and robust conceptualizations of Islamophobia. For instance, Ekman defines Islamophobia as 'hatred or animosity aimed at Islam and Muslims' (Ekman, 2015, p. 1988), Allen as 'a certain perception of Muslims, which may be expressed as hatred toward Muslims' (Allen, 2017), Bleich as 'indiscriminate negative attitudes or emotions directed at Islam or Muslims' (Bleich, 2011, p. 1581), Semati as 'a single unified and negative conception of Islam' (Semati, 2010, p. 267), Hopkins as 'anti-Islamic feeling' (Hopkins, 2008, p. 54), Moten as 'dislike towards Islam and Muslims' (Moten, 2012, p. 155) and Luqiu and Yang as 'an overall negative view of Muslims' (Luqiu & Yang, 2018, p. 1). Indeed, Hussain suggests that negativity and hatred towards Muslims is so widespread and conceptually important that Islamophobia should be renamed "misoislamia" (Hussain, 2012) whilst Aguilera-Carnerero and Azeez similarly coin the term 'Islamonausea' (Aguilera-Carnerero & Azeez, 2016).

Negativity against Muslims is a recurrent feature of tweets in the dataset. Negativity manifests in many different ways, of which the four most common are:

1. Expressing hostility (such as by using expletives)

2. Calling for, or committing to engage in, actions against Muslims (such as physical violence, property destruction, 'invasions' of Mosques and even genocide)
3. Representing Muslims negatively. Includes:
    a. Ascribing to Muslims negative traits (such as being lazy or uncultured)
    b. Associating Muslims with negative behaviours (such as terrorism or female genital mutilation)
4. Responding to Muslims negatively (such as by expressing fear or distrust or disgust or not welcoming them)

Many tweets do not fit neatly into just one category but cross several, for instance a tweet such as 'f*cking Muslims, all a bunch of terrorists' both expresses hostility (by using the word 'f*cking') and presents Muslims negatively (by associating them with terrorism). As such, these four manifestations are best viewed as cross-cutting aspects of tweets rather than distinct types. The first manifestation (expressing hostility, often through the use of expletives) occurs particularly frequently with the other manifestations. Relatively few tweets use swear words alone (although this does occur, for instance in tweets like, 'Fuck allllll Muslims'). In general, the tweets are highly affective. Many tweets use expletives, emotive language, and convey a visceral sense of hatred. This suggests that any Muslim who observed them would be deeply affected and suffer considerable emotional harm (Tell Mama, 2015). This aspect of the data was anticipated given previous research (Amiri et al., 2015; Awan, 2014; Feldman, 2015; Ingham-Barrow, 2018). The second manifestation is by far the least prevalent in the data – but is also often viewed as the most socially concerning (Tell Mama, 2015).

The third and fourth manifestations are both very prevalent in the data. Tweets in these manifestations can be further subcategorized based on two criteria. First, is whether the

hate that is expressed is relative or absolute. Most tweets express absolute negativity against Muslims in that they make a standalone statement (such as 'Muslims are awful') but some also express relative negativity whereby Muslims are compared with some other group or society in general (e.g. 'Muslims are so much dirtier than the rest of us'). Second, is whether the hate is directed *to* Muslims (i.e. when Muslims are targeted as victims) or if it is *about* Muslims/Islam but no victim is named (i.e. when Muslims are discussed / talked about). These two dimensions (relative/absolute and targeted/absent) can be used to better understand the nature of anti-Muslim negativity found in tweets, as shown in Table 3. In the dataset many tweets express both absolute and relative negativity (which is also discussed above apropos the 'difference' ideal type). Interestingly, very few tweets refer to named individuals or use @ mentions to target Muslims – which is fortunate, given the harm that directly targeting individuals is likely to cause (Kumar et al., 2018). Indeed, the only Muslims that are named are Islamist terrorists, such as Abdelhamid Abaaoud, or Islamist hate preachers, such as Anjem Choudary.

Types of Blatant Islamophobic speech	**Absolute**	**Relative**
Negativity to Muslims	"You @[user] are nothing but a terrorist, should be kicked out of the UK"	"@[user], your kind are always more violent than the rest of us"
Negativity about Muslims	"So scared about what Muslims are doing to the UK"	"Islam is a more violent religion than Christianity"

Table 3, Different types of negative speech against Muslims

Conceptualizing Islamophobia in terms of negativity offers two important benefits. First, is that negativity is broad enough that many different behaviours and attitudes can be situated within it. This should reduce the risk of conceptual 'stretching' as the concept does not need to be constantly adjusted as new forms of Islamophobia emerge in society. This makes it highly suitable for empirical research – both in the context of studying hate

speech on social media but also in other settings. For instance, physical assault and harassment can be considered Islamophobic not because they are harmful and unpleasant in their own right (which is true irrespective of who they are directed against) but because when they are directed specifically against Muslims they use violence as a means of expressing anti-Muslim negativity. Secondly, it explicitly captures the fact that it is harmful to experience Islamophobia; even Islamophobia which is articulated but not directly experienced by Muslims (such as unseen Islamophobic tweets) still cause harm. Including 'negativity' as a definitional feature ensures that Islamophobia is recognised as a *problem* or social evil, and as such worthy of disapprobation.

One concern with conceptualizing Islamophobia in terms of negativity is that it risks subsuming non-prejudicial criticisms of Islam and Muslim practices. This is concerning as it is important that in a free and tolerant society individuals are able to express opposition to Islam *qua* religion (Malik, 2009). Many researchers express concerns about Islamic doctrine on liberal and democratic grounds or because they oppose the nature and impact of institutional religion. Summarizing the main issue at stake here, Imhoff and Recker warn against 'confounding prejudiced views of Muslims with a *legitimate critique* of Muslim practices based on secular grounds.' (Imhoff & Recker, 2012, p. 811) Thus, the problem is not *whether* non-prejudicial criticisms of Islam can be said to exist (they certainly can) but how they can be separated from Islamophobia if it is defined in relation to negativity. That is, without additional specification this definition risks leaving no space for non-prejudicial 'legitimate critiques'.

4.4 | Towards a definition of Islamophobia

The six ideal types discussed in the previous section provide a map of existing conceptualisations of Islamophobia. They can be summarized as outlined in Table 4.

Ideal type	Definition
Fear and anxiety	Islamophobia consists of fear or anxiety towards Muslims or Islam
Threat	Islamophobia consists of viewing Muslims or Islam as a threat
Stereotypes	Islamophobia consists of creating and reproducing stereotypes about Muslims or Islam
Difference	Islamophobia consists of constructing and accentuating differences between (i) Muslims or Islam and (ii) the rest of society
Dominance	Islamophobia consists of the systematic domination and exclusion of Muslims and Islam
Negativity	Islamophobia consists of negativity directed against Muslims or Islam

Table 4, Ideal types of Islamophobia

The data indicates that some of the most common approaches to defining Islamophobia (namely, the ideal types of threat and stereotypes) are indeed prevalent – but this does not mean that they are appropriate conceptual bases. For instance, exploitation and power are certainly aspects of Islamophobia at a societal level – but are ill-suited to understanding individual behaviours and are intractable bases of a definition for methodologically individual empirical research. Difference is constitutive of Islamophobia but also of its opposite, and as such although it is an important dimension of Islamophobia it cannot be its defining quality.

Negativity is a necessary aspect of Islamophobia. Indeed, it can be viewed as the conceptual basis of four of the other ideal types identified here. Fearing Muslims or viewing them as a threat (ideal types one and two) supervenes on viewing Muslims

negatively, stereotypes are harmful insofar as they express (generalised) negativity and difference is constitutive of social life, only becoming Islamophobic when it is constructed negatively. Thus, negativity constitutes a conceptual axiom. It does not require further philosophical argumentation as there is nothing conceptually prior to it. However, it suffers from one key limitation; it risks being too broad in that, on its own, it suggests that any critique of Islam is Islamophobic. Accordingly, there is a need to properly specify the 'negative' ideal type before it can be used robustly in empirical research.

One solution is to specify that anti-Muslim negativity is only prejudicial if its 'unjustified' or 'unwarranted'. For instance, in a study of immigration politics, Sniderman et al. argue that 'for a negative characterization of a group to qualify as prejudice, it must be, if not erroneous, at any rate unwarranted.' (Sniderman, Peri et al. 2000, p.18). Normative terms such as 'unjustified' and 'unwarranted' have become fairly widespread, particularly in non-academic discourses around Islamophobia (Runnymede 1997, Runnymede 2017). A less contentious and more conceptually robust approach is to consider whether the anti-Muslim negativity is *indiscriminate* (Bleich 2011). Indiscriminate anti-Muslim negativity occurs when derogatory claims are made about all Muslims, such as; 'All Muslims are scum' or 'Every Muslim should be kicked out!' Arguably, indiscriminateness gets at the right idea but with the wrong terminology. This is because many tweets in the dataset are highly discriminate in that they differentiate between different actors and practices. For instance, tweets which attack a specific Muslim or a specific group of Muslims (such as Muslim women or certain Muslim sects) target only one facet of the broader 'Muslim' identity rather than all Muslims. Tweets which attack one particular action undertaken by Muslims are also very discriminate – but nonetheless still could be considered Islamophobic.

A better, and closely related, concept to use for specifying Islamophobia is 'generality'. This can be best explained by drawing on the pioneering philosophical work of Laclau into the nature of political communication (Laclau, 2005a, 2005b). He argues that every political act has both a universal and particular dimension. All universality must be articulated via something particular (Laclau, 1996). For instance, abstract ideals of justice such as 'freedom' can only manifest in society through specific claims, such as activists who campaign for the right to life or on environmental issues. This formulation can be applied to understand the nature of Islamophobia, namely the fact that the very general 'universal' account of Islamophobia identified here (that Islamophobia pertains to negativity against Muslims) must always manifest in some content – content which, by its very nature, must be *particular*.

Laclau's work also shows that particularism can manifest in the form of *ambiguity*. Statements which are more ambiguous foreground the contingent nature of social reality and indicate that the speaker is open to dialogue – which, in Laclau's terminology, means that they are more 'ethical' (Laclau, 2005b). Statements which are ambiguous express *particularism* by showing that abstract universal notions (in this case, *negativity*) could be manifested differently; it could be targeted at different people, contexts or times other targets. In the context of Islamophobia hate speech, statements which are more certain indicate the user is committed to the negativity which they have expressed. And negativity which is more certain is also more likely to cause harm to individuals who view the tweet. Thus, the *generality* axis includes within it the extent to which *ambiguity* is expressed. This further clarification of the axis is useful for contextualizing anti-Muslim negativity more robustly.

For any tweet to be considered Islamophobic it must direct negativity, at least to some extent, against Muslims in general. Importantly, this means that a tweet which directs

negativity against a person who happens to be Muslim should not be considered Islamophobia as this is – in the context of studying Islamophobia – a *purely particular* expression of negativity. Tweets which are *purely general* (e.g. 'all Muslims are all bad in all places at all times') are incredibly rare. As such, Islamophobia is best understood as a *combination* of particularity and generality. This is why 'indiscriminateness' is an inappropriate specification of negativity as it means that only the most general tweets can be considered Islamophobic. This is a very high bar which few tweets cross.

Viewing generality as a question of degrees rather than as a binary issue (*a la* indiscriminateness) offers better insight into the nature of Islamophobia on social media. It elucidates that even though tweets might be specific apropos what is being targeted (e.g. Muslim terrorists or Muslim rapists) they still often contain a substantial *degree of generality* – and, as such, can be considered Islamophobic. This is well illustrated by a prominent feature in the dataset; news stories which report on Islamist terrorist attacks. These contain a seemingly neutral description, such as 'Muslim terrorists attack London Bridge!', and a link to the related news story. These tweets might seem like objective reports of an event, but it is not clear that they *need* to explicitly reference Muslims. Why not refer simply to the person as a terrorist; or a criminal; or, as in some cases, a person with mental health problems? Indeed, this is the norm in news coverage of 'right wing terrorism', in which assailants are often only identified by their gender or mental health problems (Falkheimer & Olsson, 2015). Thus, tweets which link Muslims to terrorist attacks (or, indeed, any other derogated form of behaviour) may seem highly particular but actually contain a hidden degree of generality.

Based on the arguments made so far in this Chapter, an Islamophobic tweet (what, in line with most previous research, I call Islamophobic 'hate speech') is defined as:

Content which expresses generalised negativity against Islam or Muslims

4.5 | Strong and Weak Islamophobia

The definition provided in the previous section provides an important starting point for studying Islamophobia by delineating and clarifying its core conceptual basis. However, previous research, as well as the data used in the present work, shows that Islamophobia has many different modalities. Most noticeably, in their influential work on prejudice Pettigrew and Meertens distinguish between 'blatant' prejudice, which they describe as 'hot, close and direct' and 'subtle' prejudice, which is 'cool, distant and indirect' (Pettigrew & Meertens, 1995, p. 58). Distinguishing between qualitatively different varieties of Islamophobic hate speech offers considerable theoretical and empirical advantages over using a single category and is now a prominent feature of work by anti-Islamophobia organisations, such as MEND. However, few studies of online hate speech have explicitly sought to systematize the different modalities of Islamophobia.

The key goal in distinguishing between different modalities of Islamophobia is to balance detail with systematicity; a nuanced and highly refined set of sub-divisions may better capture the heterogeneity of Islamophobia but also risk making it difficult to systematize that heterogeneity and apply the schema to different datasets. I opt for a bipartite division, splitting tweets into just weak and strong varieties. Note that the terms 'weak' and 'strong' are analytical rather than normative – they contain no claim about which may be considered 'worse' or is experienced by Muslims as being more harmful. For instance, whilst one might expect that 'strong' is more harmful as it tends to be more de-humanizing, it could also be that Muslims find 'weak' varieties more frustrating and affective as it indicates just how widespread Islamophobia is, creating a sense of despondency (Runnymede Trust, 2017).

Many different factors and dimensions can be used to distinguish between types of Islamophobia (such as strength, setting, speaker, audience (Benesch, 2012)). Drawing on

the arguments made in the previous section, I argue that the two axes identified – (i) negativity and (ii) generality – can be used to categorize tweets into the two varieties, weak and strong. Negativity and generality form a robust basis for distinguishing different types of Islamophobia as the data shows that they are separable empirically. For instance, some tweets are highly general (in that they refer to all Muslims) but are only somewhat negative (e.g. '*All* Muslims are terrible cooks') whilst other tweets are highly particular but very negative (e.g. '*the* Muslim who bombed London Bridge is a c*nt') or express particularism through ambiguity but are nonetheless still very negative (e.g. 'I'm starting to think that Muslims are a problem in Europe, something needs to be done'). At the extremes, determining the degree of negativity and generality is reasonably easy – but in more nuanced cases it is considerably harder. In all cases, researchers need to make a subjective decision.

The two dimensions of Islamophobia can be visualised in a two-axis grid, as shown in Figure 2. In principle, any tweet which expresses generalised negativity against Muslims can be situated within this schema and the degree of Islamophobia assessed. The vertical axis shows the degree of negativity, with 'High negativity' and 'Low negativity' as the two poles. The horizontal axis shows the degree of generality with 'Particular' and 'General' as the two poles. Note that the diagram is only a schematic and the axes do not have units.

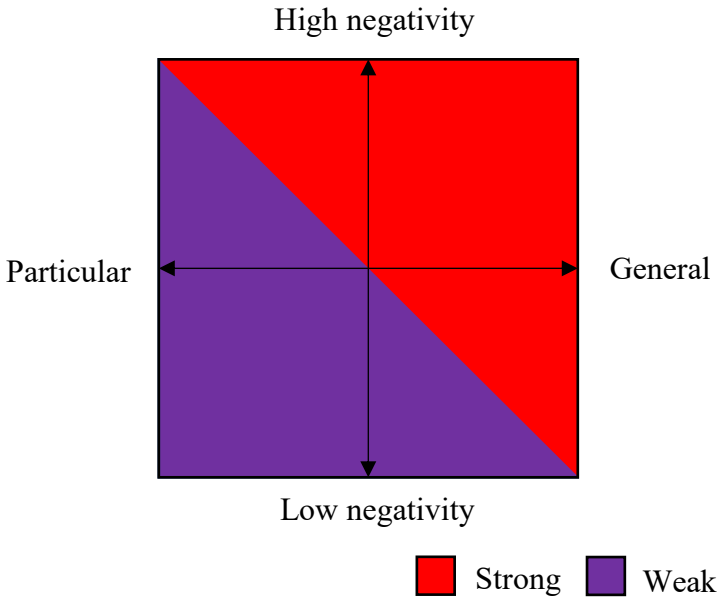


Figure 2, Characterising Islamophobic hate speech using the degree of negativity and degree of generality

Weak Islamophobia is typified by content which is highly particular and only weakly negative. Strong Islamophobia is typified by content which is both highly negative and highly general. Tweets which are a combination of these (i.e. it is only somewhat negative and only somewhat particular) might be either weak or strong – it depends upon the specificities of the tweet. Empirically, it is likely that tweets which are highly particular will be categorised as weak. This is because most of them are not so incredibly negative so as to cross over into strong. In particular, as shown in the annotation guidelines developed from this conceptual work in the following chapter, tweets which are about a single Muslim are categorised as Weak because they are so highly specific. Tweets which are highly negative but very particular and are categorised as strong tend to be those which use expletives. These issues are elucidated further in the practical application of this conceptual framework, as shown by the annotation guidelines (Appendix 5.1). Examples of the two types of Islamophobic hate speech are provided in Table 5.

Weak Islamophobia	• "Muslims are just different!" • "Muslim food smells so weird" • "Wearing a Burkha doesn't feel very #UK" • "Muslim terrorists attack London Bridge" • "Muslim radicals in the desert kill Christian hostage" • "Muslim paedo deserves jail"
Strong Islamophobia	• "Muslim men groom and rape children" • "Muslim mothers want to force FGM in the UK!" • "Typical, another bloody Muslim just blew himself up. LOSER" • "Fuck alllll Muslims" • "Muslim invasion, they're going to take over the UK" • "Top European Lawyer says that Muslims don't obey rule of law and should not be allowed to remain in Europe whilst posing a threat" • "The Police target Muslims because they're a problem, new #evidence" • "Huge rally atm against Loughborough Mosque – let's take back our country"

Table 5, Examples of Weak and Strong Islamophobic hate speech

The weak and strong categories of Islamophobia are posed as distinctive ways of understanding how different types of negativity against Muslims and Islam manifest. They can be viewed as broad 'frames' which reflect how a particular group (in this case, Muslims) are viewed in society, with implications for how that group are acted upon and activism is mobilized (Benford & Snow, 2000). Identifying the existence of weak and strong Islamophobic frames does not imply anything about the validity of those particular framings (which might, in certain circumstances, be well-evidenced or socially justified) nor whether they should be considered permissible under freedom of expression. Rather, these categories can be used as analytical tools to understand the different ways in which

Muslims are associated with negative traits, from the very weakly negative and particular to the very strongly negative and general.

Consider a 'factual' report about Islamist terrorist activity, e.g. 'Muslim terrorist attacks London bridge' – on the one hand this is a report of news. In many contexts, this is unavoidable if we are to discuss an event of nationwide importance. On the other, it can be considered weakly Islamophobic because it associates Muslims with a negative trait (terrorism), contributing to the reproduction of negative and undesirable framings. This form of weak Islamophobia may be entirely justified and, indeed, unavoidable. But this does not alter the fact that Muslim identity has been used in conjunction with a negative trait.

Crucially, the weak/strong distinction does not necessitate any assessment about the *intention* of the speaker. In the example above, the speaker may have myriad motivations in being weakly Islamophobic; they may want to spread negative views of Muslims and thereby stir opposition, or they might want to simply report on a news event, or to raise a question about multicultural integration. Irrespective of these differing intentions, all have served the same purpose: to reproduce a negative framing of Muslims. In certain contexts, the motivation behind the negative frame might be an important issue and worthy of in-depth research whilst in others it may be less important. In all cases, the utility of the weak/strong Islamophobia distinction is that it lets us, at first, identify *what* sort of negative framing against Muslims or Islam has been articulated. Only then is it possible to ask more in-depth questions about whether such frames are legitimate, useful or important and what their motivations are.

One implication of this line of argument is that it is difficult to talk about Muslims in certain contexts, and for certain practices, without risking a negative framing (and, as such, Islamophobia). We should not shy away from discussions solely because we risk

reproducing negative frames – indeed, such discussions might be necessary to address the problems facing Muslims communities. The point of the weak/strong distinction is not to label everyone who shares a weakly Islamophobic tweet an Islamophobe and to implement a *denkverbot* on free discussion. Rather, it is – at a descriptive and analytical level – to capture the fact that by sending a weakly Islamophobic tweet, individuals are reproducing a negative framing of Muslims. This is a social science phenomenon in need of study, even if such framings appear unavoidable. It is worth noting that this discussion departs considerable from a lay person's understanding of what both 'Islamophobia' and 'hate' mean, and due care should be taken to avoid using these arguments to legitimate censorship.

Weak and strong Islamophobia are, therefore, best understood as analytically distinct notions which capture different types of negative frames against Muslims and Islam. That said, they are intimately connected, and can be viewed as two levels on a single 'ordinal' scale. Both categories can be situated on a single ordinal scale because both are routed in the same two orthogonal axes: generality and negativity. In both cases, more negativity and more generality are associated with strong Islamophobia whilst less is associated with weak. It is therefore possible to claim that strong is 'more' Islamophobic because both the constituent axes of Islamophobia (negativity and generality) are higher. This does not imply anything else; strong is not necessarily morally worse or more prevalent and nor is there a pre-defined pathway of radicalization from weak to strong. The distinction is solely analytical whereby, on the terms of the definition outlined here, strong Islamophobia can be viewed as a greater form of Islamophobia than weak.

Situating weak and strong on an ordinal scale means that the order of the values is important; strong Islamophobia is greater than weak. However, they are not directly comparable such that one would be a multiple of the other and the exact size of the

difference between them is effectively unknown. As such, we cannot claim that strong is ‘twice’ as much as weak. Rather, strong is ‘more’ Islamophobic than weak, but the difference is unspecified. The fact that both weak and strong can be situated on the same scale is useful for conducting social scientific analysis empirically, as shown in the proceeding chapters.

4.6 | Conclusion

Islamophobia is an elusive concept. If nothing else, this investigation has demonstrated how varied and complex it is, and how many different ways in which it can be conceptualised. Through a close reading of the data, and in-depth critical analysis, I have answered RQ 1 and argued that the conceptual basis of Islamophobia includes two dimensions: (i) the degree of negativity and (ii) the degree of generality – and these dimensions can also be used to sub-divide hate speech into weak and strong varieties.

This conceptualization captures, and draws on, the most salient and informative features of the other ideal types of Islamophobia identified here (including, fear, threat, difference and stereotypes). Importantly, the focus on generality separates Islamophobia from intractable notions of truth and veracity. Furthermore, by rigorously investigating different ideal types of Islamophobia, I was also able to provide new insight into the empirical manifestations of Islamophobic hate speech, such as the fact that threat and fear are separate, fear is not the hegemonic form of Islamophobia, and difference is constitutive but not unique to Islamophobia.

The conceptual arguments made in this Chapter are routed in the specificities of the data, including the particular domain (Twitter), the context (UK politics) and the type of behaviour (hate speech). Nonetheless, in principle, the conceptual work undertaken here could be used to understand other types of Islamophobia in other domains, such as offline hate speech, physical assault and institutional forms of prejudice. There are good reasons to think that these behaviours are fundamentally similar to hate speech in that they all involve the targeting of Muslims and Islam. However, they may have very different modalities and forms of expression, and the schema developed here might be inappropriate – thus, a focus of future work is to explore and validate the applicability of the arguments made here.

One important area of potential application is the hateful conduct policies of social media platforms, most noticeably Twitter. As noted previously, the weak/strong distinction is posed as an analytical difference to capture the negative ways in which Muslims are framed in contemporary discourse. This should not necessarily result in such speech being censored or banned, and as such the distinction may be of less importance to a platform which, in tackling hate speech, is concerned primarily with public relations management and avoiding legal challenges. In general, social media platforms are concerned only with very overt and direct forms of behaviour and in developing action-oriented policies rather than grappling with deep conceptual problems. Thus, only the work on strong Islamophobia is likely to be relevant – even though in principle the full conceptual framework could be used to unify and provide coherence to their, at present, largely inchoate approach to handling and labelling hate speech. Perhaps a separate point of relevance for social media platforms more broadly is that this chapter shows the variety, complexity and strength of negativity expressed against Muslims. There is a need for all platforms to to not only counter well-established forms of hate, such as racism, but also this sort of religious based negativity.

Finally, the arguments made here point to how ill-suited the existing terms used to describe Islamophobic hate speech are: it is not about phobia, not (primarily) about Islam and is not (solely) about hate.. However, it is important not to risk what Sayyid terms ‘etymological fundamentalism’ (Sayyid, 2010, p. 13). ‘Islamophobic hate speech’ is widely used in academia, government and by tech companies and as such the main goal of this Chapter has been to clarify, explore and specify what it refers to through both qualitative thematic analysis and philosophical argumentation – rather than positing a new term to replace it.

Chapter 5 | Classifying Islamophobic hate speech

The purpose of this Chapter is to realise the additional research goal stipulated in the literature review:

> *To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia*

In Chapter 4, Islamophobic hate speech was defined in relation to two axes: negativity and generality. This was then used to distinguish between weak and strong varieties. Drawing on this, I define two classification tasks. First, detecting Islamophobic content. Second, distinguishing the strength of Islamophobia. I call these tasks, respectively, the binary and multi-class tasks. I follow a strategy to build the binary and multi-class classifiers in tandem. First, I create the multi-class classifier and then, second, I collapse the weak and strong categories together to create the binary classifier.

The structure of this Chapter mirrors the steps for creating a supervised machine learning classifier. In the first section, I outline the creation of a training/testing dataset, which contains labelled instances for each of the studied classes (here, none, weak and strong Islamophobic). In the second section, I engineer and test input features. In the third section, I select and test the appropriate algorithm to model the data. In the fourth section, I discuss the implementation of the two classifiers and report on performance. Accuracy for the multi-class classifier, tested on an unseen dataset, is 77.3% and balanced accuracy is 83%. Accuracy for the binary classifier is 88.3% and balanced accuracy is 89%. These results demonstrate that both classifiers can be used in the subsequent empirical chapters.

5.1 | Training/testing dataset

A key concern in previous work is building a suitable training and testing dataset, annotated with the outcome variable. In many machine learning applications, the outcome variable is easy to measure and requires little engineering. However, hate speech is notoriously difficult to study and process, not least because often the outcome variable is not pre-defined but must be extracted from the text itself (Schmidt & Wiegand, 2017). Given the work in the previous chapter to define Islamophobia, I opt to create a training/testing dataset by manually annotating a sample of tweets. This is the most widely practiced method in previous research (Schmidt & Wiegand, 2017). However, although widespread, there are noted problems with the manual annotation method. Many studies have low inter-rater and intra-rater reliability scores, which makes it difficult to trust the reliability of the dataset (Ross et al., 2017). Often this is because insufficient attention is paid to (i) developing a robust and well-specified classification schema and (ii) implementing it under appropriate conditions. In some studies, the object of study (usually, hate speech) is been poorly defined and the annotation schema lacks depth and clarity. It is unsurprising that in such cases the annotations are poor. Thus, as Ross et al. recommend for the field, 'raters need more detailed instructions for annotation.' (Ross et al., 2017, p. 1)

Benoit et al. discuss two primary methods for annotating text; (i) annotation 'by experts applying comprehensive classification schemes to raw sources' and (ii) 'crowd-sourced annotation by a large number of non-experts' (Benoit et al., 2016, p. 278). They argue that because crowdsourced annotation is scalable, agile and cost-efficient it is therefore more reproducible and scientific. Crowd-sourcing is particularly useful in cases where there are large numbers of documents to be annotated or where the features can be easily identified. It has been used effectively in recent research to annotate social media content

for hateful or prejudicial sentiment (Davidson, Warmsley, Macy, & Weber, 2017; Malmasi & Zampieri, 2017; Williams & Burnap, 2016; Wulczyn, Thain, & Dixon, 2016). However, it is not appropriate for all research and in some cases leads to poor inter-rater agreement and 'junk' annotations. For instance, Davidson et al. used Crowdflower to annotate 25,000 tweets. Three or more annotators coded every tweet, and percentage agreement was 92%. However, annotators performed very poorly in some areas. For instance, 5% of tweets were labelled 'hateful' but only 1.3% were unanimously so – meaning that only 25% of the 'hateful' tweets had full agreement. Similarly, 76% of tweets were labelled 'offensive' but only 53% unanimously so – meaning that only two-thirds of 'offensive tweets' had full agreement. These results show how even a large, well-funded and well-led study can have poor results with crowd annotations.

In a study of hate speech annotation Waseem reports that 'systems trained on expert annotations outperform systems trained on amateur annotations.' (Waseem, 2016, p. 138). A further limitation of crowd-sourcing is that it can be time-consuming, difficult and expensive to test and monitor the work of annotators (Bohannon, 2011). Given the complexity and context-specific nature of the phenomenon which is being studied in the present work – Islamophobic speech acts across both weak and strong manifestations – expert manual annotation is the most appropriate method for building a training/testing dataset, as used in several previous studies (Djuric et al., 2015; Founta et al., 2018; Gitari, Zuping, Damien, & Long, 2015; Warner & Hirschberg, 2012). This is a labour-intensive approach but also a robust one; as D'Orazio et al. note, 'although expert coding is costly, it produces quality data.' (D'Orazio, Kenwick, Lane, Palmer, & Reitter, 2016, p. 1).

A key development in recent studies of annotation is (i) to use annotators who have experienced prejudice personally and (ii) to be transparent about the identities, expertise and backgrounds of annotators. Noticeably, this goes beyond academia - commercial

organisations like FactMata, and third sector initiatives such as the Credibility Coalition, have used members of victimized communities to annotate online content (Credibility Coalition, 2018; FactMata, 2018). Waseem investigates how annotators from different backgrounds perform at hate speech annotation tasks (Waseem, 2016). He finds that individuals who have experienced forms of social oppression and prejudice directly – such as women, minority ethnic groups and transgender people – are more attuned to identifying hate speech and are more likely to reach an intersubjective consensus as to which content is hateful. This not only improves the rigour and social utility of annotations but also leads to better inter-rater agreement between annotators. At the same time, the choice of annotators should be balanced and not include individuals who are too personally effected by the phenomenon studied as this might lead them to make overly biased or emotive judgements.

5.1.1 | Implementation of training/testing dataset

Three annotators are used in the present work, including the author. For details on their backgrounds and identities see Appendix 5.1. Tweets are annotated by all three annotators using specially created annotation guidelines, which are available in Appendix 5.1 and online at `https://github.com/bvidgen`. The first draft of the guidelines was based on the conceptual analysis of Islamophobia in Chapter 4, previous hate speech annotation studies and an initial preliminary study of 200 tweets by the author. They were then developed iteratively through discussions with the annotators, a second preliminary study (outlined in Appendix 5.1) and by reading the tweets in the dataset. Ultimately, any guideline or schema is inherently limited: no guideline is exhaustive but must rely on the judgement of those who implement it. Thus, *how* the guideline is implemented can be as important as *what* is implemented. As such, the annotators communicated regularly with the author of the present work to resolve any issues and deal with complex annotations.

4,000 tweets are annotated to create a training dataset for the classifier.[4] To ensure the classifier can be applied robustly across all tweets in the present work, the annotated dataset is sampled from tweets analysed in all of the empirical chapters. Building an annotated dataset with sufficient instances of hateful content is a time-consuming endeavour, not least because in most online contexts the prevalence of hate is relatively low overall. This makes it difficult to 'build a corpus that is balanced with respect to hateful and harmless comments' (Schmidt & Wiegand, 2017, p. 7). To ameliorate this problem, Waseem and Hovy recommend increasing the prevalence of hate speech by sampling annotated data which contains associated topics, such as searching for tweets which contain relevant keywords like "Muslim" or "Islam" (Waseem & Hovy, 2016). This approach is partially adopted here, as 1,000 of the tweets are sampled using keyword searching. This reduces the representativeness of the data and may introduce considerable biases. In particular, the use of keywords may bias the dataset towards relatively 'obvious' and overt forms of Islamophobia rather than more subtle manifestations, such as hate expressed through polysemy and obfuscatory linguistic methods. This is a well-established challenge in this area of research and, given the importance of identifying sufficient instances of hate to train the classifier, is a reasonable tradeoff. Nonetheless, due caution should be used when considering the applicability of the classifiers' developed here. The sources of tweets used to create the annotated dataset are shown in Table 6. For more details on how tweets are sampled see Appendix 5.1.

[4] The number of tweets selected for annotation is based on previous work and the logistical constraints of annotation.

Source	Number of tweets
Far right seed accounts	1,000
Followers of the BNP	500
Followers of Britain First	500
Followers of the Conservatives	500
Followers of UKIP	500
Keyword search within the entire dataset of tweets (produced by followers of the BNP, Britain First, UKIP, Conservatives and Labour)[1]	1,000
TOTAL	4,000

Table 6, Sources of tweets for full annotation study

The full dataset of 4,000 tweets are annotated by all three annotators. I test the annotations for both inter-rater reliability and intra-rater reliability. Inter-rater reliability measures how consistent different annotators are. I calculate percentage agreement, Fleiss' kappa and Krippendorf's alpha for all three annotators (McHugh, 2013; McHugh, 2012). Inter-rater reliability scores are very high across all three measures, indicating strong agreement between annotators; percentage agreement is 89.9%, Fleiss' kappa is 0.837 and Krippendorf's alpha is 0.895. These values compare well with previous studies (Schmidt & Wiegand, 2017; Waseem, Davidson, Warmsley, & Weber, 2017) and suggest that the annotation schema, and its implementation, is sufficiently robust for use in the present work. Intra-rater reliability measures how internally consistent annotators are. This is an important measure as often annotators' evaluations shift over time, usually due to either fatigue or better understanding of the dataset and annotation guidelines. I measure intra-rater reliability on 100 tweets for each of the annotators and report very high values in all three cases, between 95% and 97%. Full details of the annotation process and testing are provided in Appendix 5.1.

In cases where annotators disagree (389 out of 4,000 tweets, 9.7%), the majority decision is used to assign tweets to classes. For instance, if two annotators annotate a tweet as 'Not

Islamophobic' and one annotator annotates it as 'Weak Islamophobia' then it is assigned to the 'Not Islamophobic' class. Of the 4,000 tweets in the annotated dataset, 894 tweets are labelled either weakly or strongly Islamophobic. A balanced dataset is important as otherwise the model can be skewed towards performing well on the most prevalent categories and poorly on the less prevalent. This can make it difficult to interpret the results. To create a balanced dataset for training the multi-level classifier, I reduce the number of tweets labelled 'not Islamophobic' from 3,106 to 447. The final multi-level classifier dataset consists of 1,341 tweets. This is shown in Table 7. For the binary classifier, the weak and strong classes are combined. This is shown in Table 8.

Category	Number of Instances
Not Islamophobic	447
Weak Islamophobia	484
Strong Islamophobia	410

Table 7, Number of annotated tweets in each class in the final dataset for multi-class classification

Category	Number of Instances
Not Islamophobic	447
Islamophobic	894

Table 8, Number of annotated tweets in each class in the final dataset for binary classification

The 1,341 tweets do not directly reflect the sources of the training data. Table 8.1 shows the breakdown of tweets by source, updated from Table 6 to reflect the origins of the 1,000 tweets identified through keyword searching. This shows that, in general, the far right accounts are over-indexed in the final sample whilst the mainstream parties are under-indexed. Importantly, tweets from followers of all parties are included. Noticeably, tweets from followers of the Labour party are considerably less prevalent than other parties (comprising just 1% of the final sample). This is concerning as it may limit the

applicability of the classifier to tweets from followers of this party. The lack of tweets from Labour followers is because they were not specifically sampled which, in turn, is due to the timing of the research and the early creation of the classifier. Nonetheless, given noted similarities in the supporters of Labour and the Conservatives (both of which are large moderate and liberal parties with a nationwide mandate and a track record of entering government in Westminster), it is expected that the classifier will still be applicable to these users' tweets. Importantly, the broad political context and timing of all the tweets is the same, which should increase the classifiers' applicability. In future research projects, this omission will be addressed and the full data studied sampled using a stratified method to ensure representative coverage.

Source	**Number of tweets in original 4,000**[5]	**% of tweets in original 4,000**	**Number of tweets in final 1,341**	**% of tweets in final 1,341**
Far right seed accounts	1,181	30%	549	41%
Followers of the BNP	848	21%	335	25%
Followers of Britain First	701	17%	256	19%
Followers of the Conservatives	577	14%	67	5%
Followers of UKIP	620	16%	121	9%
Followers of Labour	73	2%	13	1%
TOTAL	4,000	100%	1,341	100%

Table 8.1, Sources of tweets for 1,341 tweet training/testing dataset

[5] Values are updated from Table 6 to show how the tweets in the 'Keyword search within the entire dataset of tweets (produced by followers of the BNP, Britain First, UKIP, Conservatives and Labour)' segment break down across the other sources, including followers of Labour.

5.1.2 | Baseline algorithm accuracy

The distribution of tweets over classes is used to calculate baseline accuracy measures for both tasks (Witten, Frank, & Hall, 2011). First, I use the random classification algorithm; the probability of classifying tweets into the right category is calculated based on the categories' prevalence. Second, I use the 'zero rule' algorithm; all tweets are assigned to the most prevalent category (i.e. for the multi-level task, weak Islamophobia). Particularly in tasks with uneven class sizes, the zero-rule algorithm performs better, making it a more robust baseline comparison than random classification. The baseline performance of both classifiers is shown in Table 9. Baseline performance is low for the multi-class task, ranging from 33.49% to 36.09%, showing its difficulty. For the binary task baseline performance is higher, ranging from 57.55% to 69.43%.

Baseline measure	Accuracy
Multi-level classifier – random probability	33.49%
Multi-level classifier – zero rule	36.09%
Binary classifier – random probability	57.55%
Binary classifier – zero rule	69.43%

Table 9, Baseline accuracy for both classifiers

5.2 | Feature selection

Feature selection refers to the choice of input variables used to train the classifier. In many cases features are selected using 'brute force' computation via a grid search with little consideration for *why* they have been included. Models in which variables are selected without any theoretical justification may perform well in initial testing but only due to overfitting. Overfitting can be defined as when 'a classifier is tuned to the *contingent* characteristics of the training data rather than the *constitutive* characteristics of the categories.' (Sebastiani, 2002, p. 15) This is a problem not only because overfitted

models are computationally complex and hard to interpret (Biran & McKeown, 2017) but also because they cannot be generalized to unseen dataset and as such are largely unsuitable for empirical research applications (Dietterich, 1995; Domingos, 2012). Accordingly, in the present work, both performance optimization and theoretical justification are considered when selecting input features.

An important development in previous research has been to leverage user-level information to improve classification. This is particularly effective in cases where textual data is limited or the context – including who speaks, where and with what authority – plays an important role. The intended meaning and social impact of an otherwise ambiguous social post can be more easily inferred if the user's past behaviour is taken into consideration as this shapes the conversational dynamic in which content is shared. Dadvar et al. recently showed this with a study of cyberbullying on YouTube which leveraged user level information to more accurately tag bullying and offensive comments (Dadvar, Trieschnigg, Ordelman, & De Jong, 2013). This has been developed in subsequent research, measuring the 'bulliness' of users and explicitly modelling the role of small-scale conversational contexts in determining whether posts are harmful (Dadvar, Trieschnigg, & Jong, 2014). Similar approaches in the literature (i) identify hateful users and then build a training dataset by taking all of their content (Kwok & Wang, 2013) and, separately, (ii) use various content and user-level features to predict which users produce hateful content (Ferrara, Wang, Varol, Flammini, & Galstyan, 2016; Ribeiro, Calais, Santos, Almeida, & Meira, 2018).

A key flaw of these approaches is that users who are casual, temporary or one-off haters will most likely be missed or their behaviours under-represented – unless the threshold for being classified as 'hateful' is incredibly low (which would, in turn, overestimate the level of hate speech). Users who are less hateful, or just less active, are more likely to be

misclassified as there are fewer signals to detect. This is a substantial problem in the context of social media where behaviour can spread contagiously and users may, with the right prompts, engage in atypical behaviours (Romero, Meeder, & Kleinberg, 2011). For instance, recent research by Cheng et al. shows that the idea that trolls are 'born not made' is misplaced as 'anyone can become a troll.' (Cheng et al., 2017) They report that in the appropriate setting trolling behaviour can spread contagiously from person to person, even if many of the individuals involved have not trolled previously and seemingly do not have a troll-like disposition. This is particularly an issue with Islamophobia, which Baroness Warsi warns has now passed the 'dinner-table test'; Islamophobia is no longer the purview of just a radicalised extreme minority but is often expressed casually in 'the most respectable of settings and by the most respectable of people' (Runnymede Trust, 2017, p. v). Accordingly, it is crucial that the behaviour of even casual infrequent Islamophobes is studied as well as the behaviour of those whose prejudice is explicit or recurrent. For this reason, I opt to *not* use user-level features as inputs to the classifier, including data (such as users' descriptions), meta-data (such as the date the users' account was created) and network data (such as how well connected the user is to other individuals who engage in Islamophobia).

5.2.1 | Surface and derived features

Features can be divided into 'surface' features, which are extracted easily from the text, and 'derived' features, which require additional computation and transformation (Lai, Guo, Cheng, & Wang, 2017). In a meta-study of previous research Schmidt and Wiegand report that nearly all hate speech classifiers use surface features such as URLs, punctuation, and capitalization (Schmidt & Wiegand, 2017). Bag of Word (BOW) term unigrams are particularly well-used and have been shown to perform highly at certain tasks, such as detecting overt racism (Greevy & Smeaton, 2004; Kwok & Wang, 2013).

However, the predictive power of unigrams can be limited if there are many infrequently appearing terms, which is often the case even after terms have been stemmed. Term sparsity is particularly a problem with user-generated content, such as tweets, as spellings tend to be idiosyncratic and slang is used (Derczynski, Ritter, Clark, & Bontcheva, 2013). In addition to unigrams, term n-grams of length two or more have been widely used in previous research, although these suffer from an even greater problem of sparsity (Hee et al., 2015; Schmidt & Wiegand, 2017). I anticipate that surface features, such as punctuation, capitalization and BOW term unigrams and n-grams, will be useful input features for the classifier.

Characters and character n-grams are widely-used derived input features. An interesting recent finding is that character n-grams can be more predictive than term n-grams for identifying abusive language as mis-spellings are partly mitigated (Badjatiya, Gupta, Gupta, & Varma, 2017; Mehdad & Tetreault, 2016; Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016). Sentiment, which typically refers to the degree of positivity/negativity expressed in a document, is also often used as an input feature for classifying hateful tweets (Giatsoglou et al., 2017). Gitari et al report that extracting the 'polarity' of sentiment in tweets improves hate detection as it helps in identifying 'subjective sentences', which are defined as sentences which express feelings, views or beliefs (Gitari et al., 2015). In a study of sexism on Twitter Jha and Mamidi find that 86% of 'Hostile' tweets contain negative sentiment and only 3% contain positive (the remainder being neutral) (Jha & Mamidi, 2017).

The risk with models which use sentiment is that they might perform well at capturing the emotive or 'angry' types of hate speech but perform less well at the colder – and no less harmful – varieties. A similar problem is posed by the use of sentiment analysis in Burnap et al. (Burnap et al., 2015), though arguably its use in this case is more appropriate

given that the classifier was developed specifically for Twitter activity following a highly charged football match. Furthermore, although sentiment is a fast-improving area of computational text analysis, with off-the-shelf sentiment dictionaries, such as the Linguistic Inquiry and Word Count dictionary (Tausczik & Pennebaker, 2010) and SentiStrength (Thelwall, Buckley, & Paltoglou, 2012) now easily available, the accuracy of sentiment classifiers remains somewhat limited (Mäntylä, Graziotin, & Kuutila, 2018), particularly in cases with unusual emotive expressions, such as sarcasm (Maynard & Greenwood, 2014). Accordingly, given that the data consists of tweets, sentiment may be too noisy a signal to include as an input feature.

5.2.2 | Language syntax

A growing area of research points to the importance of the syntax of language for detecting hate. Burnap and Williams report how prejudice is often expressed without using hateful or derogatory terms (Alorainy et al., 2018; Burnap & Williams, 2016). For example, the phrase 'send them home' does not use any derogatory terms and would most likely not be classified as hateful by a simple keyword classifier, yet is clearly exclusionary and expresses negativity towards the targeted group. Burnap and Williams argue that hateful behaviour on social media can often be identified by the use of certain relational grammatical structures which 'Other' excluded groups. These, accordingly, can be used to identify hateful and abusive content. Burnap and Williams test for this by implementing a lexical parsing model which uses the Stanford dependency parser to automatically extract typed dependencies in tweets (de Marneffe & Manning, 2008). This differs considerably from a Part-Of-Speech tagger, which merely identifies the grammatical category of each term (rather than their grammatical position with a statement) using either rule-based or probabilistic approaches (Brill, 1992). Burnap and Williams report that using typed dependencies leads to a 10% reduction in false negatives

for classifying racist speech versus a baseline of only using BOW unigrams (Burnap & Williams, 2016, p. 8). Similar results have been reported by Silva et al., who use sentence structure to capture hate (Silva et al., 2016).

Burnap and Williams' work has been extended in Alorainy et al. (Alorainy et al., 2018), who create an 'othering lexicon' which contains 'two-sided' othering language. This is language in which the first pronoun refers to the self and the second pronoun to the other. Alorainy et al. use the othering lexicon to identify even indirect and subtle exclusionary content, implementing it with sentence embeddings via the paragraph2vec algorithm. Across four types of identity (religion, disability, race and sexual orientation) F-measure scores are reported as 0.93, 0.95, 0.97 and 0.92, which suggests this method is currently best in class. Focusing on the syntactic structure of tweets is best suited for dealing with general expressions of hate and offence, as it captures a more universal feature of negativity expressed through language, rather than Islamophobia specifically. Given that there are likely to be other overlapping targets of hate in the tweets (such as immigrants and ethnic minorities), this could reduce precision by creating many false positives. Thus, given also the technical challenges of implementing a syntax-based approach for classification, in the present work this option is not explored.

5.2.3 | Word embeddings

Recently, word embeddings have been widely used to remediate the problem of term sparsity in text. The output of word embedding models can be used as inputs for supervised classifiers, trained for a variety of tasks (Lai, Xu, Liu, & Zhao, 2015), including hate speech and offensive language detection. The key concept behind word embeddings is that 'you shall know a word by the company that it keeps' (Firth, 1957, p. 11). That is, the meaning of words can be uncovered by observing which other words

they are frequently used with. For instance, using the distributional hypothesis one can work out that 'dog' and 'canine' are related simply because they occur in similar contexts without any prior hardcoded knowledge of language. Word embeddings represent linguistic units (most often words but also sentences and paragraphs) as low-dimensional dense vectors, which are learnt using neural networks (Gambäck & Sikdar, 2017). The real power of word embeddings is that words which are used in similar ways (i.e. they occur in similar contexts) have similar vector representations. These can then be used for a variety of tasks, such as grouping words and documents together, information retrieval and words/document recommendation, and word embeddings 'maths', such as the now infamous 'King + Woman = Queen' or more nuanced maths such as 'Paris – France + Poland = Warsaw' (Vylomova, Rimell, Cohn, & Baldwin, 2015). In this second example, the difference between Paris and France captures the idea of 'capital city' which is then added to the country Poland to get Warsaw (the country's capital).

Word embeddings have been widely used in previous hate speech detection. Badjatiya et al. report on how word embeddings trained on a corpus of 2 billion tweets improve classification of hateful speech (Badjatiya et al., 2017). Djuric et al. similarly use sentence embeddings, through the paragraph2vec algorithm (Dai, Olah, & Le, 2015; Le & Mikolov, 2014) to learn low-dimensional document representations (Djuric et al., 2015). They report a modest improvement in the Area Under the Curve (AUC) against two baselines, one using only BOW unigrams and the other using BOW unigrams weighted by term frequency inverse document frequency. Gambäck and Sikdar study how character n-grams and word embeddings can be combined to improve accuracy, although they report the highest F1 score (0.78) for just word embeddings alone (Gambäck & Sikdar, 2017). Noticeably, Alorainy et al. use paragraph embeddings with their 'othering lexicon' to drastically improve performance at a multi-category

classification task (discussed above) (Alorainy et al., 2018). Overall, previous research strongly suggests that word embeddings are an important input feature.

The two most widely used algorithms for transforming words into vectors are word2vec (Le & Mikolov, 2014; Mikolov, Chen, Corrado, & Dean, 2013) and global vectors (GloVe) (Pennington, Socher, & Manning, 2014). Both algorithms work by learning vectors to represent words based on how they co-occur with other words. Both can be used for the two main tasks in semantic analysis, (i) predicting a target word given a set of context words (via a continuous bag of words or 'CBOW' model) and (ii) predicting context words based on a target (via a 'skip-gram' model). In its most basic implementation, word2vec trains a shallow feedforward neural network to predict target words from their context; vector representations are randomly initialised and are then updated over many iterations, their weights and biases adjusting to maximise target word prediction. GloVe is similar but also uses global statistical information about word co-occurrences. The main advance of this model is that it considers the ratios of word co-occurrence probabilities rather than the word probabilities alone. This should improve how well the vectors capture semantic meaning and, in particular, the authors argue that gloVe outperforms word2vec at word analogy tasks (Pennington et al., 2014). Because it makes use of corpus' underlying co-occurrence statistics, gloVe is considered a 'count' model, whilst word2vec is solely a 'prediction' model (Baroni, Dinu, & Kruszewski, 2014).

FastText is a recently released embeddings model from Facebook which creates vector representations based on character n-grams rather than words (Bojanowski, Grave, Joulin, & Mikolov, 2016; Joulin, Grave, Bojanowski, & Mikolov, 2016). As with word2vec and gloVe, it is based on a shallow neural net. FastText is considered better at finding representations of rare words as it uses character n-grams – which are likely to

be more prevalent than the words which they form. It also performs well with unseen words (provided that they are formed of character chunks which are present in the seen words). FastText has already been used in some speech classification tasks but the results are mixed (Badjatiya et al., 2017; Jha & Mamidi, 2017; Kumar et al., 2018; Park & Fung, 2017; Taylor, Peignon, & Chen, 2017). Noticeably, of particular relevance for the present work, Badjatiya et al. compare FastText with gloVe for hate speech classification and find that it does not improve accuracy (Badjatiya et al., 2017). The greatest advantage offered by FastText is that it is considerably faster than other models to implement and thus is well-suited to real-time applications (Joulin et al., 2016), however this is not a key concern in the present work.

I anticipate that all three of the word embeddings models would perform similarly in the downstream NLP task of the present work, Islamophobic hate speech classification. I opt to use the gloVe algorithm, given the strengths of how it was designed and its performance in previous work. Pre-trained word embeddings have been widely used in previous research for a variety of purposes, including sentiment analysis (Giatsoglou et al., 2017), information retrieval (Zuccon, Koopman, Bruza, & Azzopardi, 2015) and even political ideology detection (Iyyer, Enns, Boyd-Graber, & Resnik, 2014). One advantage of pre-trained models is that they are easy to implement and are very robust because they have been trained on a vast corpus of data. For instance, Google has trained the Word2Vec algorithm on 100 billion words from its Google News dataset (Google, 2018). For most researchers, it is infeasible to collect this quantity of data.

Pre-trained embeddings have several limitations which can reduce accuracy when used in classification work (Kamkarhaghighi & Makrehchi, 2017; Rezaeinia, Ghodsi, & Rahmani, 2017). First, some words in the target dataset (in the present work, the dataset of 1,341 annotated tweets) may be missing from the pre-trained model. This is

particularly likely in contexts, such as social media, where slang and mis-spellings are common. Second, language is context specific and the text used to train the model may not reflect how words are used in the new text being studied. For instance, Wikipedia is often used as a training set for word embeddings. But it is likely that language use on Wikipedia is more formal, considered and grammatical than in other contexts, which can limit the applicability of models. Third, pre-trained embeddings are unlikely to take into account the polysemic nature of language. For instance, the word 'beetle' refers to both a car and an animal. With pre-trained models it is unclear which word meaning was most common in the text used for the training. Specially-trained models are similarly unable to distinguish between multiple word meanings but because they are trained in just one context are more likely to capture the most salient word meaning. A model trained on, and used to study, car reviews would likely capture the car meaning of the word 'beetle'. This is not the universally valid 'right' meaning of the word but is likely to be the most useful given the specific context. Thus, given these limitations, it is anticipated that a *newly trained word embeddings model* (i.e. one that is trained on the corpus of tweets collected for this thesis) will perform best.

5.2.4 | Implementation of feature testing

I test models containing input features engineered and extracted from the annotated dataset, using the Naïve Bayes algorithm with ten-fold cross-validation. Naïve Bayes is used because previous research indicates that it outperforms most other off-the-shelf algorithms for text classification tasks (Fernández-Delgado, Cernadas, Barro, Amorim, & Amorim Fernández-Delgado, 2014; Wainer, 2016). It is also relatively simple to implement as it does not require much parameter optimisation and is deterministic, producing the same results each time it is implemented. I measure the performance of

models with accuracy. Accuracy is defined as the number of true positives plus the number of true negatives, divided by the total number of instances.

First, I create a text only model, using one-hot encodings for each term. This is where the text is represented in a document-term matrix with counts for each term's occurrence – inevitably, with large number of documents, one-hot encodings are very sparse. Second, I create a model using 50 surface-level and derived non-text features (e.g. presence of URLs, sentiment scores, part of speech tags). These are described in detail Appendix 5.2. Third, I create a combined model that uses both one-hot encodings and all 50 non-text features. Fourth, I create a model using pre-trained gloVe word embeddings, trained on two billion tweets (Stanford, 2018). Fifth, I create a model using newly-trained word embeddings (trained on the full corpus of 140 million tweets collected as part of this thesis). Due to its complexity, the word embeddings model is tuned extensively for (i) the extent of text cleaning, (ii) term frequency minimum, (iii) size of word window, (iv) number of vectors, (v) n-gram variations, (vi) word vector calculations and (vii) the number of tweets used. The tuned parameters are shown below and example scripts for calculating and visualizing the parameters are available online at `https://github.com/bvidgen`.

1. Text cleaning = cleaned text + stop words removed, but terms not stemmed
2. Term frequency minimum = 5
3. Size of word window = 10
4. Number of vectors = 200
5. N-grams = unigrams only
6. Word vector calculation = vector means based on binary occurrence of terms
7. Number of tweets = ~140 million

Finally, sixth, I create a model which uses the newly-trained word embeddings as well as all 50 of the non-text features in Model 2. The accuracy of the six models is shown in Table 10.

Input feature model	Accuracy
Model 1: Text only (one-hot encoding)	30.07%
Model 2: Non-text features	49.96%
Model 3: Text + non-text features	30.36%
Model 4: Pre-trained word embeddings	63.20%
Model 5: Newly trained word embeddings	69.13%
Model 6: Newly trained word embeddings + all non-text features	65.20%

Table 10, Accuracy of models with different input features for multi-class classification

The pre-trained word embeddings model considerably outperforms the text-only model, with accuracy almost twice as high (63.20% in model 4 compared with 30.07% in model 1). This provides compelling evidence that word embeddings are the most appropriate text-based input feature for the classifier. Furthermore, I find that the newly trained word embeddings considerably outperform the pre-trained word embeddings, with accuracy higher by 5.9 percentage points (69.13% for model 5 compared with 63.20% for model 4). This suggests that the benefits of having a model which is trained on tweets which are contextually-specific outweighs the cost of having a smaller dataset. This is in line with previous work, such as Lai et al., who find that 'corpus domain is more important than corpus size.' (Lai, Liu, He, & Zhao, 2016, p. 8) Non-text features introduce considerable noise when added all at once and reduce accuracy (65.20% in model 5 compared with 69.13% in model 6). However, I also conduct initial testing in which variables are included on a case-by-case basis. The preliminary results suggest that certain non-text features can optimize the classifier to increase accuracy. I opt to complete this testing in full only once the choice of algorithm has been finally decided.

5.3 | Choice of Algorithm

An algorithm can be understood as a set of fixed instructions which is formally defined and performs a set of actions on an entity, thereby transforming its state (Belinski, 2001). The choice of algorithm can materially affect classifiers' performance and, as such, is an important consideration. For most classifications tasks there is no need to develop proprietary algorithms as off-the-shelf freely available algorithms, such as naïve Bayes, Support Vector Machines (SVM) and random forests, perform highly (Kotsiantis, 2007). Whilst algorithms' performance is always context-specific, previous studies have benchmarked the most widely-used algorithms to guide researchers in choosing one. Fernández-Delgado tests 179 algorithms from 17 families on 121 classification tasks (Fernández-Delgado et al., 2014). He reports that the random forest family of algorithms perform best, followed by SVM and neural networks. A similar result is reported by Wainer on a smaller set of algorithms and classification tasks (Wainer, 2016). In the field of hate speech detection, Schmidt and Weigand find that most practitioners use SVM (Schmidt & Wiegand, 2017). However, across different use cases various algorithms have been found to perform best. For instance, Burnap and Williams report that random forests outperform SVM (Williams & Burnap, 2016), whilst Warner and Hirschberg opt for SVM (Warner & Hirschberg, 2012), Kwok and Wang use Naïve Bayes (Kwok & Wang, 2013) and both Davidson et al. and Nobata et al. use regression models (Davidson et al., 2017; Nobata et al., 2016). Neural networks are also increasingly being used as algorithms to classify hateful and prejudicial content (Alorainy et al., 2018; Badjatiya et al., 2017; Gambäck & Sikdar, 2017; Mehdad & Tetreault, 2016).

5.3.1 | SVM

The SVM algorithm works by learning a hyperplane which can be used to data into non-overlapping groups. It is often easy to separate low-dimensional data into separate groups with a straight line in linear space. However, with higher dimensional non-linear data, it is usually necessary to transform the data into a new space with a non-linear kernel function. In the new space it is then possible to accurately separate the data into groups. SVM works by implementing this process, accounting for complex mappings between data when separating them. It has traditionally been used only for binary classification, but in recent times has been adapted for multi-class tasks (Hsu & Lin, 2002; Zhang & Zhou, 2014). I use the 'one against one' multi-class strategy. In this approach, separate SVM classifiers are created for each pair of classes and then the results aggregated, rather than just creating one classifier for each class against all others (as with the main alternative strategy, 'one against all'). As such, whilst more computationally expensive, the 'one against one' strategy is expected to optimize performance (Milgram, Cheriet, & Sabourin, 2006). The SVM algorithm is optimized by adjusting the choice of kernel, gamma values and regularisation 'C', as recommended in previous research (Ben-Hur & Weston, 2010; Bennett & Campbell, 2000).

5.3.2 | Deep learning

Deep learning comprises a suite of algorithms inspired by the biological neural networks of brains. They have received considerable attention in computer science because of their high performance at difficult classification tasks with super high-dimensional data, such as images. Najafabadi et al. describe how deep learning is 'motivated by artificial intelligence emulating the deep, layered learning process of the primary sensorial areas of the neocortex in the human brain, which automatically extracts features and

abstractions from the underlying data.' (Najafabadi et al., 2015, p. 2) In practice, deep learning works by taking inputs (that is, features extracted from the data), feeding them into layers of 'neurons' which weight and process them, and then returning an output. A big challenge in deep learning is creating the appropriate architecture; Brown et al. argue that with neural networks, 'performance […] depends heavily on the underlying system architecture' (Hermundstad, Brown, Bassett, & Carlson, 2011, p. 1). Numerous options can be customised, from the type of neural network (recurrent and convolutional neural networks are the most widely used for text analysis) to the setup, such as the activation function, to the hyperparameters, such as the number of layers and epochs (Goldberg, 2017).

Neural networks have been shown to perform remarkably well at even difficult classification tasks, including hate speech. One limitation is that deep learning algorithms are 'data hungry' and require large volumes of labelled training data, which can make them inappropriate for some hate speech tasks where it is difficult to create a large training dataset (Lecun, Bengio, & Hinton, 2015). Nonetheless, given their high performance in prior research, it is anticipated that a deep learning algorithm will perform best in the present work. Accordingly, a multi-layer perceptron is implemented, a deep learning architecture which is feed forward and 'shallow' (i.e. it has few layers of neurons). It is optimized by adjusting the number of epochs, optimization function, activation function and learning rate.

5.3.3 | Implementation of algorithm testing

Six different algorithms are tested on the annotated dataset using the newly trained word embeddings model as an input (Model five in Table 10 above): Naïve-Bayes, random

forests (with trees = 10, 100 and 1,000), logistic regression, decision trees, SVM and deep learning. The results of algorithm testing are shown in Table 11.

Algorithm	**Accuracy**
Naïve-Bayes	69.13%
Random Forests (trees = 10)	65.40%
Random Forests (trees = 100)	68.72%
Random Forests (trees = 1000)	67.94%
Logistic Regression	69.13%
Decision Trees	61.23%
SVM with kernel = 'radial' + 'C' = 2 + gamma = 0.01	72.17%
Deep Learning with epochs = 100 + activation function = 'relu' + optimization function = rmsprop, learning rate = 0.001	71.14%

Table 11, Accuracy of different algorithms on newly trained word embeddings model

All six algorithms perform well, with accuracy ranging from 61.23% to 72.17%. Interestingly, increasing the number of trees in the random forests algorithm initially increase accuracy (from 65.40% to 68.72% when the number of trees increases from 10 to 100) but then reduces it (from 68.72% to 67.94% for 100 to 1000 trees), which is most likely due to overfitting on the training sets. The comparatively strong performance of Naïve-Bayes (69.13%, fourth highest) validates its use for input feature testing earlier in this Chapter. The two highest performing algorithms are SVM and deep learning. They outperform all the other algorithms by at least 2 percent.

The performance of SVM compared with deep learning in text classification has long been a point of debate (Zaghloul, Lee, & Trimi, 2009). Although deep learning has been heralded as the future of machine learning, several recent studies suggest that SVM can outperform it for certain tasks (Korba & Arbaoui, 2018; Liu, Choo, Wang, & Huang, 2017). In this case, the SVM outperforms the deep learning algorithm, with accuracy higher by 1.03 percentage points (72.17% compared with 71.14%). This is surprising given the extensive work undertaken to adjust the many hyperparameters and settings of

the deep learning algorithm. It is most likely due to the relatively small size of the annotated dataset, which may have constrained the deep learning algorithm. Thus, contrary to my initial expectations, I opt to use SVM for the Islamophobia classifier. The hyperparameters fitted to optimize the SVM ('C' = 2 and gamma = 0.01) are well-suited to the empirical applications of the present work; they indicate that the data is separated by a wide and smooth hyperplane, which minimizes the risk of overfitting.

5.4 | Results and discussion of performance

The final Islamophobia classifier draws on the results of the testing in the previous two subsections. The main input is a new word embeddings model trained on the full corpus of 140 million collected tweets. As discussed above, a 'one against one' multi-class SVM algorithm is used. The classifier is optimized by including additional features, identified via a grid search (discussed in Appendix 5.2). One to seven additional input features are tested, in order to identify the specific combination of additional input features which maximises accuracy. The marginal increases in accuracy from including additional features is shown in Table 12. Overall, the marginal gains of adding input features decreases as the total number of input features increases, although the rate of decrease is not monotonic.

Model	Word embeddings	One feature	Two features	Three features	Four features	Five features	Six features	Seven features
Highest accuracy	72.48%	73.38%	73.94%	73.96%	74.38%	74.55%	74.60%	74.59%
Percentage point increase	/	0.9	0.56	0.02	0.42	0.17	0.05	-0.01

Table 12, Marginal increase in accuracy from including additional features

The final model consists of word embeddings plus six additional features:

> Word embeddings + count of mentions of Mosques + presence of HTML + presence of RT + count of part of speech: 'conjunction' + count of named entity recognition: 'location' + count of named entity recognition: 'organisation'

The features included in the final classifier are consistently included in the most accurate models across all rounds of testing. This provides evidence that they are constitutive rather than contingent signals for detecting Islamophobia, and thus likely to be generalizable (Sebastiani, 2002). All of the features can also be justified on theoretical grounds. Mentions of Mosques most likely reflects that individuals are discussing Islamic practices or possibly Islamophobic terrorist attacks (which often target Mosques). This is a feature which I specially engineered for this project using data from Wikipedia and which could be of use in other studies of online Islamophobia. The presence of HTMLs suggests that users are linking to content outside of Twitter, such as news stories and blogs, which indicates a level of purposiveness, potentially associated with expressing hatred against Islam and Muslims. The presence of RTs is interesting as it suggests that much hateful content is typically retweeted from others; users are more comfortable sharing rather than creating Islamophobic hate. Potentially, this is because it is perceived to give them some distance from the source of hatred. The presence of locations and organisations can also be linked to Islamophobia, as often Islamophobia manifests in terms of negativity directed at particular places, locations and groups. The inclusion of the part of speech 'conjunction' can also be explained theoretically; conjunctions link clauses together in a sentence through words like 'but', 'or' and 'and'. Weak forms of Islamophobia use conjunctions to express implicit and indirect negativity against Islam. As such, this input feature likely provides important signal for distinguishing between the different strengths of Islamophobia.

5.4.1 | Multi-class classifier performance with cross-validation

The accuracy of the multi-class classifier in ten-fold cross-validation on the annotated dataset is 74.60%. Folds are partitions in the data which enable testing within a single dataset – using 10 folds means that I split the data into 10 partitions and then train a model on 9 partitions and test against the tenth. I then repeat this so that each of the 10 partitions is, in turn, the testing set, ensuring full coverage of the data. I then calculate additional performance metrics, drawing on research in information retrieval (Parikh, Parikh, Mathai, Chandra Sekhar, & Thomas, 2008; Sokolova & Lapalme, 2009; Velez et al., 2007) Each of the metrics computes a different combination of the true positives (positive values which are correctly classified as such), true negatives (negative values which are correctly classified as such), false positives (negative values which are incorrectly classified as positives) and false negatives (positives values which are incorrectly classified as negatives). All of the results are shown in Table 13.

Fold	Accuracy	Balanced accuracy	Precision	Recall	F1 score	Specificity
1	0.796	0.846	0.795	0.798	0.797	0.893
2	0.76	0.808	0.75	0.736	0.743	0.88
3	0.736	0.808	0.74	0.75	0.745	0.866
4	0.721	0.792	0.714	0.724	0.719	0.86
5	0.718	0.774	0.686	0.685	0.686	0.863
6	0.746	0.808	0.74	0.742	0.741	0.873
7	0.702	0.785	0.699	0.721	0.71	0.85
8	0.79	0.845	0.793	0.793	0.793	0.896
9	0.756	0.809	0.736	0.736	0.736	0.881
10	0.735	0.798	0.733	0.729	0.731	0.867
Mean	0.746	0.807	0.739	0.741	0.740	0.873

Table 13, Performance of multi-class classifier in cross-validation over ten folds

Recall[6] is the number of true positives divided by the number of true positives plus the number of false negatives. It measures how well the classifier performs at identifying all relevant instances. Precision[7] is the number of true positives divided by the number of true positives plus the number of false positives. It measures how much noise or 'junk' is included alongside the relevant instances. The importance of these metrics depends on the setting and application; in cases where false negatives can be hugely harmful, such as screening for cancer (as not identifying the cancer means the patient will not receive adequate treatment), then recall is important. In cases where false positives impose costs, such as when the police stop and search suspects (as citizens falsely identified as suspects will experience severe disruption to their lives and likely emotional harm) then precision is important. The F1 score averages both these values, giving a useful and interpretable summary. For the multi-class task, I aggregate the recall and precision scores (and, as such, the F1 scores) for each level using the macro-aggregation strategy described by Sokolova and Lapalme. Values are first calculated for each class and then the per-class agreement is averaged, with each class treated equally (Sokolova & Lapalme, 2009). The multi-class classifier performs similarly for recall and precision (0.741 and 0.739 respectively), and as such has a similar F1 score (0.74). This is encouraging as it means that it does well at balancing the need to identify relevant instances with identifying noisy junk, and as such can be easily applied to real world 'wild' data.

Balanced accuracy is a relatively new metric put forward by Velez et al. They argue that 'the presence of imbalanced classes is an issue for data mining and classification' as it renders traditional metrics for assessing classifier performance less informative (Velez et

[6] Recall is also known as 'Sensitivity'
[7] Precision is also known as 'Positive predictive value'

al., 2007, p. 308). This is particularly a problem for small datasets, where even small differences in class size can have considerable impact. To ameliorate these issues, Velez et al. combine specificity and sensitivity into a single metric ('balanced accuracy'). They find this is a robust measure of performance for many empirical applications as it is less biased and involves no additional data manipulation, such as over- and under- sampling. Subsequent research has shown the practical utility of using balanced accuracy to measure performance (Carrillo, Brodersen, & Castellanos, 2014), and it is now also used by commercial machine learning platforms, such as DataRobot (DataRobot, 2018). Balanced accuracy of the classifier is particularly high (0.807), which provides further evidence that the classifier does well at balancing identifying positive and negative instances.

Specificity[8] is the number of true negatives divided by the number of true negatives plus the number of false positives. It measures how many of the true negatives have been identified as negative. High specificity means that, for each level, most of the tweets which should be excluded from the class are correctly excluded. Specificity for this classifier is high at 0.873. This is expected given that for each level there are approximately twice as many tweets which are not in that level. Thus, to an extent, high specificity is an artefact of the design of the classifier. Nonetheless, this is important for the applying the classifier to empirical data as it ensures results can be considered robust.

[8] Specificity is also known as the 'True negative rate'

5.4.2 | Multi-class classifier's performance on unseen data

To check the multi-class classifier's performance against real data 'in the wild' it is applied to an unseen dataset of 109,488 tweets produced by 45 far right Twitter accounts during 2017 (see Section 1 of Chapter 4 data collection overview). Then, 100 tweets are randomly sampled from each of the classes (none, weak and strong Islamophobia) to create a combined dataset of 300 tweets. This is annotated blind by the three annotators used to annotate the original training dataset, using the annotation guidelines developed for this Chapter. For all tweets I take the majority of decision to decide the annotation (in 95% of cases all three annotators are in perfect agreement). The results of this testing, and how well it compares with the previous cross-validation testing, are shown in Table 14.

	Accuracy	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**	**Specificity**
Tested on unseen data	0.773	0.830	0.778	0.773	0.776	0.887
Difference with ten-fold testing	+ 0.027	+ 0.023	+ 0.039	+ 0.032	+ 0.036	+ 0.014

Table 14, Performance of multi-class classifier on unseen data

Interestingly, the classifier performs equally well – if not slightly better – across all of the metrics on the unseen data; accuracy is 77.3% and balanced accuracy is 83%. The small uplift in performance indicates the robustness of my approach and its generalizability, which is most likely due to my selection of theoretically-informed input features. These results compare well with previous research, indicating the classifier can be in the empirical part of the present work as an explorative application. In particular, precision (0.778) is considerably higher than the minimum threshold of 0.7 for applying classifiers in empirical work recommended by van Rijbsergen (as reported in (van

Rijsbergen, 1979; Williams & Burnap, 2016)). The classifier's performance compares well with previous studies. For instance, on similar tasks, Malmasi and Zampieri achieve accuracy of 78% but on a dataset in which over half the values are non-offensive, Burnap and Williams achieve accuracy of 77% and Kumar et al. an F1 score of 0.64 (Kumar et al., 2018; Malmasi & Zampieri, 2017; Williams & Burnap, 2016). Davidson et al. achieve higher performance (including an F1 score of 0.90) but on a far more lopsided dataset.

To provide additional insight into the performance of the multi-class classifier relevant metrics are calculated for each of the three levels (none, weak and strong). Accuracy is not reported for each level as this is the same as precision. The results are reported in Table 15. Across nearly all metrics, performance is best for the category of none Islamophobia, followed by strong and then weak. Noticeably, the F1 score is considerably lower for weak (0.683 compared with 0.839 and 0.793) This indicates that the model performs less well at correctly identifying weak Islamophobia. This is also reflected in both the lower precision (0.687), which indicates that many instances of weak Islamophobia are mis-classified as either none or strong, and the lower recall (0.680), which indicates that many instances identified as weak are actually either none or strong.

Islamophobia	**Balanced accuracy**	**Precision**	**Recall**	**F1 score**	**Specificity**
None	0.890	0.778	0.910	0.839	0.870
Weak	0.762	0.687	0.680	0.683	0.845
Strong	0.837	0.869	0.730	0.793	0.945

Table 15, Performance of multi-class classifier across the three classes on unseen data

The lower performance of the classifier on weak Islamophobia is somewhat expected given that it is the middle category and as such has much overlap with the other two – they, in contrast, are easier to categorize as many tweets are at the extremes (i.e. they are either entirely non-Islamophobic or overtly so). The performance of the classifier on the

unseen dataset of 300 tweets is shown in Table 16. Qualitative investigation of the tweets shows that, in many cases, the None tweets express hatred and prejudice against other groups, such as immigrants. Some also discuss Muslims and Islamic practices but without expressing any negativity. Such tweets likely have similar input signals to the classifier, making them hard to separate. In particular, the few cases which are severely misclassified (i.e. the 4 none in strong and the 1 strong in none) are all cases with considerable ambiguity: the 4 None all express hatred against other groups and the 1 Strong uses no profane or aggressive terms. Making nuanced distinctions such as these is a limitation which needs to be further investigated.

		Predicted			
		None Islamophobic	Weak Islamophobic	Strong Islamophobic	
Actual	None Islamophobic	91	22	4	117
	Weak Islamophobic	8	68	23	99
	Strong Islamophobic	1	10	73	84
		100	100	100	**300**

Table 16, Contingency table for the multi-class classifier on unseen data

5.4.3 | Binary classifier's performance on unseen data

For the binary classifier the categories of weak and strong are collapsed into a single category of 'Islamophobia' (as discussed above in Section 5.2). Accuracy is 88.3% and balanced accuracy is 89%. The high performance of the binary classifier is somewhat anticipated given that it collapses the weak and strong Islamophobic classes together, which was the biggest problem with the multi-class classifier. Performance metrics are shown in Table 17. Both recall and precision are high, but recall is noticeably higher than precision (0.95 compared with 0.870). The model has comparatively low specificity (0.777). These results show that few Islamophobic tweets are incorrectly classified as None but that some of the tweets which are classified as Islamophobic are in the wrong category. Put simply, the model is more likely to slightly overestimate rather than underestimate the prevalence of Islamophobia – but not excessively as the F1 score of 0.91 shows. Responding to this problem will be a key focus of future work.

Accuracy	Balanced accuracy	Precision	Recall	F1 score	Specificity
0.883	0.89	0.870	0.950	0.910	0.777

Table 17, Performance of binary classifier on unseen data

5.5 | Conclusion

In this Chapter I outlined the development of two machine learning classifiers to classify Islamophobic hate speech, based on the conceptual work undertaken in Chapter 4. As such, the additional research goal has been met:

> *To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia*

The classifiers were developed using best practice in the field; they were first trained using cross-validation and then tested on an unseen dataset. Three parts of developing a supervised classifier were discussed in detail: (i) creating a training/testing data set, (ii) identifying relevant features, and (iii) selecting the optimal algorithm. In the final section the performance of both the multi-class and binary classifier was reported on. Both classifiers have sufficient accuracy and robustness to be used in the subsequent empirical chapters.

The methodology used here, in particular the attention paid to the development of robust annotation guidelines, is likely to be of interest to other researchers in the field. However, one limitation of the methodology is that the training/testing dataset is relatively small, with only 4,000 tweets. This is due to the labour-intensive nature of the expert annotator coding process. An area for future development is to create a larger and more varied training dataset. This may also affect the type of algorithm that can be used in the model, as more data might enable use of a deep learning classifier, which could increase accuracy. The classifiers developed here are context specific to both the platform (Twitter) and the types of users (followers of UK political parties). As such, they are not necessarily generalisable to other contexts and domains, although the methodology followed could be adapted. This is an important limitation of any hate speech

classification system – it is necessarily bound by time, space and context. In much previous research this is inadequately recognised, with classifiers presented as though they are universally valid. Here, caution is advised. The classifiers can only be used in other research applications if they undergo further validation.

Chapter 6 | Islamophobia and the far right

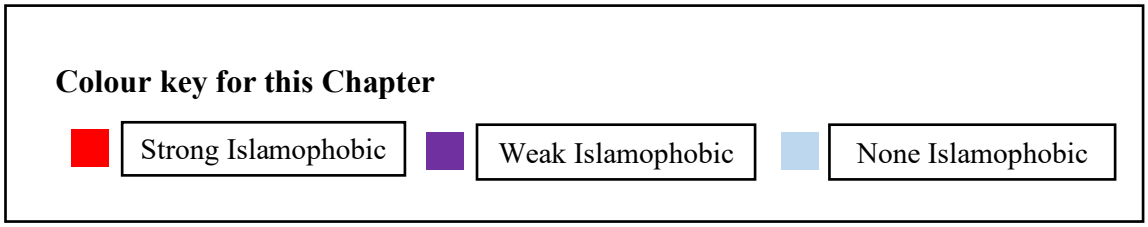
The purpose of this Chapter is to investigate far right social media users, specifically Twitter followers of a far right party (the BNP) to address the second research question in this thesis:

> RQ 2: To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?

The purpose of addressing this RQ is to develop existing academic research apropos the nature of Islamophobia within the far right, and the different ways in which Islamophobia manifests on social media. In particular, the findings feed into existing work on far right radicalization and offer a critical perspective on the view that the far right has created 'walls of hate' on social media (Awan, 2016).

In the first section, I describe the dataset of tweets collected from followers of the BNP. In the second section, I discuss Islamophobia within the far right and analyse the prevalence of Islamophobia. I then develop a typology of Islamophobic users and find that followers of far right parties are far more likely to either never tweet Islamophobically or tweet very Islamophobically. In the third section, I identify six different user trajectories by building a latent Markov model. This shows that far right behaviour is heterogeneous and that users vary considerably in their Islamophobia, from those who are Escalating to De-Escalating Islamophobes, and those who are Never to Extreme Islamophobes. In the conclusion, I discuss the implications of the results, as well as limitations and future extensions.

Throughout this Chapter, I apply the supervised classifier developed in Chapter 5. Use of the colour red indicates strong Islamophobia, purple indicates weak Islamophobia and light blue indicates none Islamophobic.



6.1 | Data overview

The dataset used in this Chapter comprises tweets sent by followers of the far right political party the BNP. The BNP is well-known for its prejudicial views, which have variously been described as Islamophobic (Eatwell & Goodwin, 2010; Wood & Finlay, 2008; Zúquete, 2008), racist (Richardson & Wodak, 2008, 2017), anti-Semitic (Copsey, 2007; Edwards, 2012), homophobic (Commerer, 2010; Severs, 2017), anti-Immigrant (Ford & Goodwin, 2010; John et al., 2004; Margetts et al., 2004) and sexist (Gottlieb, 2004). The party is widely criticised in the British press, which has described it in the past as the 'British *Nasty* Party'. Nonetheless, it is the most successful far right party in the UK, with two Members of the European Parliament (MEP) elected in the nation-wide 2009 European elections and several councillors during the 2000s. It is arguably the only party in the UK's history which challenges the widely held view that the UK is a case of 'far right failure' (Ignazi, 2003), and has been extensively researched in academic literature (Atton, 2006; Brown, 1995; Copsey, 1994; Eatwell, 2006; Eatwell & Goodwin, 2010; Ford & Goodwin, 2010; Goodwin, 2010, 2011, 2013a; Goodwin, 2008; John et al., 2004; Macklin, 2013; Margetts et al., 2004; Renton, 2004),

The BNP has long been dominated by its leader and one-time MEP Nick Griffin, a figure who has received much attention in both academia and the news, often appearing on popular primetime TV shows such as Question Time (Anstead & O'Loughlin, 2011; Edwards, 2012; Goodwin, 2011). He was the BNP's leader from 1999 to 2014, its period of greatest success, after which he was expelled from the party and replaced with Adam Walker. Since the BNP's high in 2009 it has lost a considerable amount of support. At the 2010 general election it won 563,743 votes or 1.9% of the total (although no BNP Members of Parliament (MP) were elected due to Britain's first past the post system); in 2015, the BNP received just 1,667 votes. During this period, prominent figures either left

the party, such as the MEP Andrew Browns and London Assembly representative Richard Barnbrook, or lost their elected positions, such as the 12 out of 13 members of the Barking and Dagenham local council who were BNP members in the mid 2000s. Determining the level of support for the BNP is difficult, particularly as previous research indicates that its constituency of potential or 'latent' supporters may be far greater than the number of current voters and party members (Margetts et al., 2004).

The BNP is an important focus of research into the far right on social media; it is not only the most successful far right party in British history but also one of the most prominent UK-based far right parties on Twitter. Figure 3 shows the number of followers the BNP has, compared with other prominent far right groups (identified from Hope Not Hate's 2015 and 2017 State of Hate reports (Hope Not Hate, 2015, 2017)). Twitter is one of only three social media platforms which the BNP uses (the other two are Facebook (215,971 Likes) and YouTube (11,812 subscribers)).[9] Furthermore, other popular far right groups, including Britain First and the EDL, were banned from Twitter in December 2017 and as such are unsuitable focuses of this study. Equally, For Britain was only founded in October 2017 and thus is also an inappropriate focus. Due to its prominence, longevity and large number of online supporters, as well as the considerable body of relevant offline research, I focus on the BNP as a way of operationalizing RQ 2 and investigating followers of far right parties on Twitter.

[9] Data is collected on 1st November 2018.

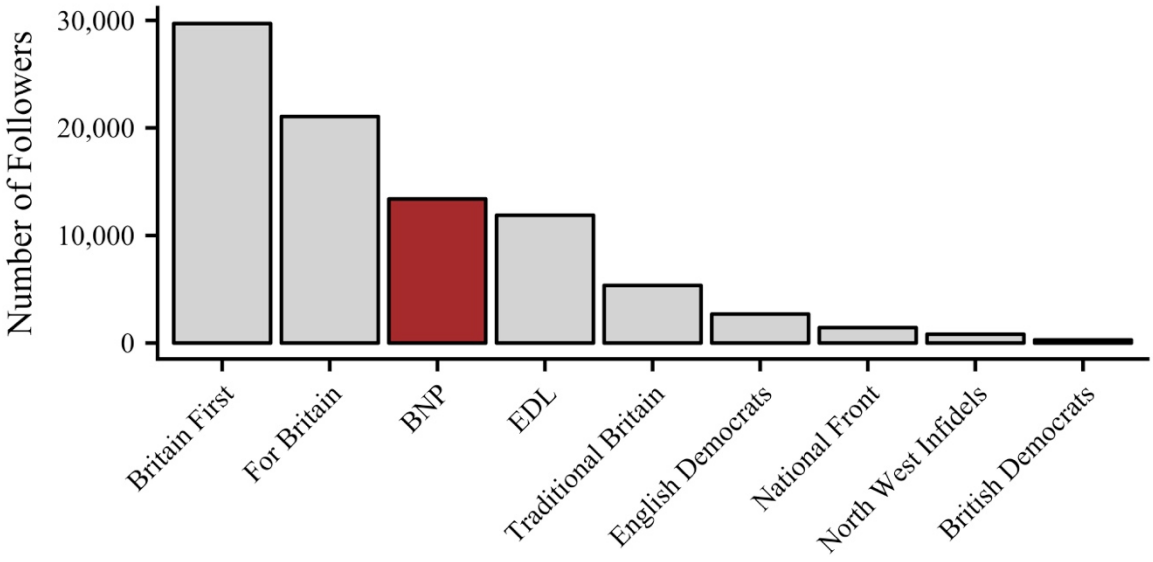


Figure 3, The number of Twitter followers for prominent far right parties[10]

6.1.1 | Dates

For followers of the BNP, a period of one year is studied, from 1st April 2017 to 1st April 2018. One year is chosen to account for any temporal patterns linked to seasonal trends, such as possible increased tweeting around Christmas or New Year (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Li, Goodchild, & Xu, 2013). This period covers several important political events in the UK, including the General Election on 8th June 2017, Local Elections on 4th May 2017, Manchester Arena bombing on 22nd May 2017, London Bridge terror attack on 3rd June 2017 and the progression of the European Union (Withdrawal) Act of 2018 through the UK parliament. All tweets produced by followers of the BNP during this period are collected (in accordance with the data collection process outlined in Chapter 3), which amounts to over 10 million tweets.

[10] Data is collected on 1st August 2018 for the BNP, For Britain, English Democrats, Traditional Britain Group, British Democrats and North Western Infidels. Britain First, the EDL and the National Front were suspended from Twitter on 18th December 2017, and the number of followers recorded for them are from this date.

6.1.2 | User sampling

At the start of the period (1st April 2017) there are 13,002 followers of the BNP and at the end (1st April 2018) there are 13,951. Of the original 13,002 users, 11,785 (90.6%) are still followers at the end (1,217 followers ceased following). I call these users the 'persistent followers'. Given that social media followership patterns can change rapidly, I only focus on these users. During the period, 6,611 of the 11,785 persistent followers (56%) tweet at least once and as such can be considered active during the period studied. These active users send 10,229,137 tweets in total, which is an average of 1,547 tweets each. The standard deviation is 4,429 and the range is 1 to 65,373. This indicates that bots are present in the dataset, which is addressed below. For the remainder of this Chapter, only the active persistent followers of the BNP are studied (n = 6,611). From hereon all references to 'users' and 'followers of the BNP', unless otherwise qualified, refers only to them.

6.1.3 | Language

The 6,611 users tweet in many different languages, which may partly reflect the perceived growing 'internationalisation' of the far right (Doerr, 2017a, 2017b; Macklin, 2013; Mammone, Godin, & Jenkins, 2012). Users set their own language, and it is provided automatically by the Twitter API. The number of tweets produced in each language is shown in Appendix 6.1. English is the dominant language, accounting for 76.8% of the tweets. The second most prominent language is 'Undetermined', which is an option selected by users who do not want to explicitly state their language. 'Undetermined' accounts for 618,952 tweets. Manual inspection shows that 'Undetermined' overwhelmingly consists of tweets in English. I only keep tweets in 'English' and 'Undetermined' given that the Islamophobia classifier developed in Chapter 5 is tuned

only for English language. The size of the dataset reduces by 1,749,762 tweets to 8,479,375. Removing just non-English *tweets* ensures that users who produce a mix of both English and non-English language tweets are kept in the dataset. Nonetheless, 77 users – who only tweet in languages other than English and Undetermined – are removed, reducing the number of users from 6,611 to 6,534.

6.1.4 | Bots

After non-English language tweets have been removed, the average user sends 1,298 tweets during the period. The average number of tweets per user is likely inflated by the presence of highly active users, many of which are likely to be 'bots'. Bots can be defined as a 'computer algorithm that automatically produces content and interacts with humans on social media' (Davis, Varol, Ferrara, Flammini, & Menczer, 2016, p. 1) – although many researchers also note that bots are not always purely automated but can involve a degree of human input, particularly with regard to content creation and user-interaction (Mønsted, Sapieżyński, Ferrara, & Lehmann, 2017). Bots often do not act on their own and so are best understood in terms of 'bot nets'. These comprise several thousands of remotely controlled accounts which can simultaneously focus their efforts on a single online interaction or bit of content (Soltani et al., 2014).

Bot detection is a notoriously difficult challenge given the sophistication of bot strategies and the ethical limitations of many bot detection methodologies (Thieltges, Schmidt, & Hegelich, 2016). With regards to social media, different approaches for bot detection have been adopted. Davis et al. define a supervised machine learning algorithm, called 'bot or not', which assigns users a probability based on how likely it is that they are a bot (Davis et al., 2016). This method draws on 1,000 input features, which can be grouped into six categories: network, user, friends, temporal, content and sentiment. This method,

which the authors have made available open-source, has been used in other works and marks an important step forward in studying Twitter bots (Varol, Ferrara, Davis, Menczer, & Flammini, 2017; Woolley & Guilbeault, 2017). However, it is possible that their method over-estimates the prevalence of bots. Of a sample of 900,00 users whose 'botiness' was evaluated, over 90% of users had a 0.75+ probability of being a bot. More broadly, it is difficult to evaluate the method's effectiveness given that there does not a exist a single 'gold standard' dataset and because the details of the algorithm (including how the 1,000 input features are engineered and weighted) is not made available.

Kollanyi et al. define a rule-based approach whereby accounts which post at least 50 times per day are deemed highly automated (Kollanyi et al., 2016). This approach is crude but effective. It recognises the difficulty of distinguishing between (i) bots and (ii) genuine users with idiosyncratic or somewhat irregular features and behavioural patterns (Larsson & Hallvard, 2015). Rather than seek to make such fine-grained distinction, this approach specifically targets just one type of bot (highly active ones). Indeed, the purpose of this method can be best understood as not just the removal of highly automated bots but, rather, the removal of all highly active accounts – irrespective of whether they are fully automated bots, semi-automated accounts or hyper-active genuine users. Removing such users is important because they often tweet atypically and can bias statistical analyses due to their large volume. That is, the behaviour and dynamics of the small number of users who tweet in high volumes may be entirely unrepresentative of the rest of the cohort – but, because they have such high volume, these users disproportionately impact the overall analyses. This is well shown by Axel and Stieglitz's study of metrics to analyse Twitter data, in which the top 1% of active users often accounted for over 50% of all the tweets sent (Bruns & Stieglitz, 2013).

One limitation of rule-based approaches is that bot-owners can respond to them and set bots to tweet just below the limits.[11] The choice of 50 is also entirely arbitrary – and as the distribution of tweets per bots is likely scale-free, it is not possible to statistically identify a threshold from the data alone. I opt for a lower threshold than Kollanyi et al. and set it to 40 tweets per day per user (14,600 in total during the period studied). This is appropriate given that the goal is to remove not just bots but also all high-volume tweeters. 128 users meet this bot-detection criterion (1.96% of the total), tweeting 23,191 times each on average. Note that many of the users which have already been removed from the previous sampling criteria are likely to have been high volume and semi-automated tweeters, such as bots. Removing the 128 accounts categorised as bots reduces the number of users by 1.95% to 6,406. The number of tweets in the dataset reduces considerably by 35% from 8,479,375 tweets to 5,510,893.

6.1.5 | Data summary

The final BNP dataset covers on year (1st April 2017 to 1st April 2018) and comprises 5,510,893 tweets sent by 6,406 users.

[11] This point was made to be my Sam Woolley, one of the authors of the Kollanyi et al. paper, in a private conversation.

6.2 | Islamophobia within the far right

The goal of this section is to understand the prevalence of Islamophobia across followers of the BNP. I use the classifier outlined in Chapter 5 to annotate all tweets in the dataset (n = 5,510,893). The four plots in Figure 4 visualize key aspects of the dataset.

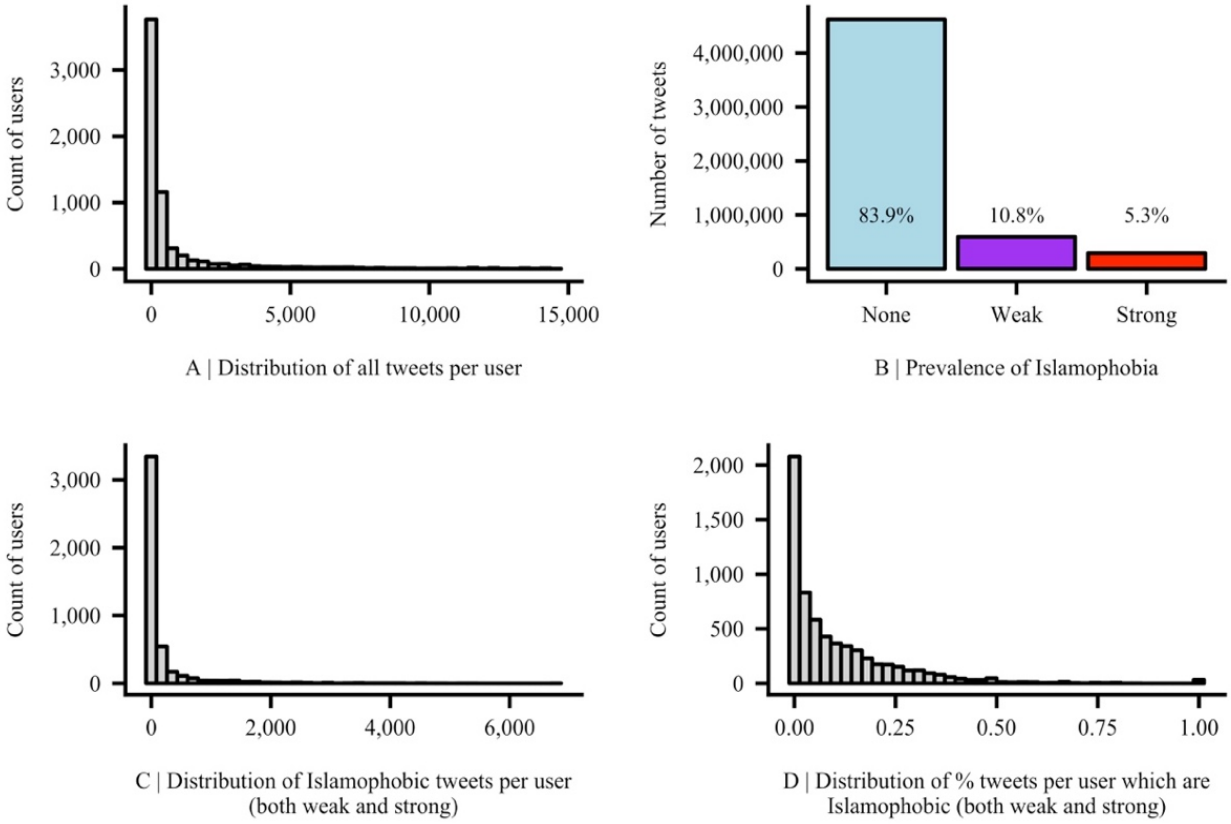


Figure 4, (A) Number of tweets per user, (B) Prevalence of Islamophobia, (C) Number of Islamophobic tweets per user and (D) Probability of a tweet being Islamophobic per user

Panel A shows the distribution of all tweets per user. The maximum number of tweets is curtailed at 14,600 due to the sampling process outlined in Section 1 above. The distribution is long-tailed. Panel B shows the prevalence of Islamophobia within all tweets (calculated using the binary classifier). The overall prevalence is 16.1%, split between 10.8% weak and 5.3% strong. This indicates that over twice as much of the far right Islamophobia on Twitter is subtle, nuanced and specific than either highly general

or highly negative. This is a surprising finding given previous research in both online and offline contexts which characterises Islamophobia in the far right as vitriolic and aggressive. It also highlights the importance of distinguishing between different strengths of Islamophobia in empirical research.

For panels C and D, the binary classifier is used, whereby strong and weak Islamophobic tweets are collapsed into a single category. Panel C shows the distributions of Islamophobic tweets per user, and panel D the distribution of the percentage of tweets per user which are Islamophobic. Panel C and D show that the overall prevalence of Islamophobia (reported in Panel B as comprising 16.1% of all tweets) is driven largely by a small number of highly Islamophobic users. This is also reflected in Figure 5, which shows the number of Islamophobic tweets versus the total number of tweets sent by each follower of the BNP. Note that the axes are logarithmic and as such users who do not send any Islamophobic tweets (n = 1,843) are not shown. A large number of users send a high volume of Islamophobic tweets and for many users Islamophobic tweeting constitutes a large *percentage* of the total number of their tweets.

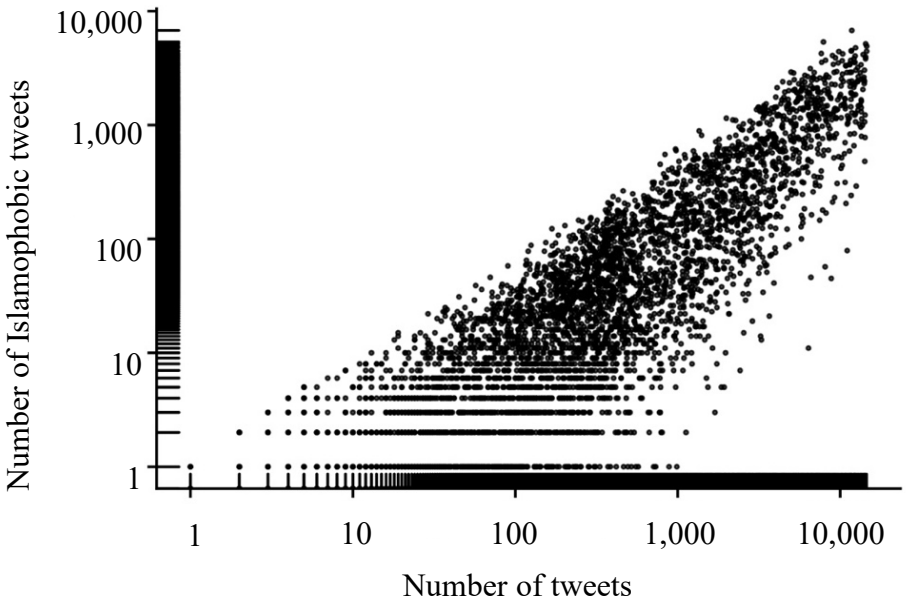


Figure 5, Number of Islamophobic tweets versus the number of tweets for followers of the BNP

6.2.1 | Typology of Islamophobic users

To gain better insight into the different types of Islamophobic users I establish a typology of 7 different types of users based on their tweeting behaviour. This typology is a useful coarse-grained tool for understanding the dynamics of Islamophobic behaviour, in terms of differences between users, within the far right. The 7 types are based on the three levels of the multi-class classifier (none, weak and strong Islamophobic tweets) and constitute a collectively exhaustive and mutually exclusive set. The seven types are:

1. None only: Users who only tweet none Islamophobically
2. Weak only: Users who only tweet weak Islamophobically
3. Strong only: Users who only tweet strong Islamophobically
4. None and Weak: Users who only tweet both none and weak Islamophobically
5. None and Strong: Users who only tweet both none and strong Islamophobically
6. Weak and Strong: Users who only tweet both weak and strong Islamophobically
7. None, Weak and Strong: Users who tweet none, weak and strong Islamophobically

Identifying which type each user belongs to is a three-step process. First, the number of tweets in each level of the multi-class classifier is counted. Second, the count is reduced to a binary evaluation showing *whether* or not the user has sent a tweet in each level. Third, the levels which the user has sent at least one tweet within are checked, and from this the user is assigned to a type. For instance, if a user sent just one strong Islamophobic tweet and one weak Islamophobic tweet they would be considered part of type 6 (users who tweet both weak and strong). If another user sent just one strong Islamophobic tweet and 65 weak Islamophobic tweets they would also be considered part of type 6. Thus, the minimum number of tweets a user must send in each level to be considered representative of that level is just one. As there are potentially three levels a user can tweet in (none, weak and strong), I only include users in the sample if they have tweeted at least three times, which reduces the number of users from 6,406 to 6,018. This approach simplifies the actual tweeting patterns of each user, reducing them to just seven different types. Each user is assigned to one, and only one, type.[12]

To assess how many users are assigned to each type I calculate (1) the actual number of users in each of the seven types (using the three-step process outlined above) and (2) the number of users in each type based on a random expected distribution. This allows me to assess statistical significance, and whether the observed values indicate that users' behaviour is driven by particular concerns and dynamics. I calculate the random counts by taking (i) the probability that each tweet falls into one of the three levels (none, strong and weak Islamophobic), which is given by the overall prevalence of each level, and (2)

[12] Note that I adopt this approach out of parsimony. It is not possible to compute a single score for each user – for instance, I cannot meaningfully claim that the magnitude of 'Strong' Islamophobia is twice as great as 'Weak' Islamophobia. As such, the values for these separate levels cannot be combined within a single scale.

the number of tweets produced by each user. For each user the probability that they fall into each of the seven types is then calculated. The sum of the probabilities is 1 as each user has tweeted at least once. These probabilities are summed over all of the users to calculate an overall distribution. An example calculation is shown in Appendix 6.2. The results of comparing the empirical distribution of users with the expected random distribution is shown in Table 18. Note that I only calculate values for users who send at least 3 tweets (n = 6,018).

Type	Actual number of users (3 + tweets only)	Expected number of users (random)	Difference, actual vs. expected	0.95 Confidence interval for the expected numbers[13]	Information about the type			
					% strong	% weak	% none	Number of tweets per user
None only	1,494 (24.82%)	402 (6.68%)	1,092 +272%	364 – 440 *sig*	0%	0%	100%	38
Weak only	7 (0.12%)	0 (0%)	7 +600%[14]	0 *sig*	0%	100%	0%	4
Strong only	3 (0.05%)	0 (0%)	3 +200%[6]	0 *sig*	100%	0%	0%	4
None and Weak	889 (14.77%)	643 (10.68%)	246 + 38.3%	596 – 690 *sig*	0%	2.7%	97.3%	136
None and Strong	181 (3.00%)	208 (3.46%)	-27 - 7.7%	180 – 236 N.S.	3.3%	0%	96.7%	52
Weak and Strong	1 (0.017%)	1 (0.017%)	0	0 – 3 N.S.	57.1%	42.9%	0%	5
None, Weak and Strong	3,443 (57.21%)	4,764 (79.12%)	-1,321 -27.73%	4,702 – 4,825 *sig*	5.5%	11.1%	83.4%	1,546
TOTAL	6,018	6,018						

Table 18, Comparison of expected and actual number of users for each type

Table 18 shows some important divergences between the actual number of users in each type and the expected number based on the random calculation. Comparing the actual and expected numbers of users in each type is important to show the extent to which user patterns are following a social dynamic rather than just a random distribution of behaviours. Five of the seven types are significantly different as the values fall outside the 0.95 confidence interval. Noticeably, the number of users who only tweet in None

[13] Confidence intervals are calculated by taking the standard error (the variance divided by the square root of the number of instances) and approximating the categorical with a normal distribution, using z = 1.96 for a 95% confidence interval, where the confidence interval equals the Expected value +- z * standard error. The variance for each category is given as p * (1 – p) where p is the probability of that category occurring.

[14] Calculations are based on rounding up the expected values to 1.

only is far *higher* than anticipated with a difference of 1,092 users (1,494 compared with 402), a 272% increase. Equally, the number of users who tweet in None, Weak and Strong is far *fewer* than anticipated, with a difference of 1,321 users (4,764 compared with 3,443), a 27.73% decrease. For other types of users (including Weak only, Strong only, None and Weak, and Weak and Strong) there are more actual users than expected in the random calculations, of which just Weak and Strong is not significantly different. This suggests that far right users on Twitter are far more *heterogeneous* than anticipated and that more of them are on the extremes; more users tweet Islamophobically and more users do not tweet Islamophobically. This means that the far right cannot be characterised simply using the overall prevalence of Islamophobia as there are important user-level variations in how users behave. The results of this analysis, in which the expected and actual numbers of users in each type are compared, is also presented visually in Figure 6.

An additional finding is that very few users *only* engage in Islamophobic behaviour; the Strong only, Weak only and Weak and Strong types account for just 11 out of the 6,018 users studied, less than 0.25% - and these users typically send very few tweets (on average, just 4.5 across all three types). Thus, even within the far right very few users are *solely* Islamophobic. This indicates that more attention needs to be paid to understanding the internal variations within the None, Weak and Strong type, as this is both the most prevalent type (comprising 3,443 users) and most active (sending 1,546 tweets on average).

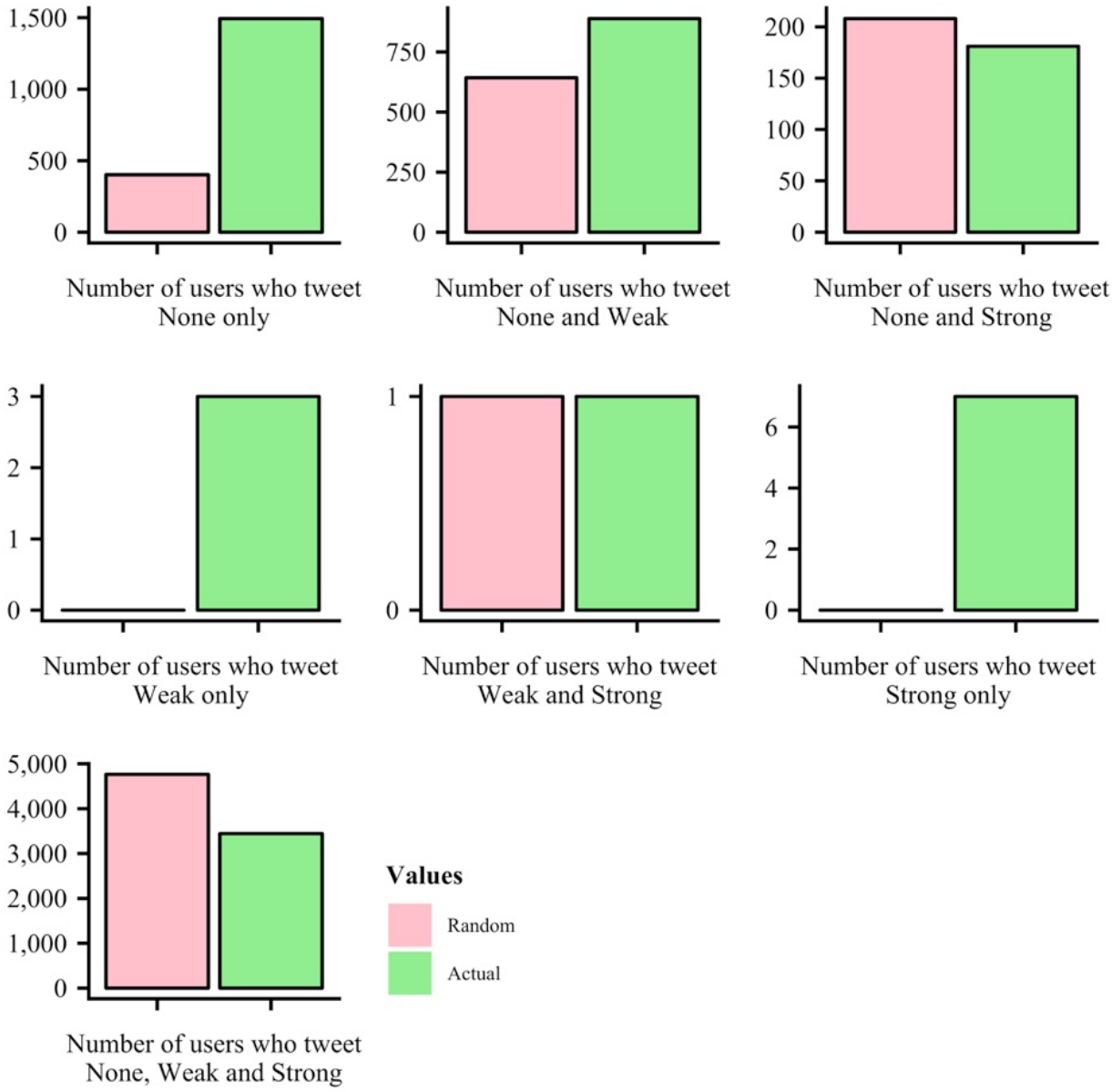


Figure 6, The random and actual number of users for the seven different types of user behaviour

6.3 | Trajectories of Islamophobia

The previous section demonstrates that, when considered over the whole period, users are highly heterogeneous with regards to their Islamophobic tweeting. Building on this finding, in this section I study how users' Islamophobic tweeting varies over time to identify different *trajectories* of Islamophobia within the far right. For instance, two users might both be in type 7 (tweeting none, weak and strong Islamophobically). However, they could exhibit very different trajectories, whereby one user's Islamophobia is escalating over the period and the other is de-escalating. Furthermore, users in different types – for instance, users in type 3 (strong only) and type 6 (weak and strong) – might exhibit very similar overall trajectories of behaviour, even though they send tweets in both levels of Islamophobia. It is important to take these differences into account to better understand the nature of Islamophobia within the far right. Accordingly, in this section I model different trajectories of user behaviour, using the timestamps of the annotated tweets to capture longitudinal changes.

6.3.1 | Statistical modelling

Longitudinal data poses several difficulties for statistical modelling. Common approaches include the use of fixed effects and random effects models. In this case, the data consists of a time series in which a single item (the degree of Islamophobia, covering none, weak and strong manifestations) is measured repeatedly. The goal of the modelling is to uncover *different* trajectories for the users, using the single item time series. As such, two types of model are particularly well-suited, both of which model single item time series and include different latent states within the model: Growth curve mixture modelling and latent Markov chain modelling. Both can be used to separate users into different classes. I opt for latent Markov chain modelling as this is better-suited to

categorical variables for handling large volumes of users (Curran, Obeidat, & Losardo, 2011).

Latent (or 'hidden') Markov (hereafter known as 'LM') modelling is an extension to the traditional Markov chain model. A Markov chain model is a stochastic model which can describe a sequence of events in which the probability of each event depends *only* on the previous event. For this reason, Markov chains are described as having 'memorylessness' because they satisfy the 'Markov property' whereby the conditional probability of future states depends only on the present state. A one-dimensional random walk is an example of a Markov chain; in a random walk, the future position of the walker depends only on its current position rather than its prior behaviour (Spedicato, Kang, Yalamanchi, & Bhargav, 2017). The traditional Markov chain model estimates the behaviours that users are likely to engage in at each time point, providing a probability for each behaviour. All users are part of one overarching class, and as such just one probability is calculated for a given point in time. The LM model extends this by assuming the existence of K latent states. The latent states must be defined in advance and must be a finite (i.e. countable) number (Spedicato et al., 2017). The LM model then estimates both the behaviours associated with each latent state and the transitional probabilities between states. Thus, for a user in a given latent state at time t_j the model estimates the probability that they will engage in a particular behaviour (e.g. sending a none, weak or strong Islamophobic tweet) – which is based upon the latent state they are in – and the probability that their state will change at time t_{j+1} (which is commonly described as a 'regime change'). In most LM models, the transitional probabilities are time homogeneous and so do not change as time passes. This assumption can be adapted for different use cases to create more complex models (Bartolucci, Farcomeni, & Pennoni, 2010). The parameters in Latent Markov models are typically estimated using maximum likelihood estimation via

the expectation-maximization algorithm (Bartolucci et al., 2010; Bartolucci, Pandolfi, & Pennoni, 2015).

As with many statistical models, the assumptions of the Markov chain are not necessarily 'true' in that they are unlikely to accord with the realities of human behaviour. In particular, the primary assumption is unlikely to hold; it is very unlikely that users' Islamophobic tweeting behaviour is genuinely memoryless. However, hidden Markov models are remarkably effective at approximating users' behaviour. For instance, a study by Druce et al. used LM modelling to study how users engaged with a medical monitoring app as part of the 'Cloudy with a chance of pain' public health project. Druce et al. identified different trajectories of user behaviour, including 'tourists', 'low', 'moderate' and 'high' engagers (Druce, McBeth, et al., 2017; Druce, Veer, et al., 2017). Within these clusters, the researchers found different levels of reported pain, medical conditions and treatment experiences. It is likely that users' behaviour is not memoryless, in that their past experiences of using the app – and the habits they develop associated with the app – likely inform their future engagement with it. However, using the Markov chain model it is possible to identify different patterns and latent states of behaviour.

LM modelling is a highly robust and well-established method for handling longitudinal data. It has been developed for categorical outcomes, and as such is highly suitable to the dataset used here. LM models can suffer from dependency issues with multivariate data (such as when separate dependent variables are measured and latent states identified for different combinations of them) but this is far less of a concern with univariate outcomes, as is the case here (Song, Xia, & Zhu, 2017). LM models account for (i) how users shift *between* states and (ii) how their behaviour varies *within* states. This means they are highly appropriate for capturing different trajectories of Islamophobia, and as such will enable me to answer the RQ addressed in this Chapter. Specifically, I will be able to

identify heterogeneity of Islamophobia within the far right, including pathways towards and away from extremist behaviour. LM modelling can also be used to calculate the number of users assigned to each trajectory, and as such their prevalence. This is crucial for not only seeing how users vary but also which variations are the most prominent.

6.3.2 | Fitting the latent Markov model

In this section I outline the three steps undertaken to fit the LM model: (1) measuring time, (2) measuring user behaviour and (3) fitting the number of latent states. I justify the choices made for each of these three steps as they are crucial inputs into the LM model. Additional details are provided in Appendix 6.3

6.3.2.1 | Measurement of time

Studying users' behaviour on Twitter longitudinally poses a considerable difficulty in that users not only vary qualitatively (i.e. with regard to the extent of Islamophobia that they express) but also quantitatively (i.e. with regard to the temporal dynamics of how frequently and regularly they tweet). Unlike most longitudinal studies, where individuals are measured at pre-defined equally spaced intervals, in this case individuals express Islamophobia at very different times. The actual timestamps of tweets cannot be used as this would create a LM model with millions of different 'events', few of which line up with each other. One solution to this problem is to use a coarse-grained time window, such as 1 day, which agglomerates the tweets sent within window. However, this risks introducing considerable bias into the results due to the varying volume of tweets sent by users over time (i.e. on some days the volume of tweets is high but on others it is low). As such, I opt to scale the time period by the *overall* volume of tweets. Note that three alternative strategies are discussed in detail in Appendix 6.3.

Scaling time by the overall volume of tweets is implemented by taking the total number of tweets sent by all users (in this case, 5,510,893 tweets) and dividing it into *T* periods. For instance, if *T* is set to 100 then each time period *t* consists of 55,109 tweets. The linear time periods that *t* covers will vary according to how active users are. For this data, setting *T* to 100 results in values of *t* which range from 1.7 days to 8.7 days. Then, users' tweeting behaviour within each time period *t* is measured. This approach is counter-intuitive but ensures that (i) the number of time periods without a value is minimized as users are, in effect, afforded more time to send a tweet in periods when the overall volume of tweeting is low and (ii) users are compared across the same time intervals; t_x covers the same time period for every user – it is just that the linear length of t_x is not the same as the linear length of t_{x+1}. I opt to use this approach as it is the best-suited for taking into account the varying volume of tweets sent across time, thereby ensuring that users' behaviour can be meaningfully compared at separate time periods. One limitation is that, as with all the approaches considered here, the choice of *T* is likely to have a considerable impact on the ensuing statistical analyses but is also inherently arbitrary. A model with a smaller value of T is less nuanced but is also less susceptible to one-off events which may lead to temporary variations in users' behaviour. To account for this, I fit four separate models with different values of *T*: 10, 25, 50 and 100 (reported in Appendix 6.3 and discussed below).

6.3.2.2 | Measurement of Islamophobia

The second consideration with studying time is how to measure users' behaviours within each time period. This is a difficult task because the dependent variable (Islamophobic tweeting) is ordinal but not interval. Strong Islamophobia is not a multiple of weak Islamophobia and nor are weak and strong situated on the same probability spectrum. This reflects the conceptual arguments made in Chapter 4 regarding the differences

between weak and strong Islamophobia, which are embodied in the classification work in Chapter 5. Weak and strong Islamophobia are *not* determined by a single probability, whereby tweets with a low-to-medium probability of containing Islamophobia are classified as weak and tweets with a high probability are classified as strong. Instead, weak and strong are separate varieties of Islamophobia, each of which are assigned a *separate* probability. Accordingly, a mean value cannot be computed for each users' behaviour in any given time period *t*.

I measure Islamophobia by taking each users' strongest single expression of Islamophobia in any time period. I treat this as representative of that users' behaviour (for that time period). For instance, if a user sends at least one tweet that is strong Islamophobic during t_x then that is how their behaviour is characterised for t_x. If they send at least one weak Islamophobic tweet but none strong then their behaviour is characterised as weak. It is only characterised as none if they send no weak or strong tweets. This strategy is a simple solution which ensures that strong Islamophobic tweets are well represented in the LM model. It is also theoretically robust since what is of greatest interest is *whether* users have engaged in Islamophobic behaviour rather than whether the *majority* of their behaviour is Islamophobic. Accordingly, I opt to use this measurement strategy. In practice, this means that if a user is measured as sending strong Islamophobic tweets at a given time period *t* this means only that they have sent at least one strong Islamophobic tweet during *t* and not that the majority of their behaviour is Islamophobic. This is discussed further in Appendix 6.3.

An additional issue is that, even though I am using a varying time period scaled by the overall volume of tweets, it is likely that some users will have time periods when they do not send any tweets. Rather than treating time periods without any tweets as missing data, I assign them a score of none Islamophobic. An alternative would be to establish a four-

tiered categorisation whereby users can be in one of four states: not tweeting, not Islamophobically tweeting, weak Islamophobically tweeting and strong Islamophobically tweeting. I opt not to do this as users who do not send any tweets in a time period are still not engaging in Islamophobia. The impact of this decision is shown in Figure 7. The left-hand panel shows the behaviour of a random sample of users where periods in which they have not tweeted are left blank and the right-hand panel shows the same users' behaviour where periods without tweets have been labelled none Islamophobic.

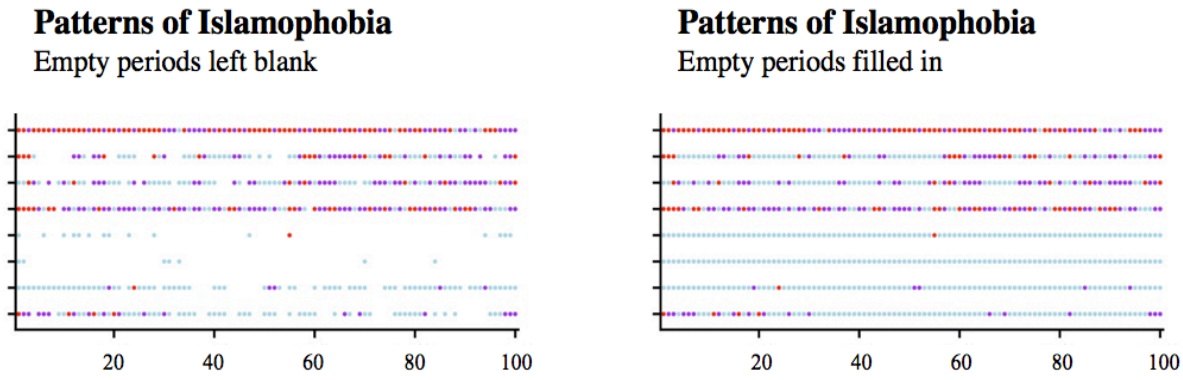


Figure 7, Patterns of Islamophobia measured over 100 time periods with missing tweets filled in as none Islamophobic

6.3.2.3 | Number of states

The LM model fits *K* latent states to the data. As discussed above, the number of latent states must be provided as an input to the model. *K* can be optimized by evaluating model quality on the data using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC are widely-used metrics for evaluating model quality in statistics and data science, and are well-established for fitting LM models (Bartolucci et al., 2010). They take into account both models' goodness of fit and their simplicity, penalising more complex models which have more parameters. One risk with using AIC

and BIC is that they are likely to suggest a larger value for the optimal number of states, and as such are best interpreted as an upper limit or maximum on the value of *K* (M. J. Green, 2014). The model fitting process and results are described in detail in Appendix 6.3.

I fit LM models for 1 to 6 latent states for time periods (*T*) of length 10, 25, 50 and 100. Varying the length of T ensures that the number of time periods, which is an arbitrary input, does not bias the selection of typified user trajectories. In Appendix 6.3, I show analysis which demonstrates the final results – the typified user trajectories discussed above – are reasonably consistent across varying values of T. This analysis is not provided in the Chapter for brevity. For $T = 10$, the optimal number of latent states is 3. Note that it is a coincidence that this matches the 3 types of behaviour (None, Weak and Strong Islamophobic tweeting): it is *not* a requirement of the model that the number of behavioural and latent states align. For models where T is higher (25, 50 and 100) The optimal number of latent states increases, roughly proportionally with the increasing number of time periods, despite the penalties imposed by both AIC and BIC. Thus, the number of latent states cannot be fixed for all models in advance. In the remainder of this Chapter, I show the results for a model with $T = 10$.

6.3.3 | Analysis of LM model

I fit an LM model with 10 time periods and for each user calculate their most probable state at each time period, thereby creating a new and simplified tweeting behavioural trajectory for each user. The model is implemented in R using the 'LMest' package, and the convergence tolerance is set to 1e-10 (Bartolucci et al., 2015). It is worth restating here the basic problem which this method overcomes; that users exhibit highly heterogeneous patterns of behaviour. Figure 8 shows the behaviour of a random sample of 30 users. Visually, the users exhibit very different patterns of behaviour, and it is not easy to separate them into different trajectories through manual inspection.

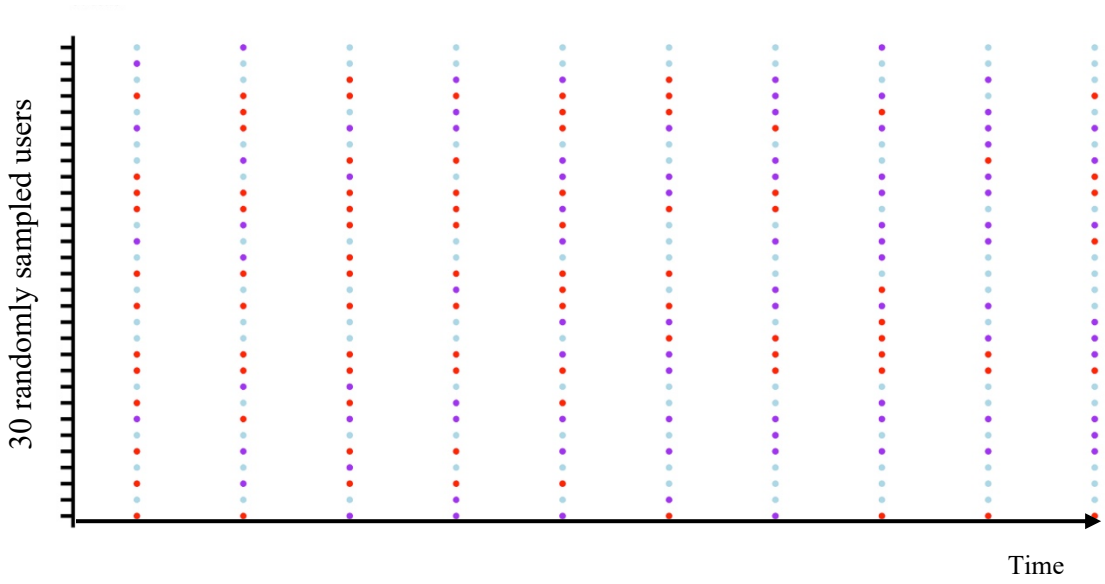


Figure 8, Behavioural patterns of 30 randomly sampled users

6.3.3.1 | Latent state probabilities

For each latent state, the LM model estimates the probability that users will engage in any of the three types of behaviour (strong Islamophobic tweeting, weak Islamophobic tweeting and none Islamophobic tweeting) and the probability of a user in that state transitioning to another. Then, for each user, the LM model estimates their latent state at each time period *t*. The behavioural probabilities for each of the three latent states are

shown in Table 19. Two of the latent states match closely with one of the three types of behaviour; users in latent state 1 have a 0.86 probability of engaging in none Islamophobic behaviour and users in latent state 3 have a 0.95 probability of engaging in strong Islamophobic behaviour. Latent state 2 is more mixed. The most probable behaviour is weak Islamophobia (0.44) but the other two states are also highly probable (0.19 for none and 0.37 for strong). This suggests that users in the middle state, whilst most likely to be weakly Islamophobic, may also exhibit behaviour at both extremes. This is understandable given that weak Islamophobia can be considered a more varied state, in which users are less committed to either fully Islamophobic or non-Islamophobic behaviour.

	Latent state 1	**Latent state 2**	**Latent state 3**
None Islamophobia	0.86	0.19	0.02
Weak Islamophobia	0.09	0.44	0.03
Strong Islamophobia	0.05	0.37	0.95

Table 19, Probability for each latent state of engaging in different types of behaviour

The utility of the LM model is that at each time point users can be in the same latent state but not necessarily exhibit the same type of behaviour. For instance, one user may tweet in each time period strong Islamophobically, which could be represented as in the left-hand plot of Figure 9. Another user might tweet strong Islamophobically most of the time but sometimes also weak Islamophobically, as shown in the right-hand panel of Figure 9. Even though the users' manifested behaviour is different their latent states might be the same. Even though there are small differences in their actual behaviour, both users can be classified on the same trajectory. The LM model identifies that these users have

similar latent states – which underpin their varied behaviour – thereby reducing the amount of variation in the data.

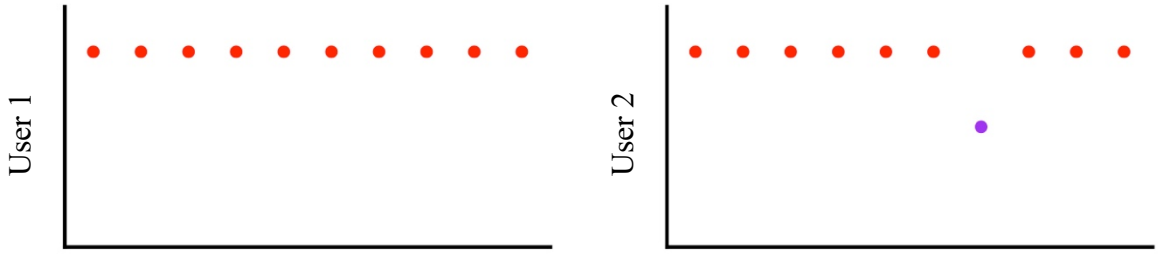


Figure 9, Example behaviour of two simulated users

After reviewing the results of initial LM model testing, I decide to separate 1,843 users who do not send any Islamophobic tweets. Initial LM models correctly assign these users to a 'none Islamophobic' latent state across all time periods. However, the models also assign users who send a few Islamophobic tweets to this state. For some empirical applications this might be a reasonable assumption (i.e. very low levels of an activity might indicate an unwillingness or disinterest in engaging in that behaviour). Nonetheless, it is inappropriate for studying Islamophobia as even one Islamophobic tweet is concerning. Although I remove these users from the LM model, I re-integrate them into the analysis below as a separate 'Never Islamophobic' typified user trajectory. Thus, the LM model is run not on the full 6,406 users but on a subset of 4,563 users.

6.3.3.2 | User trajectories

I cluster the user behavioural trajectories, based on their latent states in the LM model, into *typified* user trajectories using the k-modes clustering algorithm (Huang, 1998), which is an extension of the widely used k-means algorithm, and has been used in previous research using LM modelling (Druce, McBeth, et al., 2017). Effectively, this projects time into a spatial dimension: for the purpose of clustering, user *i* at time *j* is modelled as user *i* in dimension *j*. I provide additional details on the algorithm and fitting

the number of clusters in Appendix 6.3. I fit the latent states in the LM model into five typified user trajectories, which best balances generalisability and specificity. I add a sixth user trajectory to the five identified from the latent states in the LM model; never Islamophobic. I name the six user trajectories based on the frequency, magnitude and regularity of Islamophobic behaviour they exhibit.

The probabilities for the *latent states* for the typified user trajectories are shown in Figure 10. Each panel shows a different typified user trajectory and the probabilities associated with each latent state at each time period. In line with the colours used throughout this Chapter and the previous one, blue represents the none Islamophobic latent state, purple the weak Islamophobic latent state and red the strong Islamophobic latent state. The tone of each colour represents the probability assigned to it, whereby stronger tones are more probable (for each time period, the probabilities sum to 1). In Appendix 6.3, Section 6, I present two additional plots for the typified user trajectories; (1) the probabilities for the *behaviours* associated with each time period and (2) the empirical *prevalence* of the behaviours in each time period. All three figures show the same overall pattern for each typified user trajectory, which indicates that the six trajectories capture meaningful differences in user behaviour.

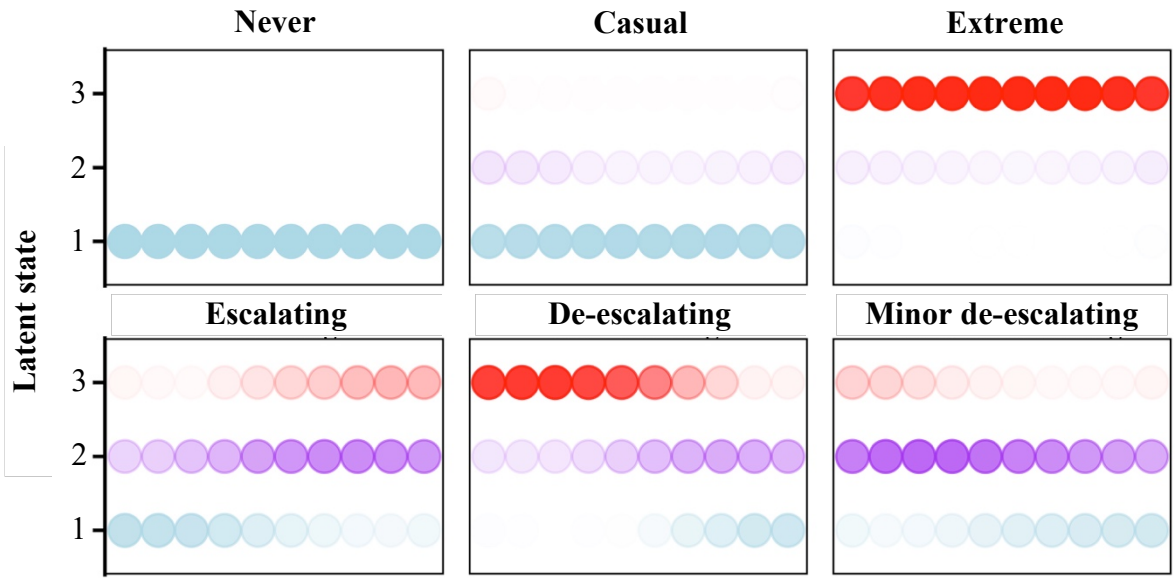


Figure 10, Typified user trajectories of Islamophobia, showing probabilities of users being in the three latent states in the LM model

The typified user trajectories show users' typical behaviour over time. One noticeable discrepancy is that there is not a symmetric opposite to the Never Islamophobic trajectory; there is no trajectory which consists *solely* of the strong Islamophobic tweeting latent state (latent state 3). This is a remarkable difference as it indicates that very few users are constantly expressing strong Islamophobia. Second, is that there are two de-escalating trajectories (De-escalating and Minor de-escalating), but only one escalating. Interestingly, the Escalating trajectory is an inverse midpoint of the two de-escalating ones; it shows a stronger change than Minor de-escalating but is not as strong as De-escalating. Casual and Extreme can be seen as mirrors of each other; behaviour is concentrated in a single latent state (respectively, latent state 1 (i.e. none) and latent state 3 (i.e. strong)) with some aspects of the middle latent state. Descriptions for each of the trajectories are provided in Table 20.

Name	Description
Never Islamophobes[15]	Users who never engage in any form of Islamophobia (whether weak or strong)
Casual Islamophobes	Users who sporadically engage in Islamophobia, only infrequently sending Islamophobic tweets (most of which are weak rather than strong)
Extreme Islamophobes	Users who nearly always send Islamophobic tweets, and overwhelmingly tend to engage in strong rather than weak Islamophobia
Escalating Islamophobes	Users whose Islamophobia is increasing over time, shifting from none to weak and from weak to strong
De-escalating Islamophobes	Users whose Islamophobia is decreasing over time, shifting from strong to weak and from weak to none
Minor de-escalating Islamophobes	Users whose Islamophobia is minorly decreasing over time, whereby most of their behaviour is weak Islamophobic but there is also a shift from strong to none

Table 20, Names and descriptions of the six typified user trajectories

The number and proportion of users assigned to each trajectory is shown in Figure 11. This varies considerably, from 313 users who are De-escalating Islamophobes (4.89%) to 2,028 users who are Casual Islamophobes (31.66%). Interestingly, the combined number of de-escalating users is considerably greater than Escalating (1,177 or 18.39% combined versus 382 or 5.69%), which shows that, during the time period studied, more users are de-escalating than escalating. Worryingly, a large number of users are Extreme Islamophobes (976 or 15.20%), which points to the existence of a committed and persistent base of Islamophobes. However, the symmetric opposite trajectory, Casual Islamophobes, constitutes almost twice as many users (2,028 or 31.66%). The fact that there are more Casual than Extreme Islamophobes suggests that the far right is not

[15] As already noted, this typified user trajectory consists of users who are removed prior to LM modelling as they do not send any Islamophobic tweets.

typified by perpetual Islamophobes but, rather, intermittent Islamophobes. These findings indicate that, overall, followers of the BNP engage in varied behaviour, which can be, to varying degrees, both Islamophobic and non-Islamophobic. Note that this internal variety does not make any of the Islamophobic behaviour that users engage in 'less' bad – all Islamophobic behaviour is harmful, indefensible and should be challenged as a social ill.

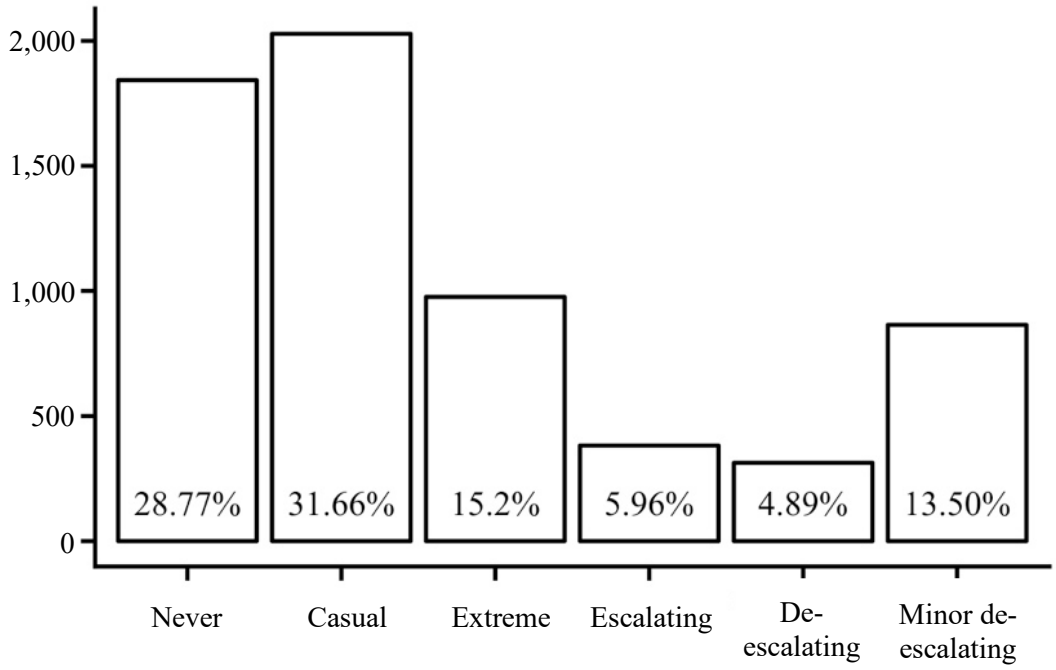


Figure 11, Number and proportion of users assigned to each typified user trajectory

To provide more insight into how individual users' behaviour can vary internally within these typified trajectories, Figure 12 shows the actual behavioural patterns of a random sample of users assigned to each trajectory. This plot demonstrates two important points. First, is that for some trajectories there is considerable variety in users' behaviour. For instance, in the Extreme Islamophobes user trajectory nearly all users behave the same, sending strong Islamophobic tweets at every time interval. For others, such as the Casual Islamophobes, users' behaviour is considerably heterogeneous, both in terms of regularity and strength of Islamophobia which is expressed.

Second, is that because the trajectories are based on latent states – rather than observed or most probable behaviours – they capture similarities across highly heterogeneous user behaviours. For instance, users who send either weak or strong Islamophobic tweets in time periods eight, nine or ten might nonetheless be considered 'de-escalating or 'minor de-escalating'. This is because their underlying trend shows a shift towards less Islamophobia, despite any noisiness in how it manifests. This means that very nuanced differences in user behaviour can be taken into consideration, and users classified appropriately. In practice, this means that users can exhibit complex pathways but still be clustered together. For instance, users in the escalating/de-escalating trajectories do not have to move monotonically between behaviours (i.e. going in one direction constantly) but might switch several times back and forth between the different levels. The LM model is useful precisely because it allows for such bidirectional movements whilst still capturing the broad direction of changes.

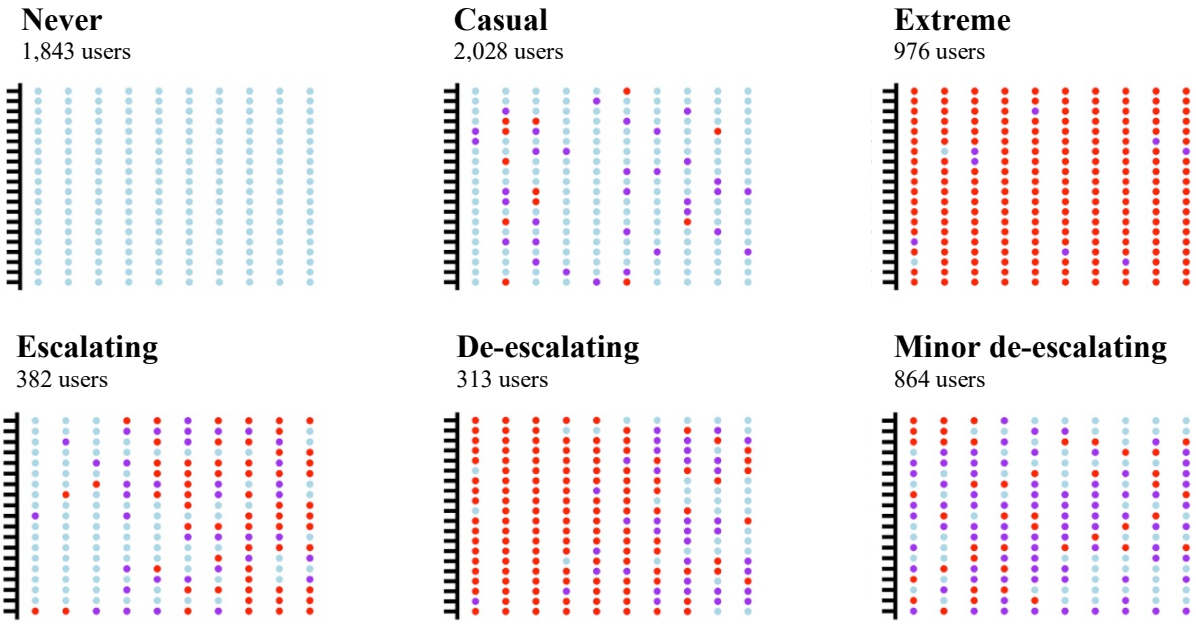


Figure 12, Random users assigned to each of the six user trajectories

Finally, it is worth noting that – as discussed above – the number of time periods used in the LM model can have considerable impact on the trajectories which are identified and how users are clustered together. In Appendix 6.3 I show that with different numbers of time periods the same users are consistently cluster together. The primary source of variation is between 'Perpetual Islamophobes' and 'Casual Islamophobes' as these categories overlap the most. Nonetheless, this analysis further verifies the robustness and utility of the model described here.

6.3.4 | Predicting Islamophobic behaviour

The LM model can be used not only to characterize and understand user behaviour, but also to *predict* future behaviour. This can be achieved by using the latent state transitions. All latent states in the LM model have a transitional probability; this is the probability of either remaining in the same state or changing to a different state. Note that for each latent state there is a positive probability of engaging in each of the three behaviours – which, in practice, means that users can stay in the same state but their behaviour can change. The transitional probabilities for each latent state are shown in Table 21. Users in latent state 1 and 3 have very high probabilities of staying in the same state (0.93 and 0.95 respectively) and very low, and similar, probabilities of moving to either of the other states. But in latent state 2, where the probability of staying in the same state is 0.86, there is a noticeable difference between the transition probabilities of 0.12 to state 1 and 0.02 to state 3. This means that it is far more likely that the Islamophobia of users in this state will de-escalate rather than escalate, which is in line with the findings discussed in the previous sub-section.

Latent state transition probabilities	Transition to state 1	Transition to state 2	Transition to state 3	Sum of probabilities
Start in state 1	0.95	0.038	0.012	1
Start in state 2	0.12	0.86	0.02	1
Start in state 3	0.04	0.03	0.93	1

Table 21, Transition probabilities for moving between each latent state

For predicting *individual* users' behaviour the transitional values in the LM model performs only as well as a baseline model of assigning each user to the same state at t_{j+1} as at t_j. This is because the LM model is memoryless and therefore only considers their prior state. For all latent states, the most probable transition is to stay in the same latent state (as shown in Table 21). However, whilst unhelpful for predicting changes in individuals' latent states, these transition probabilities are useful for predicting latent states and behaviours across the entire cohort of users. That is, LM modelling can be used to predict the *aggregate* number of users who exhibit each type of Islamophobic behaviour. From hereon in this section, all discussions of 'users' refers to users in aggregate. That is, 'users' refers to the number of user 'equivalents' which can be estimated for each time period using probabilities rather than to specific individual users.

I create a new LM model (LM_9) using data from just the first 9 time periods. The input parameters are the same as the original LM model (i.e. with three latent states). I take the latent states each user is assigned to at time period nine from LM_9 (i.e. the most probable latent state) and multiply them by the transition probabilities in LM_9. I then multiply these by the behavioural probabilities for each latent state in LM_9 and as such calculate the expected prevalence across the entire cohort of different types of behaviour. For these calculations, I remove the 1,843 users in the 'Never Islamophobes' category – the behaviour of such users can be predicted with 100% accuracy because it does not change over time. To evaluate the predictive performance of the model I compare it with the

actual number of users exhibiting each type of behaviour at time 10. To ensure this comparison is robust I also compare it against the number of users exhibiting each type of behaviour at time period 9. This is the same predictive baseline as used in meteorological studies; the behaviour tomorrow (t_{j+1}) will be the same as it is today (t_j). The results are shown in Table 22. The results show that the LM model considerably outperforms the baseline at predicting future behaviour. The baseline method labelled 184 users incorrectly, whilst the prediction from LM_9 labels only 58 incorrectly, which is a net gain of 126 users. Compared with the baseline, accuracy at predicting the number of users assigned to each of the three classes increases from 95.96% to 98.73%, an improvement of almost three percentage points. This is an impressive improvement given the already high performance of the baseline.

Tweeting behaviour	Actual number of users at time period 10	Baseline (Number of users at time period 9)	Difference between baseline performance and actual	Predicted number of users at time period 10	Difference between predicted and actual	Improvement over baseline
1 (None)	2,400	2,447	47 (1.95%)	2,429	29 (1.21%)	18 (0.74%)
2 (Weak)	679	724	45 (6.63%)	665	14 (2.06%)	31 (4%)
3 (Strong)	1,484	1,392	92 (6.20%)	1,469	15 (1.01%)	77 (5%)

Table 22, Prediction of the number of users exhibiting each behaviour at time period 10

6.3.4.1 | Prediction with less data

To further test the predictive ability of LM modelling, I fit models for just the first 6, 7 8 and 9 time periods (LM_6, LM_7, LM_8 and LM_9) and then predict the aggregate behaviour for the future time periods, up to and including time period 10. LM_{10} consists of taking the estimated latent states and associated behavioural probabilities in the complete model for each state and multiplying this out – for this reason, there is still some error in the

predicted user behaviours with this model. In Appendix 6.3, I show the performance of each model in predicting all future aggregate behaviour up to the 10th time period. Figure 13 shows the models' performance at predicting aggregate user behaviour in time period 10. The number of mis-assigned users is calculated as above by taking the discrepancy between the actual and predicted number of users for each behavioural state (none, weak and strong Islamophobia). Figure 13 shows clearly that the predictive power of the model decreases as fewer time periods are used to estimate the latent states, transition probabilities and behaviour probabilities. For instance, for time period 10 LM_9 estimates that 2,429 users are none Islamophobic; which is an overestimate of just 29 values. LM_6 estimates that just 2,242 users are none Islamophobic, which is a far larger discrepancy of 158 values. This suggests that accuracy can be increased in the future by using a longer period of data. In Appendix 7.3 I show the performance of LM models with different numbers of time periods, broken down by the three levels of Islamophobia.

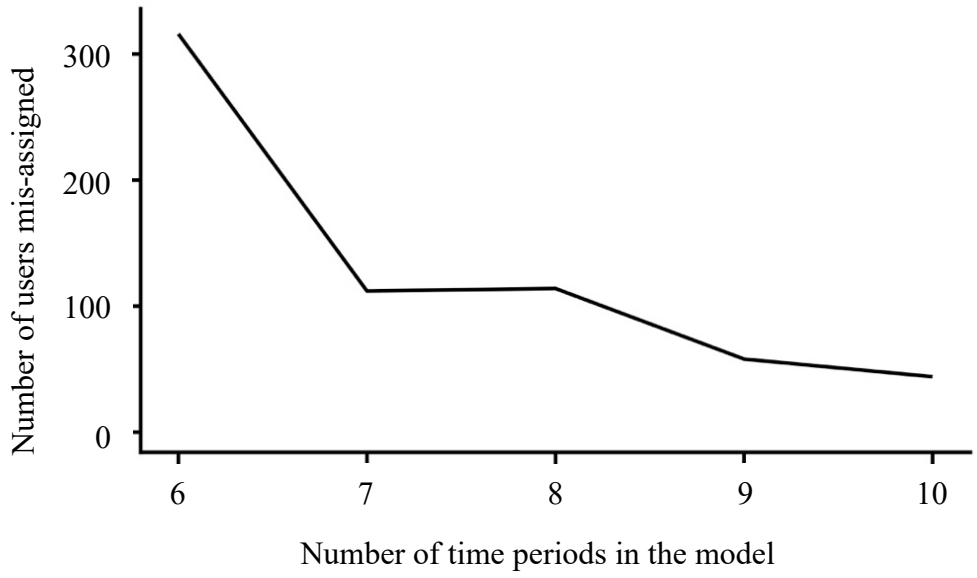


Figure 13, Models' performance at predicting aggregate behaviour in time period 10

Finally, I represent the performance of all the LM models for the three strengths of Islamophobia in Figure 14. I add back in the 1,843 users who are in the 'Never

Islamophobes' type to ensure the panels can be interpreted in line with the rest of the results. Accordingly, all 6,406 users are represented. The solid lines show the actual behaviours of the users and the dotted lines show the results for each of the four models. This visualization shows that as fewer time periods are used to estimate each LM model the predictions not only become less accurate but also more *stable*, which indicates that they are less capable of taking into account changes in user behaviours. This means not only that models which are trained on less data are less accurate but also that they are more likely to approximate the baseline method and predict that users exhibit the same behaviour in t_{j+1} as they did in t_j – which means that that they are far less nuanced. Overall, the close approximation achieved by LM_9, as well as the comparatively poor performance of LM_6, shows both the potential power of LM modelling but also the importance of having sufficient data to make the model's predictions useful.

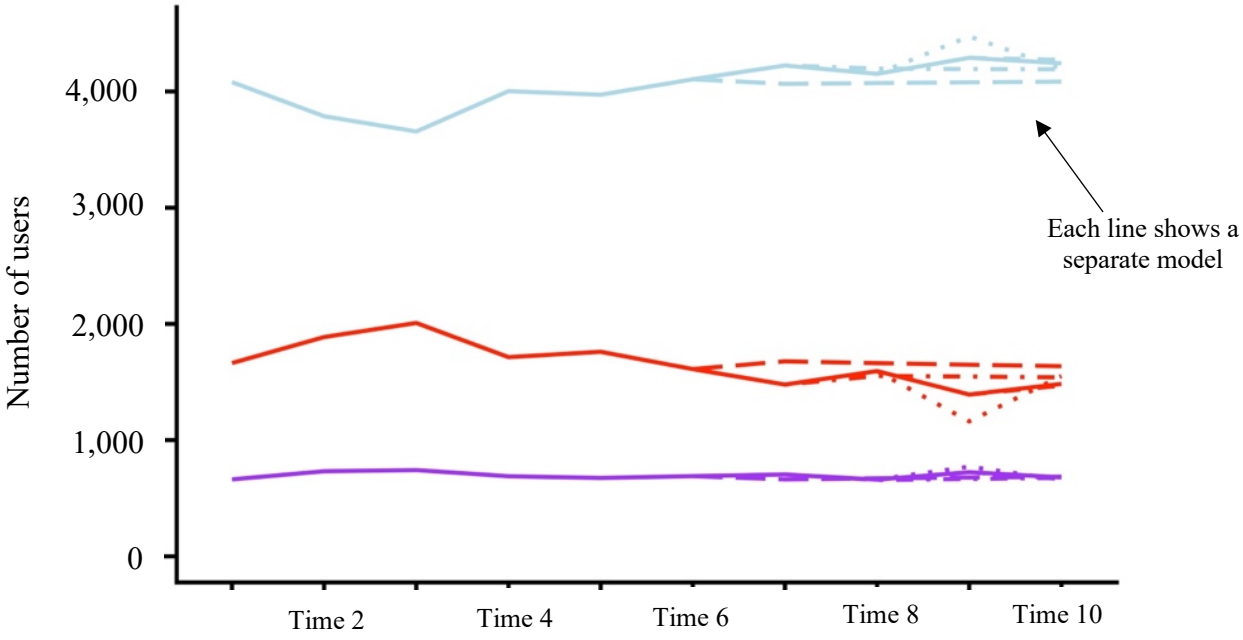


Figure 14, Predictive performance of models LM_6, LM_7, LM_8 and LM_9 versus the results from the original LM model

6.4 | Conclusion

This Chapter has sought to answer the second research question:

> RQ 2: To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?

The results demonstrate the heterogeneity of Islamophobic user behaviour within the far right. First, Islamophobic behaviour itself is varied; weak and strong Islamophobia can be observed in considerable quantities, although weak is more prevalent than strong (10.8% compared with 5.3%). Second, users are more likely to be on the extremes of behaviour, as shown in Section 6.3: far more users than anticipated are in the None Islamophobic only type and far fewer are in the None, Weak and Strong type. Third, users engage in remarkably different trajectories of behaviour, as shown by the analysis of user trajectories in Sections 6.4 and 6.5.

This Chapter has focused on followers of the BNP. Arguably, the similarities between the BNP and other far right parties, such as the National Front and Britain First (which largely share the same activists, tactics, support bases and messages (Copsey, 2007; Ford & Goodwin, 2010)), suggests that the results can be considered representative of followers of UK far right parties. It is plausible that the prevalence of each trajectory, and other details, differ – but the overall user-level dynamics should be broadly similar. However, the behavioural trajectories identified might not generalise to qualitatively different types of far right organisations, such as street movements (Castelli Gattinara & Pirro, 2018). There is no such thing as a 'typical' far right organisation; they each share certain similarities (such as prejudices against Muslims/Immigrants and a populist opposition to elitism) but there are often many confounding differences, such as the leadership structure, geographical basis of support, type of political action pursued, and

political goals. In particular, many newer far right groups attract younger and more vitriolic supporters, who might be more extreme in their behavioural patterns (Chris Allen, 2011; Hope Not Hate, 2017). As such, these non-party far right organisations are likely to exhibit different trajectories of behaviour.

The LM model provides powerful insight into users' Islamophobic behaviour, reducing the considerable heterogeneity of the actual patterns of 6,604 users' behaviour into just six typified trajectories. These typified trajectories have been named and, accordingly, can be used to uncover further insight into the nature of far right behavioural journeys This analysis marks an important step forward in quantifying different trajectories of behaviour. Although quantification can be somewhat reductive, the work here shows that different user states can be identified and their prevalence measured. This is important for ensuring that our understanding of the far right is not only theoretically insightful but also precise and empirically supported.

The 'far right' is often used as a broad brush to characterize those on the extreme of UK politics. However, this analysis shows that the far right is very varied, comprised of users exhibiting many different patterns of behaviour. Only those in the Extreme Islamophobes trajectory fit the stereotype of the constantly hateful and aggressive far right. This suggests that whilst some users certainly do create a 'wall of hate' (as evinced by those users who send many Islamophobic tweets, shown above in Figure 5), they comprise a minority of the far right. Most users engage in Islamophobia far more infrequently – and a large number (n = 1,843) never send any Islamophobic tweets. This has important implications for the UK Government working to monitor far right extremism, platforms which seek to remove hateful and harmful content, and civil society groups who want to counter and challenge the far right. Not least, it suggests that other factors, beyond Islamophobia, are what attracts users to follow and support far right parties. This builds

on the qualitative conceptual work in Chapter 4 regarding the weak manifestations of Islamophobia by demonstrating the casual, irregular and uneven ways in which Islamophobia manifests – even amongst the far right. This could be explored further in future work by qualitatively investigating the Never Islamophobes' behaviour.

This Chapter throws up as many questions as it answers; not least of which is how users enter the Extreme Islamophobes trajectory – are they Extreme Islamophobes as soon as they join Twitter or do they go through an escalation period? Furthermore, do users exhibit 'cycles' of Islamophobia, which increases and decreases over a longer period of time? What are the causes of escalation and de-escalation? In this Chapter I have not explicitly modelled the role of external events, such as terrorist attacks. Such events could be driving some of the behavioural trajectories identified. In particular, it is possible that users in the de-escalating trajectories are usually 'Casual' Islamophobes but have been 'activated' to be highly Islamophobic at the start of the period as many terrorist attacks occurred at this time. This also raises an interesting question as to why users which I have allocated to the Casual trajectory were not highly activated by the terrorist attacks at the start of the period – and why users in the 'Escalating' trajectory' were Islamophobic at the end of the period but not at the start (despite the terrorist attacks taking place then). Users in the Escalating, De-Escalating and minorly De-escalating trajectories could provide particularly useful insight into the pathways of Islamophobia within the far right in future research.

The analysis in this Chapter is based on the study of a single prejudice (Islamophobia) on a single platform (Twitter). Potentially, some of the dynamics do not reflect changes in users' behaviour *in general* but simply changes *on Twitter*. Users could engage in the same level of Islamophobia but (i) switch between online platforms, (ii) switch between identities/alters within a single platform or (iii) switch between online and offline

settings. They might also switch between Islamophobia and other forms of prejudice, such as racism and anti-Immigrant prejudice. Thus, whilst trajectories of Islamophobia are robustly identified on Twitter, it may be that these are driven by intra- and inter-platform dynamics which have not been studied. That said, there is not strong evidence that users habitually switch between platforms, identities and online/offline settings so rapidly or frequently that all of the observed behavioural changes can be attributed to it. Accordingly, whilst this issue requires further in-depth analysis, potentially with different research methods and multi-site datasets, it does not invalidate the results.

6.4.1 | Limitations

There are several limitations of the current LM model which could be further investigated and addressed in future work. First, even though the k-modes clustering algorithm has proven to be remarkably effective, it suffers from a key drawback in that it does not explicitly model the temporal sequence of items for each user. Clustering data which lies in a categorical rather than continuous space is notoriously difficult, as is explicitly modelling time. Nonetheless, recent work in information retrieval systems could be applied here to use a more theoretically-informed clustering algorithm, which might enhance the output's interpretability (Crane, 2015; Yuan, B, Chen, & Cai, 2016). Furthermore, a deterministic algorithm which does not rely on random initializations (as with K-modes) could increase the integrity of the findings – although this has been addressed here by inspecting the output from several initializations.

Second, the number of time periods impacts the trajectories identified in the model. Although I have tested for four different time periods (T = 10, 25, 50, 100) it would be beneficial to further investigate this and to see whether more nuanced user trajectories can be identified when a longer value of T is used. This, however, requires using more

data – and the third limitation is that the LM model is data hungry and performs far better with more data to estimate the probabilities. To ameliorate this, I would need to either collect data over a longer time period or on more users. Using a larger number of time periods would likely mean that more latent states would be identified, increasing the complexity of the model but also the nuanced insight that can be extracted from it. Finally, a longer time period would mean that user behaviour is more likely to be driven by external events and one-off changes, such as terrorist attacks, Government reports or other political changes. There would be more noise in the trajectories, which might require the introduction of covariates, such as how long has elapsed since an event – this is addressed in the next chapter (Chapter 7), which examines the role of terrorist attacks.

Fourth, the work here is only possible due to the multi-level classifier developed in the previous chapter, which takes into account different strengths of Islamophobia. As the latent Markov model shows, this must be taken into account to understand the complexity of user behaviour; a binary classifier would miss much of the variety in how users tweet including (i) how users' behaviour escalates and de-escalates and (ii) how users can be categorised based not just on the regularity but also the strength of Islamophobia they express. However, the quality of the modelling depends upon the accuracy of the multi-class classifier developed in the previous chapter. Improving the classifier will lead to more robust modelling of users and as such the utility of this work. In addition, the choice of taking the 'strongest' behaviour exhibited by each user in each time period as representative of that time period could be changed in future work.

6.4.2 | Extensions

There are several extensions to the current LM model which could be implemented in the future. First, is that the dependent variable could be changed from univariate to

multivariate, counting the number of tweets sent in each of the three levels in the multi-class classifier. This would provide greater insight into different user trajectories by systematically accounting for the *magnitude* of Islamophobic tweeting. That said, it would likely still require a scaling metric to take into account the different volumes of tweets sent by each user and would need further statistical checks to ensure that the dependent variables do not violate the assumptions of the multivariate LM model.

Second, is that the current LM model treats periods in which users do not tweets as none Islamophobic. This is a reasonable assumption but it could be worth investigating the impact of modelling periods when users do not tweet as a fourth level in the dependent variable (for a univariate LM model) or as a separate outcome (in a multivariate LM model). This might help to address situations where users stop tweeting – not because they are no longer engaging in Islamophobia but because either they (i) move to other social media platforms, (ii) act Islamophobically offline or (iii) use 'alters' on Twitter, sending tweets through a different account. It is very difficult to know what happens during such periods of seeming inactivity; but modelling periods where users do not send tweets as a separate form of behaviour could help to provide exploratory insight into this issue.

Third, is that the LM model performs well at accurately predicting the future behaviour of users *in aggregate*. This could be extended to model individual users' behaviour by using more advanced time-based sequence modelling, rather than just the LM model output (i.e. the latent state probabilities, behavioural probabilities and transition probabilities). Possible applications include either an Auto-Regressive Moving Average (ARIMA) model or the pattern sequence forecasting algorithm (Bokde, Asencio-Cortés, Martínez-Álvarez, & Kulat, 2016).

Chapter 7 | The Twin threat of Islamophobia

There are two goals in this Chapter. First, to investigate the nature of Islamophobia across different political parties (including both far right and mainstream parties), addressing the third research question in this thesis:

> RQ 3: To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?

Second, to investigate the role of terrorist attacks in driving Islamophobic, addressing the fourth and fifth research questions:

> RQ 4: To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?

> RQ 5: Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?

Answering these RQs will enhance scholarly understanding apropos both the nature of Islamophobic behaviour in contemporary UK politics and the causal drivers of such behaviour on social media. To answer the RQs, I use a dataset of tweets sent by followers of four political parties: UKIP, the Conservatives, Labour and the BNP as well as the multi-class and binary machine learning classifiers developed in Chapter 5. The findings build on the results of the previous chapter, Chapter 6.

In the first section, I describe the dataset of tweets collected from followers of each party. There second, third and fourth sections consist of data analysis. In the second section, I find considerable differences in the volume of Islamophobia across different political

parties by using a fixed effect regression model and appropriate statistical tests, thereby answering RQ 3. I also provide insight into the temporal dynamics of Islamophobia, showing the bursty and uneven nature of such behaviour. In the third section, I investigate the impact of Islamist terrorist attacks in driving Islamophobia. I identify a temporal process of sharp sudden escalation and then longer de-escalation which holds across all terrorist attacks, show the importance of party followership, the number of people killed and the number injured during each attack, and also investigate the role of the media in driving Islamophobia. This answers RQs 4 and 5. In the fourth section, I investigate *who* tweets during Islamist attacks, and show that the sharp increase in Islamophobia is due to a small group of hyper active tweeters rather than one-off or low-volume tweeters. In the conclusion, I revisit the research questions and evaluate the extent to which they have been answered by discussing the results, considering the limitations of the research and outlining possible extensions.

In this Chapter and the appendices, a single colour scheme is used to describe the political parties. Red is for Labour, blue is for Conservatives, purple is for UKIP, brown is for BNP and light grey is for all parties combined. This is shown in the colour key below. I do *not* use the same colour scheme as in the previous chapter, where light blue denoted none Islamophobic, purple denoted weak Islamophobia and red denoted strong. This should be taken into account when comparing results between the two chapters and interpreting the overall narrative of the thesis. In future work, the choice of colours will be re-evaluated.

Colour key for this Chapter

All parties | UKIP | Conservatives | Labour | BNP

7.1 | Data overview

Followers of four parties are studied in this Chapter: the BNP, UKIP, Conservatives and Labour. A period of one year is studied, from 1st March 2017 to 28th February 2018. All tweets produced by followers of the four parties during this period are collected (in accordance with the data collection process outlined in Chapter 3). Most UK political parties, particularly those with elected representatives in national or supra-national bodies, have a very large number of followers on Twitter, as shown in Table 23. I sample 7,500 users from each party, based on relevant power tests, using appropriate sampling methods for long-tailed distributions (outlined in Appendix 7.3). To ensure that I can disambiguate how followers of different parties vary I only include users in each sub-sample if they *do not follow* the other parties studied here. For example, users can only be included in the Conservatives sub-sample if they do not follow any of UKIP, Labour or the BNP.

Party	**Number of followers on 1st March 2017**	**Size of sub-sample**
UKIP	153,623	7,500
Conservatives	236,306	7,500
Labour	502,465	7,500
BNP	12,895	7,500

Table 23, Number of followers for each party

I only include users in the sample who are active (i.e. they have sent at least one tweet) and follow the party at both the start and end of the period. I remove tweets which are not either in English or the 'Undetermined' language. I also remove highly active accounts which are likely to be bots. These actions reduce the size of the dataset considerably from a combined total of 30,000 users to 15, 253. In total, 11,143,987 tweets

are studied. Table 24 provides a summary of the tweeting behaviour of followers for each party. The total number of tweets per party varies from 2.14 million to 3.17 million.

Party	Original size of sub-sample	Number of followers after sampling	Total Number of tweets	Average number of tweets per user
UKIP	7,500	3,497	2,691,105	770
Conservatives	7,500	3,346	2,135,850	638
Labour	7,500	4,683	3,167,564	676
BNP	7,500	3,727	3,149,468	845
TOTAL	30,000	15,253	11,143,987	731

Table 24, Summary of final dataset for followers of each party

7.2 | Islamophobia across parties

The goal of this section is to understand how the prevalence of Islamophobia varies across followers of the four different political parties. I use the classifier outlined in Chapter 5 to annotate all tweets in the dataset (n = 11,143,987). On average, 90.94% of tweets are non- Islamophobic, 6.48% are weak Islamophobic and 2.58% are strong Islamophobic. Figure 15 shows the prevalence of the three types of tweets (none, weak and strong Islamophobia) for each party (as shown by the respective colours) and also all parties together. The right-hand panel of Figure 15 zooms in on just the weak and strong tweets shown in the left-hand panel.

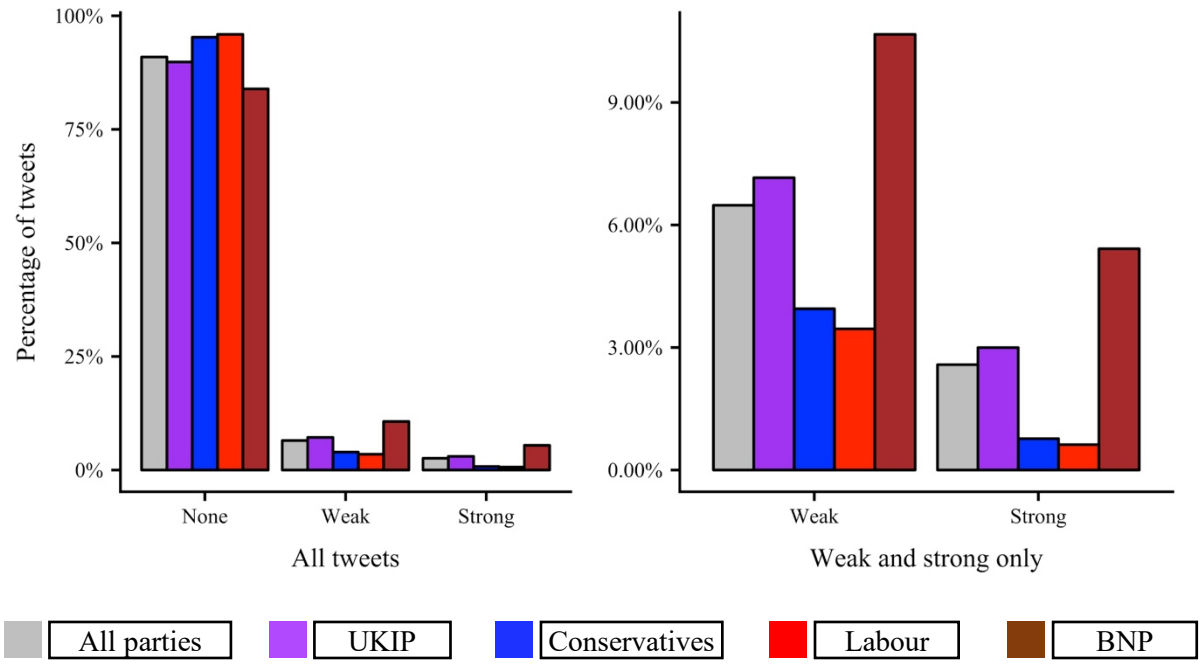


Figure 15, The percentage of tweets which are Islamophobic for each party. The right-hand panel shows just weak and strong Islamophobic tweets

Across all parties none Islamophobic tweeting is the most prevalent type of behaviour, followed by weak Islamophobia and then strong. However, there are some considerable differences between parties. First, the prevalence of each type of tweet differs by party.

BNP send the most Islamophobic tweets (~16% combined), followed by UKIP (~10%), and then Conservatives (~5% combined) and Labour (~4% combined). Second, the proportion of weak and strong Islamophobic tweets varies considerably across parties. For Labour, the multiple of weak tweets compared with strong is 5.6 (3.46% weak and 0.62% strong) whilst for BNP it is just 1.97 (10.67% weak and 5.42% strong). The more Islamophobic tweets that each party sends, the more of those tweets are strong Islamophobic rather than weak. This suggests that users who engage in more Islamophobia (quantitatively) are also likely to engage in stronger Islamophobia (qualitatively). This suggests that when followers of the BNP engage in Islamophobia they are more likely to send directly and explicitly Islamophobic tweets rather than subtle, nuanced and partial ones, as is the case with Labour. It also shows that summarising the overall prevalence of Islamophobia across the entire dataset is unlikely to well represent the data due to considerable cross-party variations in behaviour. For instance, the overall prevalence of strong Islamophobia is 2.58% - but 5.42% of the BNP's tweets are strong Islamophobic, whilst just 0.62% of Labour's are. This is a difference of 4.8 percentage points, or nearly one order of magnitude. Further details on the prevalence of Islamophobia across the four parties is given in Table 25.

Party	Total number of tweets	Average tweets per user	Percentage None (#)	Inverse Rank None	Percentage Weak (#)	Rank Weak	Percentage Strong (#)	Rank Strong
UKIP	2,691,105	770	89.85% (692)	2	7.16% (55)	2	3.00% (23)	2
Conservatives	2,135,850	638	95.29% (608)	3	3.95% (25)	3	0.77% (5)	3
Labour	3,167,564	676	95.92% (649)	4	3.46% (23)	4	0.62% (4)	4
The BNP	3,149,468	845	83.91% (709)	1	10.67% (90)	1	5.42% (46)	1
Combined	11,143,987	731	90.94%		6.48%		2.58%	

Table 25, Tweeting habits of followers of each party

I compare each party's distribution of the *number* of Islamophobic tweets per follower. I plot the density of the number of Islamophobic tweets per user with a logarithmic x-axis in Figure 16. Many users do not send any Islamophobic tweets (n = 4,795) and I remove these from the plot. On the left-hand side of the distribution (where the number of Islamophobic tweets is very low), the BNP has the lowest density followed by UKIP, Conservatives and Labour. On the right-hand side of the plot, the order is reversed, whereby BNP and UKIP have considerably higher density than Conservatives and Labour. This shows that BNP and UKIP are more likely to have followers who send a large number of Islamophobic tweets. This ordering is also in-line with the overall volume of Islamophobic tweets sent by each party, as shown in the Table above. This provides evidence that the large volume of tweets sent by the BNP is driven by a small number of very active users (in some cases, sending close to 5,000 Islamophobic tweets during the period). It also suggests the inverse; that many followers of the BNP send a similar number of Islamophobic tweets as the other parties. In Appendix 7.1, I show a plot where 1 is added to each users' count of Islamophobic tweets, thereby allowing users with no Islamophobic tweets to be included. The tail of this plot is very similar to Figure 16. Note that the distribution of the percentage of tweets per user which are Islamophobic is also very similar – but is not shown here for brevity.

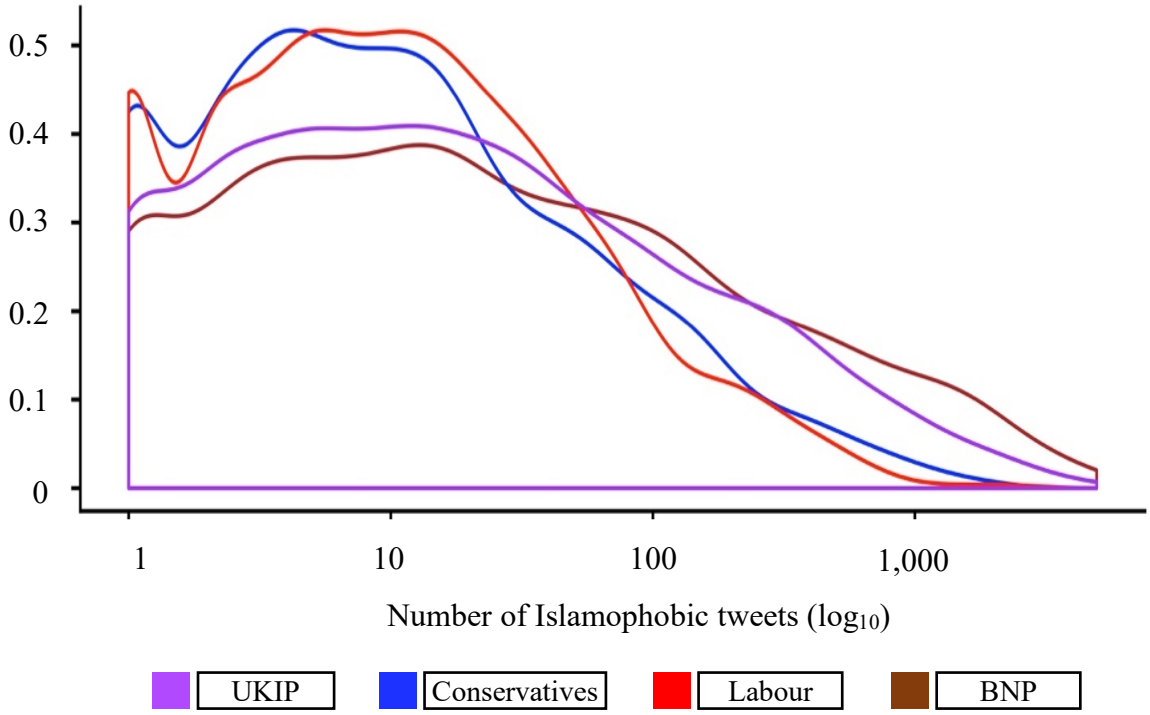


Figure 16, Density plot for the number of Islamophobic tweets for each user, split by party – zero values removed

I quantify the inequality of the distribution of the count of Islamophobic tweets for followers of each party by calculating the Gini coefficient, shown in Table 26, and plotted as a Lorenz curve in Figure 17. The Gini coefficients are high for all parties (ranging from 0.831 to 0.883), which suggests that in all parties a small number of users are responsible for most of the Islamophobic tweets which are sent. Noticeably, the coefficients are highest for the most Islamophobic parties (BNP and UKIP). The high Gini coefficient for the BNP shows that even within the far right users vary considerably and that most users do not send, proportionally, many Islamophobic tweets. It also provides further evidence that the large volume of Islamophobic tweets sent by followers of the BNP is driven by a small number of very Islamophobic users.

Party	Gini coefficient	Gini coefficient rank
UKIP	0.883	1
Conservatives	0.875	3
Labour	0.831	4
BNP	0.880	2

Table 26, Gini coefficients for each party's cumulative volume of Islamophobic tweets versus the cumulative volume of users

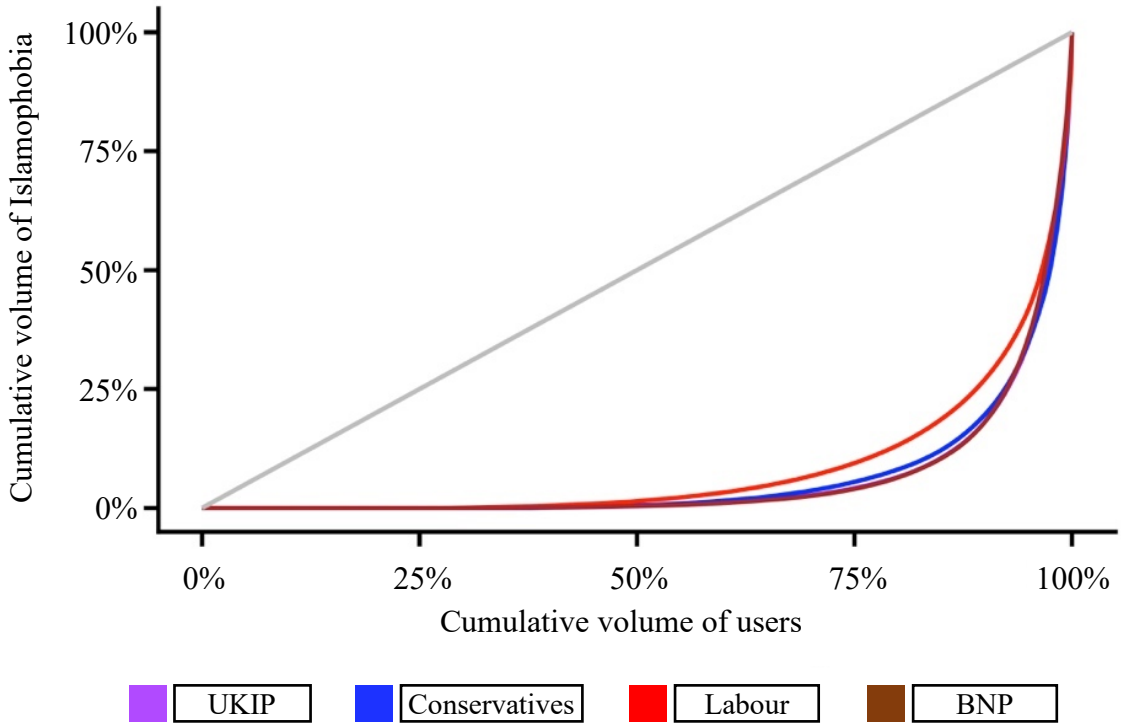


Figure 17, Lorenz curve of Gini coefficients for each party's cumulative volume of Islamophobic tweets versus the cumulative volume of users

7.2.1 | Statistical significance of party differences in Islamophobia

To better understand how Islamophobic tweeting differs across parties, I study the raw probabilities that a tweet is either Islamophobic or not.[16] I average these for each user and then aggregate by party to capture party-level differences. The results of this analysis are shown in Figure 18. Each dot is a user and the higher the dot the greater the probability of the user sending an Islamophobic tweets. The circumference of the dot represents the total number of tweets sent by each user. The black dot shows the average probability of Islamophobia for each party and the error bars show a range of 2 standard deviations, with a floor of 0. Figure 18 shows the clear differences between Conservatives and Labour on the one hand, which tend to have low Islamophobia and are tightly clustered, and the BNP and UKIP on the other, which have greater variance. Noticeably, most of the followers of Conservatives and Labour with a medium to high probability of Islamophobia have very few tweets. In contrast, many of the medium to high probability Islamophobic tweeters for both UKIP and the BNP send a large number of tweets overall.

[16] The probabilities are calculated for each tweet using the binary classifier outlined in Chapter 5. A value is assigned to each tweet for none, weak and strong Islamophobia, the total of which sums to 1. These values are calculated for all 11,143,987 tweets in the dataset. Note that the probabilities of Islamophobic behaviour are calculated rather than the prevalence of the actual assignments (which could be expressed as a percentage, thereby giving a probability). As such, the average probability of Islamophobia is higher than the average *prevalence* of Islamophobia.

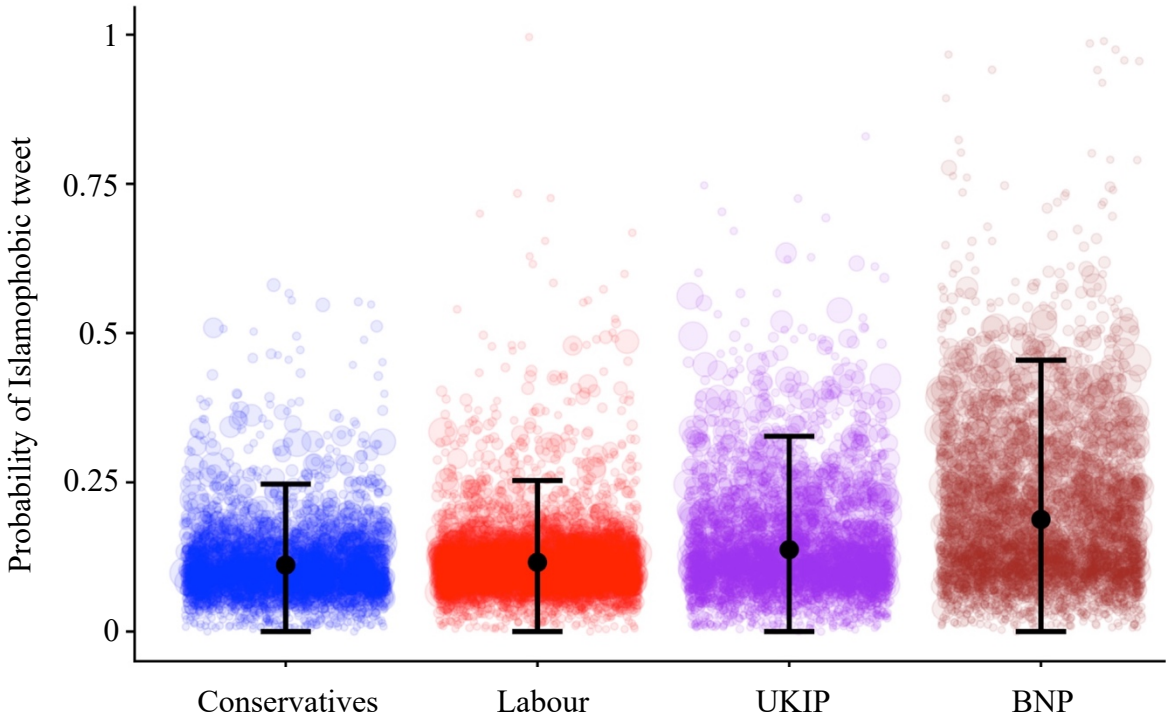


Figure 18, Probability that a users' tweets are Islamophobic (using the binary classifier) with the size of the dots showing the total number of tweets they send, split by party

The results so far suggest that there are important differences in the level of Islamophobic tweeting across parties. To verify that these differences are meaningful, I run several omnibus statistical tests and regression models, as shown in Table 27. In all cases, the independent variable is the party affiliation. Islamophobic tweeting (the dependent variable) can be measured in several different ways, and to ensure that the results are trustworthy I test for (i) different varieties of Islamophobia, namely *weak* Islamophobic, *strong* Islamophobic and *Islamophobic* tweeting (in which the values for weak and strong Islamophobia are combined), and (ii) different types of variables, namely continuous variables (using the probabilities assigned by the multi-class classifier), categorical variables (taking both the modal and strongest state) and count variables (using the raw counts of the assigned Islamophobia values). The continuous probabilities are particularly useful as they take into account the fact that the average number of tweets

sent by each user varies. Every statistical test reported in Table 27 is extremely significant. This provides very strong evidence that there is a statistically significant relationship between the party that users follow and the level of Islamophobia they engage in. Note that all of the statistical tests are omnibus tests and show whether there is a difference across all parties – they do not show which party sends more Islamophobia.

Type of variable	**Dependent variable (calculated for each user)**	**Significance level and test**[17]
Continuous (probability)	Probability that tweet is Islamophobic	***** One-way ANOVA
		***** Kruskal Wallis
		***** Fixed effects model
	Probability that tweet is weak Islamophobic	***** One-way ANOVA
		***** Kruskal Wallis
		***** Fixed effects model
	Probability that tweet is strong Islamophobic	***** One-way ANOVA
		***** Kruskal Wallis
		***** Fixed effects model
	Probability that tweets are strong and weak Islamophobic	***** One-way MANOVA
Ordinal	Modal Islamophobic state	***** Chi-square
		***** Ordered logit regression
	Strongest Islamophobic state	***** Chi-square
		***** Ordered logit regression
Count	Count of Islamophobic tweets	***** Poisson regression
		***** Negative binomial regression
	Count of weak Islamophobic tweets	***** Poisson regression
		***** Negative binomial regression
	Count of strong Islamophobic tweets	***** Poisson regression
		***** Negative binomial regression

Table 27, Statistical tests for the relationship between party and Islamophobic tweeting

[17] Significant results are demarked as N.S. == not significant, * == p(< 0.05), ** == p(< 0.01), *** = p(< 0.001), **** = p(< 0.0001), ***** = p(< 0.00001). Note that this notation for significance is used throughout the Chapter.

7.2.2 | Size of party differences in Islamophobia

Table 27 shows significant results – but this could be due to the large sample size which I use, which means that even very small differences are captured. As Meehl puts it, '[a]nybody familiar with large scale research data takes it as a matter of course that when the N gets big enough she will not be looking for the statistically significant correlations but rather looking at their patterns, since almost all of them will be significant.' (P. E. Meehl, 1990, pp. 204–205) Thus, the fact that there are significant differences in Islamophobia across followers of different parties is not surprising. What matters most is the *size* of the difference. For the remainder of this analysis I focus on the *probability* of Islamophobic tweeting as this is likely to best capture differences in party followers' *propensity* to engage in Islamophobic behaviour. It also means that differences in the overall number of tweets sent by followers of each party will not unduly bias the results.

I fit models for the probability of Islamophobia, weak Islamophobia and strong Islamophobia for the whole period combined (i.e. every tweet is modelled as a separate data instance). These models are shown in Table 28. As party is a categorical variable, three coefficients are estimated. Followership of the BNP is the baseline against which these values are calculated and as such, the BNP is not included as a coefficient in the model – but is estimated via the y intercept. Coefficients are calculated for UKIP, Conservative and Labour party followership.

Model[18]	R-squared, adjusted	y-intercept	Party (UKIP)	Party (Conservatives)	Party (Labour)
Probability (Islamophobia)	0.04, 0.04	0.2413	-0.0606	-0.1158	-0.1196
Probability (Weak Islamophobia)	0.03, 0.03	0.1527	-0.0334	-0.0636	-0.0677
Probability (Strong Islamophobia)	0.04, 0.04	0.0886	-0.0273	-0.0522	-0.0519

Table 28, Summary of OLS linear regression models

Table 28 shows that although the models are statistically significant their R-squared values are very low (between 0.03 and 0.04). The models likely perform poorly because users' behaviour has a strong temporal dimension, and without taking this into account party differences cannot be fully captured. To account for the impact of time, I implement fixed effect models on the data, which is a widely-used statistical method for multi-level data (Bell, Fairbrother, & Jones, 2018; Mummolo & Peterson, 2018; Vaisey & Miles, 2017). Fixed effect models assume that the effects of the independent variable (here, Party followership) on the dependent variable (the probability of sending an Islamophobic tweet) operate within a fixed third variable (the time interval). The inclusion of the third fixed variable allows fixed effects models to account for 'omitted variable bias'. This is when external factors bias the results and estimated coefficients because they exert a countervailing effect to the independent variables which have been studied (A. T. A. B. Snijders, 2005).

[18] For all 9 FE models, every term is extremely significant (*****) and every model is extremely significant (*****).

In the fixed effect model, I hold time constant and then, for each time interval, estimate the impact of party followership on the probability of sending an Islamophobic tweet. This differs from many fixed effects models, where typically a fourth variable is used as the 'within level of analysis' and then a pair of variables are measured over time (Dranove, 2012). Thus, there are typically at least four variables – the independent variable, the dependent variable, the fixed effect and the time stamp. Each time stamp is attached to a pair of independent and dependent variables, which each constitute a separate wave of data. Here, the time period is the fixed effect and so each separate wave of data consists of each user's party affiliation and their probability of sending an Islamophobic tweet (within each time period). In effect, there are as many waves as there are users. This allows me to measure the impact of party followership on the probability of sending an Islamophobic tweet, accounting for the impact of time.

Fixed effect models are implemented in R using the 'plm' function from the PLM package (Croissant & Millo, 2008). The call is: probability (Islamophobia) ~ party | FE (time). As with the Ordinary Least Squares (OLS) models above, BNP is used as the baseline and coefficients are estimated for the impact of UKIP, Conservatives and Labour. I fit fixed effect models for the probability of Islamophobia, weak Islamophobia and strong Islamophobia. The *length* of the time period in the fixed effect model is inherently arbitrary. As such, I model time granularities of 10,000 seconds, 100,000 seconds and 1,000,000 seconds in order to check whether the results are consistent. 10,000 seconds is approximately 2.7 hours, 100,000 seconds is 27 hours (just over 1 day) and 1,000,000 seconds is ~11.6 days. In total, I fit 9 models.

The fixed effect models are shown in Table 29. For the models of time period = 1,000,000 there is a close to 100% improvement in the R-squared values compared with the linear OLS models reported above (the R-squared values increase from 0.03-0.04 to 0.07-0.08).

This demonstrates the greater power of the fixed effects model compared with the linear OLS regression model (Dranove, 2012). Overall, the R-squared values for the models are quite low. However, all of the terms are significant and within each type of Islamophobia the differences in the coefficients for each party are quite consistent.

Model[19]	**Time period (seconds)**	**R- squared, adjusted**	**Party (UKIP)**	**Party (Conservatives)**	**Party (Labour)**
Probability (Islamophobia)	1,000,000	0.08, 0.07	-0.0456	-0.078	-0.079
	100,000	0.05, 0.05	-0.0473	-0.082	-0.084
	10,000	0.05, 0.04	-0.0505	-0.092	-0.094
Probability (Weak Islamophobia)	1,000,000	0.04, 0.04	-0.025	-0.040	-0.42
	100,000	0.03, 0.03	-0.026	-0.044	-0.047
	10,000	0.03, 0.03	-0.029	--0.051	-0.054
Probability (Strong Islamophobia)	1,000,000	0.07, 0.07	-0.020	-0.034	-0.035
	100,000	0.05, 0.05	-0.021	-0.037	-0.037
	10,000	0.05, 0.05	-0.0.22	-0.041	-0.042

Table 29, Summary of OLS fixed effect regression models

For the model which measures the overall probability of Islamophobia (both weak and strong combined) with 1,000,000 second granularity, the coefficient estimate for Labour party followers is -0.079, Conservative party followers is -0.078, and UKIP -0.046. This means that on average the probability of a UKIP follower sending an Islamophobic tweet is 0.046 less than that of a BNP follower and for the Conservatives and Labour it is 0.078 and 0.079 less respectively. The coefficient for the BNP can be estimated by taking the average fixed effect across each of the 33 time periods in this model: 0.187.

These models show that Labour and the Conservatives are remarkably similar, which is surprising given the media focus on Islamophobia within the Conservatives, and that

[19] For all 9 FE models, every term is extremely significant (*****) and every model is extremely significant (*****).

UKIP is approximately halfway between them and the far right BNP. Noticeably, the differences between the coefficients for each party are consistent across the three independent variables (Probability of Islamophobia, Weak Islamophobia and Strong Islamophobia). This suggests that the differences between parties hold across the different strengths of Islamophobic tweeting.

7.2.3 | Islamophobia over time

The goal of this subsection is to understand how the prevalence of Islamophobia varies *over time* for followers of the four political parties. Figure 19 shows tweets and Islamophobia over time. Panel A shows the total number of tweets sent each day, smoothed with time windows for 7 days and 30 days. Panel B shows the prevalence of weak and strong Islamophobia for each day. The bottom line (dashed) shows strong Islamophobia and the top line (dotted) shows weak. Panel C shows the total number of tweets sent by each party and panel D shows the total number of Islamophobic tweets (both weak and strong) sent by followers of each party.

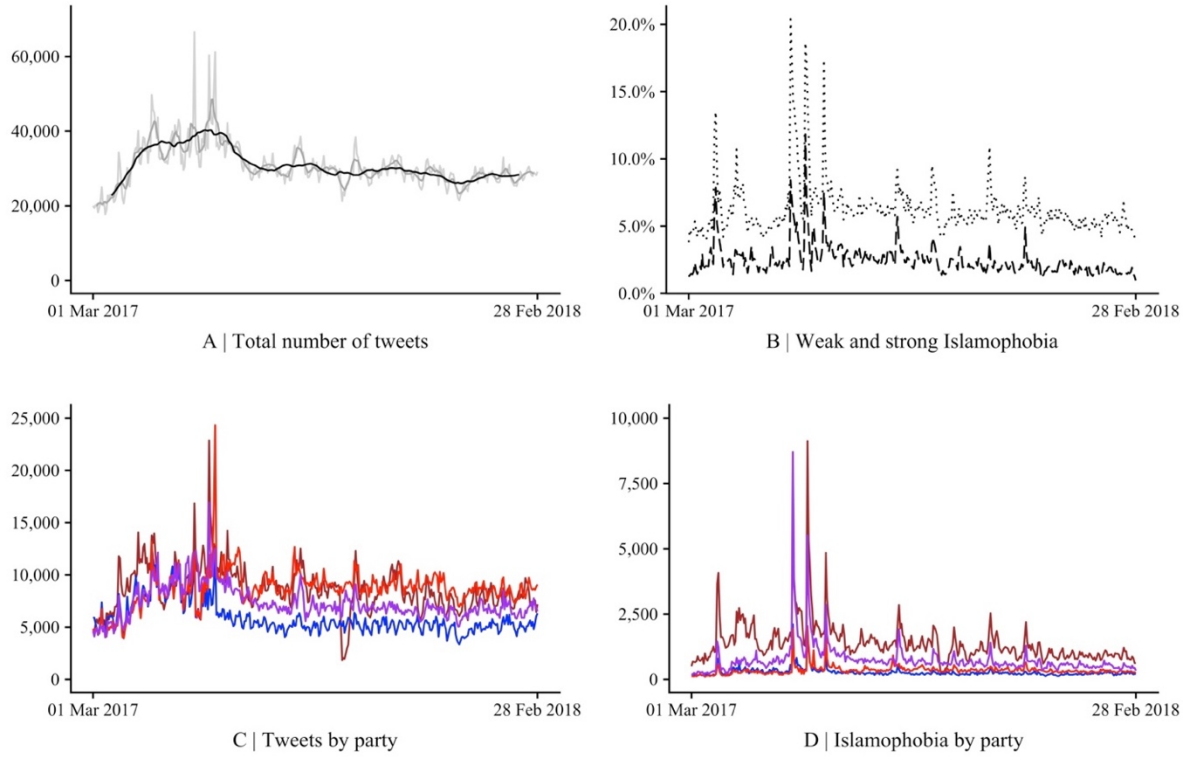


Figure 19, Number of tweets and Islamophobia over time

There are some striking similarities between the panels. Noticeably, weak and strong Islamophobia (shown in panel B) follow a very similar pattern, which suggests that at the aggregate level they are driven by similar temporal dynamics. This shown by the very high correlation coefficient of 0.939. The overall volume of tweets contains some large

peaks, in particular at the start of the period (from April to June 2017). It is then far more consistent after this period, with only minor random perturbations. The range of tweets sent each day is 17,726 to 66,578, with a mean of 30,531.

The *percentage* of tweets which are Islamophobic (shown in panel B) follows a similar pattern to the overall *number* of tweets (shown in panel A). This is intriguing as it suggests that in periods when a high volume of tweets is sent overall, then Islamophobic tweets (which, *ceteris paribus*, would increase only proportionally with volume), take up a greater proportion of the overall volume – and so *very* large spikes in the actual volume of Islamophobic tweets occur. The correlation between the total volume of tweets and the number of Islamophobic tweet sent each day is 0.71, which is reasonably strong but also suggests that there is a large aspect of the number of Islamophobic tweets which is not dependent on the overall volume of tweets.

Visual inspection of panel C shows that the fluctuations in the total volume of tweets holds for each of the parties, which all follow broadly similar peaks and troughs. To quantify this, I take the total volume of tweets sent by followers of each party on each day and then calculate the correlation coefficient for each pair of parties. The average of these correlations is only 0.518, which suggests that the relationship across all parties is only moderate. However, some of the correlation coefficients for each pair of parties are far higher. For UKIP and the BNP, the coefficient is 0.677 and for UKIP and the Conservatives it is 0.687.

Visual inspection of panel D suggests that all of the parties follow different dynamics. Whilst they might vary in volume (shown both here and in the previous section), they follow similar temporal dynamics, with the increases and decreases in the volume of Islamophobia closely-aligned. To quantify this, I take the volume of Islamophobic tweets sent by followers of each party on each day and then calculate the correlation coefficient

for each pair of parties. This is shown in Table 30. Parties are far more aligned in terms of the volume of Islamophobic tweeting than the volume of tweets overall: the average of the correlation coefficients in Table 30 is 0.808, compared with 0.518 for the overall volume of tweets by party (an increase of 0.29). The correlation between UKIP and Conservatives is highest at 0.899. Indeed, the correlation coefficients for the BNP, UKIP and Conservatives are all high (from 0.838 to 0.862) and the biggest gap is between them and Labour. In particular, the lowest correlation coefficient is for the BNP and Labour at just 0.718. Nonetheless, overall, this suggests that similar factors drive the temporal dynamics of Islamophobic tweeting across all parties. In the following section I provide detailed analysis of a key driver of Islamophobic tweeting – Islamist terrorist attacks – and investigate the extent to which this driver affects followers of different political parties.

	BNP	**UKIP**	**Conservatives**	**Labour**
BNP	1			
UKIP	0.838	1		
Conservatives	0.862	0.899	1	
Labour	0.718	0.789	0.738	1

Table 30, Correlation of volume of Islamophobic tweets sent by followers of each party on each day

7.3 | Impact of Islamist terrorist attacks on Islamophobic tweeting

The goal of this subsection is to investigate how Islamist terrorist attacks are related to the prevalence of Islamophobic behaviour on Twitter. Four Islamist terrorist attacks occurred during the period studied in the UK: Westminster attack on 23rd March 2017, Manchester Arena on 22nd May 2017, London Bridge on 3rd June 2017 and Parsons Green on 15th September 2017. Further details about the attacks are provided in Appendix 7.2. Prior to the period covered there were no terrorist attacks which could bias results (the most recent prior Islamist attack occurred on the 22nd May 2013 and far right attack occurred on the 16th June 2016).

Figure 20 shows the total volume of Islamophobic tweets sent each day, with solid grey vertical lines to show when UK Islamist terror attacks occurred (n = 4). Terrorist attacks in Europe (n = 12) and the USA (n = 1) are shown with dotted grey lines. Visual inspection shows a strong relationship between the occurrence of terrorist attacks and peaks in Islamophobia. I opt to not focus on non-UK terrorist attacks as Figure 20 shows they are less closely related to peaks in Islamophobia. The occurrence of terrorist attacks appears bursty, with the time differences between attacks potentially following a long-tailed power law distribution (Goh & Barabási, 2008; Karsai, Kaski, Barabási, & Kertész, 2012). However, there is insufficient data to test this quantitatively. Noticeably, there is a long stable period in the last three months of the data (from 1st December 2017 to 28th February 2018) during which Islamophobia varies little and there are few Islamist terror attacks.

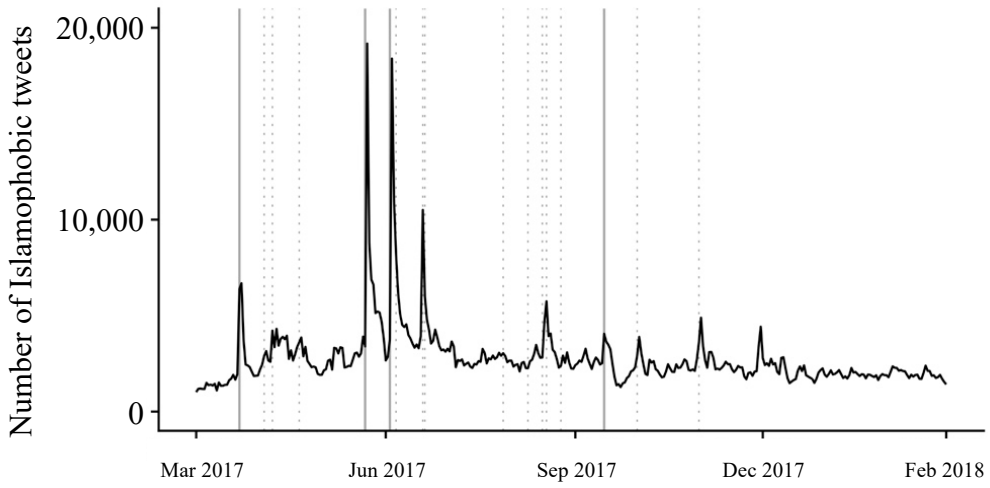


Figure 20, Islamist terror attacks from 1st March 2017 to 28th February 2018

Studying the impact of terror attacks on the prevalence of Islamophobia is difficult as the overall volume of tweets varies considerably. Figure 21 shows the changes in Islamophobia before and after the four UK Islamist terror attacks in the UK. On all three panels, the grey lines represent Islamophobic tweets and the black line shows the average of these. Note that the lines are calculated for followers of all parties in aggregate. The plots all show a period of 15 days, with '0' denoting the day of the attack.

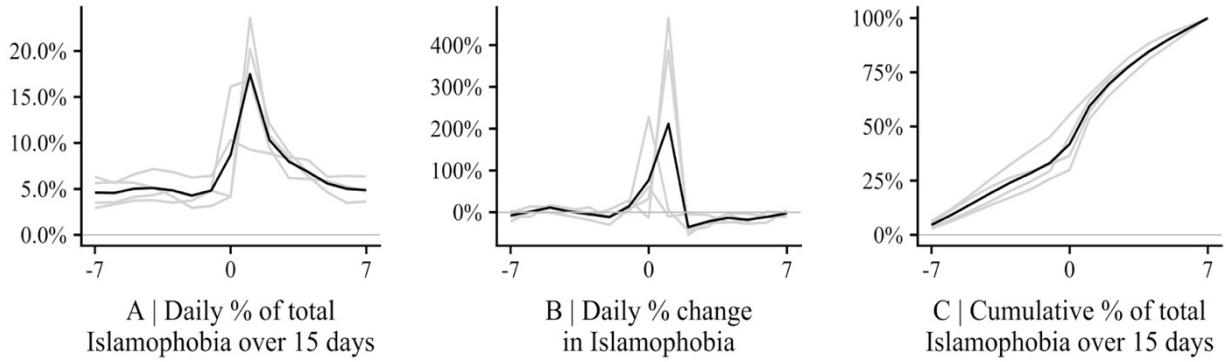


Figure 21, Changes in Islamophobia before and after Islamist terror attacks in the UK

Panel A shows how many of the Islamophobic tweets which are sent over the 15-day period are sent on each day – for instance, on day +1 approximately 16% of all

Islamophobic tweets sent during the 15 days are sent. Noticeably, day +7 is similar to day -7, which suggests that whilst the impact of terrorist attacks is large, it is also reasonably short-lasting. Panel B shows the percentage change in the number of Islamophobic tweets over the period, calculated on a daily basis. Note that the y-axis runs to 400%, showing the magnitude of the changes which occur. For the four attacks (shown in grey), there is a difference of one day for when the prevalence of Islamophobia peaks, which is likely due to the timing of the terror attacks. Parsons Green occurred in the morning, Westminster in the afternoon and both Manchester and London Bridge in the evening – for the latter two, many of the Islamophobic tweets were sent the day afterwards. Panel C shows the cumulative percentage of Islamophobic tweets sent during the 15-day period. There is a subtle sigmoid or 'S-shape' to this curve. This shows the same dynamic as the other panels; the level of Islamophobia surges suddenly when the terrorist attack happens but then decreases sharply too, returning to its baseline level in a matter of days. These results provide strong preliminary evidence that Islamist terror attacks have a marked effect on the overall level of Islamophobia which is expressed.

In all three panels there is one grey line which is a noticeable outlier. This is for an attack with a very different dynamic to the other three: the Parsons Green attack on 15th September 2017. There are two noticeable differences regarding this terrorist attack. First, no individuals were killed (Although 30 were injured). Second, preliminary analysis of the Nexis news database suggests that there was far less media coverage. This is a very interesting outlier and raises important questions about *why* this particular terrorist attack did not generate an Islamophobic response – particularly as it was committed by a prime target for far right hatred: a young Iraqi man who had arrived in the UK illegally with links to ISIS.

In Appendix 7.2, I provide in-depth visual analysis of Islamophobic tweeting during each of the four Islamist terrorist attacks. These are not shown here for brevity.

7.3.1 | Segmented regression model overview

Qualitative inspection of Figures 19, 20 and 21 highlight the dramatic impact of terrorist attacks in driving Islamophobia. To quantify their impact, I fit a segmented regression (also known as 'piecewise' regression) model. This is a widely-used tool for quasi-experimental interrupted time series analysis (Kontopantelis, Doran, Springate, Buchan, & Reeves, 2015; Mcdowall, Mccleary, Meidinger, & Hay, 1980; Wagner, Soumerai, Zhang, & Ross-Degnan, 2002). It enables researchers to estimate the impact of a phase change, typically due to the occurrence of an event, in a time series of data by fitting different model coefficients for different time phases, separated by breakpoints. A key advantage of segmented regression is that researchers can calculate the *change* in parameters, such as the slope. This makes inferences more robust by accounting for the underlying trend prior to the studied event (Bernal, Cummins, & Gasparrini, 2017). Segmented regression will allow me to see how the rate of Islamophobic tweeting varies at different points during, before and after Islamist terrorist attacks. The dataset used in this Chapter meets the requirements of segmented regression: the data has been collected regularly over time, is organised in equally spaced intervals, and there are sufficient data points to ensure both statistical power and seasonal/circadian rhythms are accounted for (Wagner et al., 2002). Segmented regression has been used effectively in similar previous work, such as Garcia-Gavilanes et al.'s work modelling attention decay on Wikipedia following plane crashes (García-Gavilanes, Tsvetkova, & Yasseri, 2016).

I study an equal number of days before and after the *peak* of Islamophobia following each attack – rather than the date/time of the attack itself. This is because the attacks occur at

different times of day, and as such the peak of Islamophobia happens at different points; in two cases it occurs on the same day and in the other two it occurs on the subsequent day. Basing the time period around the peak of Islamophobia ameliorates this issue and ensures that the results are comparable across all four attacks. For each attack, I take a period of 11 days prior and subsequent to the peak of Islamophobia. The choice of 11 days is driven by the data; 12 days after the peak in Islamophobia following the Manchester Arena attack the London Bridge attack occurred. As such, using a longer time period would confound the estimates as the model would also incorporate the impact of this event. However, this nonetheless poses the problem that the period prior to the London Bridge attack does not constitute a fair comparison as it is also the aftermath of the Manchester attack. I resolve this issue by simulating data for this part of the London Bridge time series, after conducting appropriate robustness checks.[20] This is discussed in Appendix 7.2.

Given the high correlation between weak and strong Islamophobic tweets, I use the binary classifier, in which weak and strong Islamophobic tweets are collapsed together. As already discussed, Islamophobia within tweets can be measured in different ways, primarily as either (i) a *probability* or (ii) a *count*. I fit models for both measurements and find that they follow very similar temporal dynamics (reported in Appendix 7.2). I focus on the count of Islamophobia as (i) it is more interpretable and (ii) it is the most pressing practical and policy-focused measurement for understanding Islamist terrorist attacks, given that it is the overall *volume* of Islamophobic tweeting which is likely to inflict the greatest harm on Muslims, rather than the proportion of tweets.

[20] Terrorist attacks are unexpected and unpredictable events. As such, there is no need to adjust the time series for the period *after* the Manchester arena attack as users are entirely unaware of the impending London Bridge attack.

Poisson regression is often used to model count data as it has greater power for this task than OLS regression (Warton, Lyons, Stoklosa, & Ives, 2016). One limitation of Poisson regression is that coefficient estimates are biased when the data is 'overdispersed', which is when a variable's variance is greater than its mean (J. Martin & Hall, 2016). Overdispersion is accounted for in the negative binomial model by the inclusion of a dispersion term. In the present work, the conditional variances for the counts of Islamophobic tweets (subset by each party) are considerably greater than the conditional means and, accordingly, I use a negative binomial model. This has been used in very similar previous work and is well-suited to Twitter data (Burnap et al., 2014; King & Sutton, 2013). It is not necessary to use a zero-inflated negative binomial regression model as none of the time periods have 0 tweets (J. Martin & Hall, 2016).

I fit negative binomial segmented regression models for three different time granularities: 1,000, 10,000 and 100,000 seconds.[21] All three models follow very similar dynamics and have similar coefficients (shown in Appendix 7.2). I focus on 10,000 seconds for the remainder of this analysis as 100,000 seconds (~27 hours) is insufficiently granular whilst 1,000 seconds is both computationally inefficient and the model does not perform as well. I use three breakpoints in all of the models – this is a key hyperparameter (Jaromir Antoch, Jan Hanousek, Lajos Horvath, Marie Huskova, 2017), and I fit three breakpoints by both qualitatively inspecting the time series and applying Bai and Perron's breakpoint algorithm, implemented using the 'breakpoints' function from the 'strucchange' package in R (Bai & Perron, 2003; Zeileis et al., 2015).

[21] 1,000 seconds is approximately 17 minutes, 10,000 seconds is approximately 2.7 hours, and 100,000 seconds is 27 hours (just over 1 day).

7.3.2 | Segmented regression model results

Results of model fitting for time granularity of 10,000 seconds are shown in Table 31. The simplest model (Model 1) is plotted in Figure 22.[22] Four more complex models, which include terms for the number injured and the number killed, are also reported. The pseudo R-squared values are high for all models, ranging from 0.529 to 0.788 (using the Cox-Snell estimation). The five models have very similar coefficients for the slopes, and very similar break points. This indicates that the underlying temporal dynamics which drive Islamophobia following terrorist attacks are consistent. The impact of the number killed is positive in all models, showing that the level of Islamophobia is driven by the nature of the Islamist terrorist attack; attacks which inflict more harm motivate a more powerful response. The number injured has a more ambiguous association with Islamophobia. In model 3, it is associated with a small increase but in models 4 and 5 it is associated with a small decrease. In these models, the coefficient for the number killed is higher than in model 2 and the pseudo R-squared values are higher (by ~0.1); the number killed has the greatest impact when controlling for the number injured. This is most likely because the ratio of number killed to number injured varies considerably across attacks, and the attack where no people are killed but 30 are injured (Parsons Green) is associated with a far smaller increase in Islamophobia than the others. Model 5, which includes an interaction term for the number injured and number dead, is very similar to model 4. Given that this term increases model complexity, I focus on model 4: time + the number killed + the number injured.

[22] To plot the data, I take the average count of Islamophobic tweets in each 10,000 second time interval (calculated over the four attacks), and show the average fitted line (also calculated over the four attacks).

Statistic	MODEL 1[23]	MODEL 2	MODEL 3	MODEL 4	MODEL 5
Estimated breakpoints	-10, 0, 30	-10, 0, 30	-10, 0, 30	-10, 0, 30	-10, 0, 30
y-intercept	331 *****	261 *****	257 *****	433 *****	862 *****
Number Killed	/	1.025 *****	/	1.138 *****	1.134 *****
Numbed Injured	/	/	1.0035 **	0.979 *****	0.961 *****
Number Dead: Number Injured	/	/	/	/	1.001 *****
Slope 1	1.002 *	1.0017 *	1.0018 *	1.002 *	1.002 **
Slope 2	1.524 *****	1.564 ***	1.552 **	1.488 ***	1.477 *****
Slope 3	0.965 *****	0.969 *****	0.965 *****	0.969 *****	0.964 *****
Slope 4	0.993 *****	0.995 *****	0.994 *****	0.995 *****	0.994 *****
Change vs. slope 1	1.520	1.561	1.549	1.486	1.474
Change vs. slope 2	0.633	0.619	0.629	0.652	0.652
Change vs. slope 3	1.029	1.027	1.029	1.027	1.031
Breakpoint 1	-7.207 *****	-6.841 *****	-6.963 *****	-6.998 *****	-7.001 *****
Breakpoint 2	-3.698 *****	-3.707 *****	-3.630 *****	-3.613 *****	-3.359 *****
Breakpoint 3	29.556 *****	-36.673 *****	30.418 *****	36.999 *****	29.733 *****
Convergence	65 iterations	30 iterations	10 iterations	1,350 iterations	1,399 iterations
Pseudo r-squared (Cox Snell)	0.529	0.648	0.593	0.771	0.788

[23] All of the reported coefficients or slope, intercept and the other terms, both here and in the appendixes, are exponentiated from the underlying values – which model the change in the log of the dependent variable. For instance, for the y-intercept estimated by Model 1 is 5.801. Exponentiated, this is 330.63, which can be rounded to 331. Exponentiation raises 2.718282 to the reported value. In this case, the calculation is: 2.718282 ^ 5.801, which is 330.63. Note that negative values exponentiate to less than 1.

Pseudo r-squared (Nagelkerke)	0.529	0.648	0.593	0.771	0.788
Pseudo r-squared (Pearson)	0.42	0.536	0.496	0.644	0.640
Dispersion parameter	3.6483	4.236	3.9423	5.3194	5.2699

Table 31, Summary of negative binomial segmented regression models

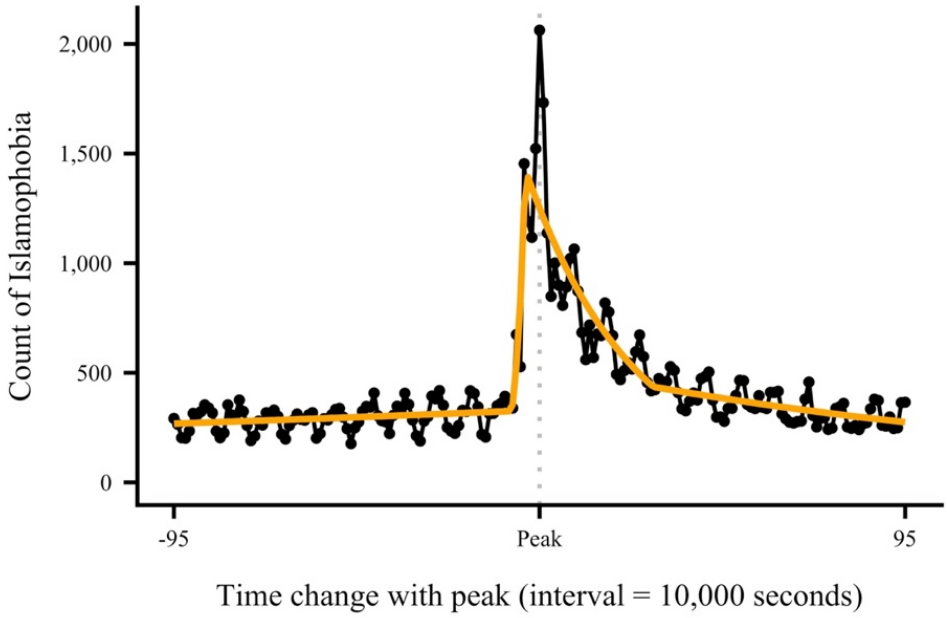


Figure 22, Negative binomial segmented regression model with time granularity of 10,000 seconds (model 1)[24]

[24] Note that the colour orange is used for the segmented regression line because this colour has not been used elsewhere in the Chapter.

7.3.3 | Dynamics of Islamophobic tweeting during Islamist terrorist attacks

The initial slope in model 4 has an exponentiated coefficient of 1.002. This means that for every 10,000 seconds that passes there is an average increase in the count of Islamophobic tweets of 0.2%. This is a very small increase, and the slope is insignificant. As such, I characterise this period as unchanging; in periods when there is not an Islamist terrorist attack, fluctuations in the number of Islamophobic tweets which are sent is largely random, moving only slightly from the y-intercept. I validate this argument by studying the period from 1st December 2017 to 28th February 2018 – as shown above in Figure 19, during this 3-month period no Islamist terrorist attacks take place. I take a time granularity of 10,000 seconds and fit a negative binomial regression model for the number of Islamophobic tweets against time (full results are not shown here for brevity). Both the y-intercept and slope coefficient are significant ($p < 0.0001$) but the slope coefficient is only -0.0002321. When exponentiated, this indicates that for each 10,000 second interval that passes there is a percentage decrease in the number of Islamophobic tweets sent of just 0.02%. This provides strong evidence that during periods when there are no Islamist terror attacks (including the periods just prior to an Islamist terror attack), the level of Islamophobia does not fluctuate considerably.

The first breakpoint is 7 time periods before the peak in Islamophobia. This can be understood as the number of time periods it takes from the terrorist attack occurring for the number of Islamophobic tweets sent to peak. After this breakpoint the exponentiated slope is 1.488. This means that for every 10,000 second interval the number of Islamophobic tweets increases by 48.8%, or almost half. This can be interpreted as a super linear scaling factor, showing the proportional rate of change. Crucially, this means that the rate of change is compounded; the 48.8% growth in the volume of tweets is

multiplicative with each new time period. In effect, the rate of Islamophobia is not only increasing but also *accelerating* during this phase (L. M. A. Bettencourt, Lobo, Helbing, Kuhnert, & West, 2007; Luís M.A. Bettencourt, 2013; West, Brown, & Enquist, 1999). Even though this phase lasts just 3.5 time periods (~11 hours), the level of Islamophobia quadruples.

The second breakpoint is at 3.6 time periods before the peak in Islamophobia – the fact that this is before the peak is an artefact of the modelling process and this point can be understood conceptually as the Islamophobic peak. After this breakpoint, the coefficient of the slope is 0.969 (a change of 0.652) – crucially, this is a change in sign from positive to negative, marking the start of a long period in which the volume of Islamophobia decreases. After ~40 time intervals (~4.5 days) there is another breakpoint when the rate of deceleration slows and continues for ~60 time intervals (~7 days). During this period the exponentiated slope coefficient is 0.995. Note that the slowing rate of deceleration is also indicated by the positive *change* in slope (the exponentiated coefficient of which is 1.027).

The overall cycle of Islamophobia can be summarised as follows. First, a period of baseline Islamophobia where there are only minor fluctuations in the level of Islamophobia. Then, a short period of rapid acceleration (approximately 1 day) in which the volume of Islamophobia quadruples. Third, there is an extended period of deceleration, consisting of a period of stronger deceleration (~4.5 days) and then a longer period of less intense deceleration (~7 days). By the end of this period, the level of Islamophobia returns to approximately the baseline level at the start – surprisingly, the baseline does not increase in the aftermath of attacks. This cycle of escalation/de-escalation is shown in Figure 23. Note that this cycle of Islamophobic tweeting might be

interrupted by the occurrence of multiple terrorist attacks in short succession – as with the Manchester arena and London Bridge attacks.

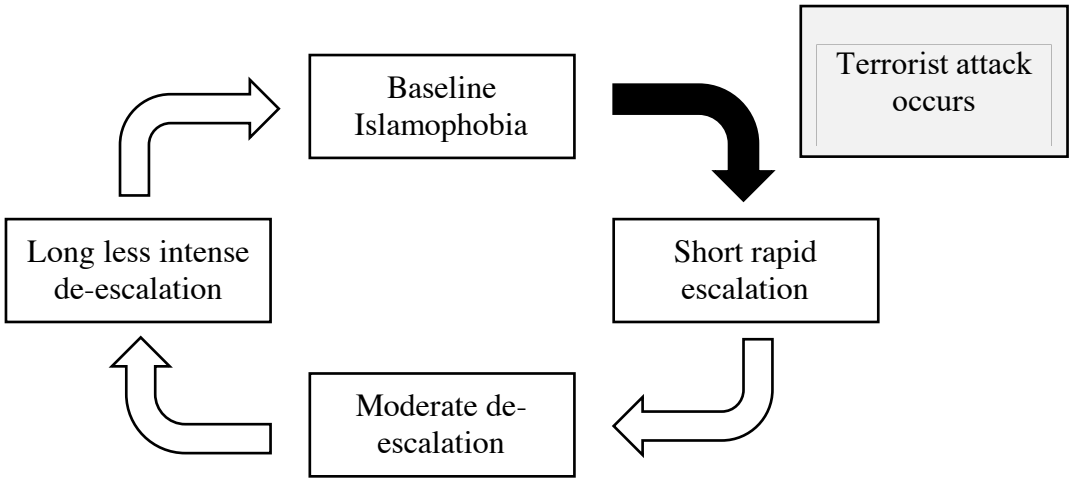


Figure 23, Typical progression of Islamophobia following a terrorist attack

7.3.4 | Impact of party followership

In this subsection I estimate the impact of party followership for two of the models analysed in the previous section: model 1 (which only models the impact of time) and model 4 (which models time + the number killed + the number injured). Coefficients for these models (6 and 7, which are models 1 and 4 with Party included) are reported in Table 32. All of the coefficients in both models are significant, and the pseudo R-Squared values are far higher than in models 1 and 4 (0.676 and 0.799 versus 0.529 and 0.771). The slope coefficients and breakpoints indicate the models follow broadly the same temporal dynamics as outlined above. The y-intercepts are lower in models 6 and 7 than models 1 and 4 (140 and 227 vs. 331 and 433). This is because the dependent variable in the models is the count of tweets for just each party rather than all parties combined, and so the total is far lower. Model 7 has considerably better fit than model 6 (pseudo r-squared of 0.799 versus 0.676), which is in line with expectations given the dramatic

improvement in fit from model 1 to model 4. Model fitting details are reported in Appendix 7.2.

Statistic	MODEL 1	MODEL 4	MODEL 6	MODEL 7
Estimated breakpoints	-10, 0, 30	-10, 0, 30	-10, 0, 30	-10, 0, 30
y-intercept	331 *****	433 *****	140	227 *****
Number Killed	/	1.138 *****	/	1.172 *****
Numbed Injured	/	0.979 *****	/	0.973 *****
Party (UKIP)	/	/	0.5886 *****	0.554 *****
Party (Conservatives)	/	/	0.245 *****	0.226 *****
Party (Labour)	/	/	0.259 *****	0.247 *****
Slope 1	1.002 *	1.002 *	1 N.S.	1.0007 N.S.
Slope 2	1.524 *****	1.488 ***	1.154 *****	1.144 *****
Slope 3	0.965 *****	0.969 *****	0.962 *****	0.964 *****
Slope 4	0.993 *****	0.995 *****	0.992 *****	0.994 *****
Change vs. slope 1	1.520	1.486	1.154	1.143
Change vs. slope 2	0.633	0.652	0.834	0.842
Change vs. slope 3	1.029	1.027	1.031	1.031
Breakpoint 1	-7.207 *****	-6.998 *****	-9.465 *****	86.267 *****
Breakpoint 2	-3.698 *****	-3.613 *****	-0.008 *****	95.992 *****
Breakpoint 3	29.556 *****	36.999 *****	29.194 *****	125.897 *****
Convergence	65 iterations	135 iterations	7 iterations	6 iterations
Pseudo r-squared (Cox Snell)	0.529	0.771	0.676	0.799

Pseudo r-squared (Nagelkerke)	0.529	0.771	0.676	0.799
Pseudo r-squared (Pearson)	0.42	0.644	0.488	0.579
Dispersion parameter	3.6483	5.3194	2.1156	2.5391

Table 32, Negative binomial segmented regression models with party followership

In model 7, the y-intercept is 227, which represents the baseline of Islamophobic tweets sent during each 10,000 second time period by the BNP. The exponentiated coefficients for the parties are: 0.554 for UKIP, 0.226 for Conservatives and 0.247 for Labour. The coefficients show that the values for UKIP are approximately half of the BNP, and the values for Conservatives and Labour, in turn, approximately half of UKIP. The differences between the coefficients for each party are remarkably similar to the differences shown in the OLS regression models in the previous section. This suggests that the impact of party on the level of Islamophobia is broadly similar during both periods around Islamist terrorist attacks and other periods. Thus, whilst behaviour changes considerably around Islamist terrorist attacks (showing a huge spike straight away after) the differences between parties are consistent. This is somewhat expected given the high correlation coefficients for the volume of Islamophobic tweets sent each day by followers of each party (reported above).

The fitted values from model 7 are plotted in Figure 24. Note that the scales vary, and that the peak for the BNP is approximately five times greater than that of both Conservatives and Labour. A set-scale figure is shown in Appendix 7.3. Noticeably, the dynamics of UKIP and the BNP are very similar, with a very sharp and high peak in Islamophobia followed by a two-phase period of de-escalation (as described above). In contrast, for Conservatives and Labour the peaks are less high (relative to the starting level of Islamophobia) and less sharp. The de-escalation process appears sharper, particularly for Labour, suggesting that the impact of Islamist terrorist attacks is far

shorter for followers of these parties. Overall, the process of escalation and de-escalation in Islamophobia holds for each party.

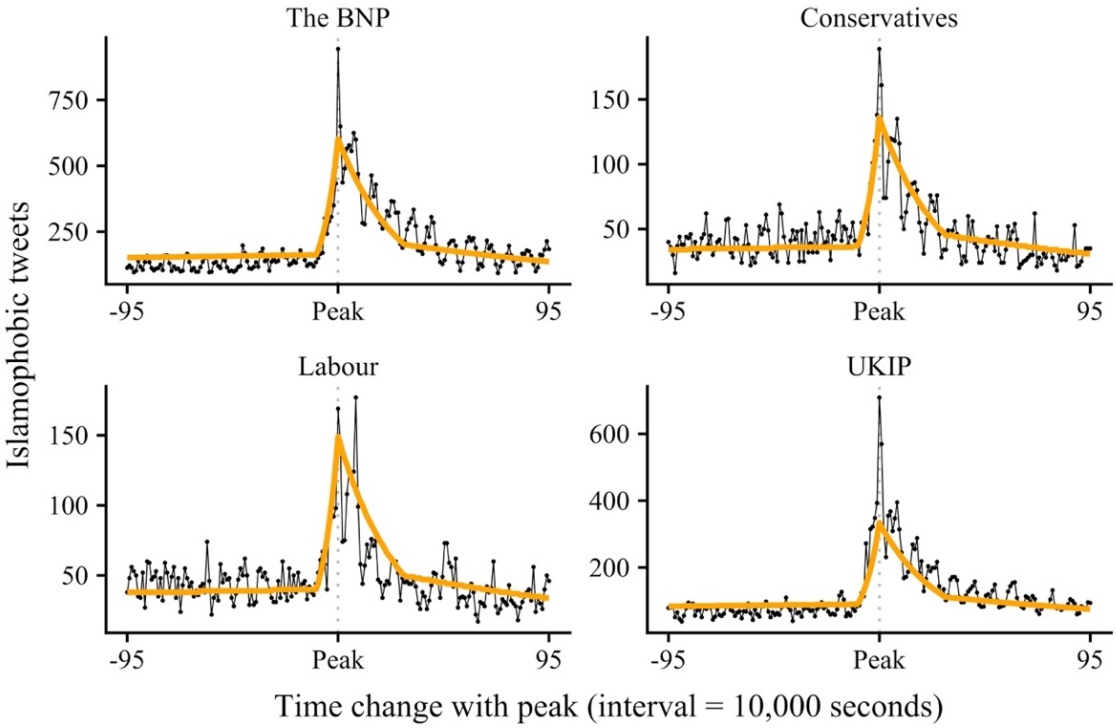


Figure 24, Model 7 fitted values for each party

7.3.5 | Confounding: the media's impact on Islamophobia

There is a risk of confounding with the variables for the number killed and the number injured, whereby the number killed/injured might motivate a greater and more incendiary media response – which could, in turn, be the actual driver of more Islamophobia. Accordingly, I collect a dataset of news stories about terror incidents and examine its relationship with the prevalence of Islamophobia. I search for the wildcard term 'terror!' in the Nexis database of news stories, covering the period 1st March 2017 to 22nd February 2018. After cleaning, the dataset consists of 13,814 unique news stories. Figure 25 shows the number of news stories for each day. This follows a similar dynamic as the number

of Islamophobic tweets sent each day. The correlation coefficient is 0.83, which indicates a very strong relationship.

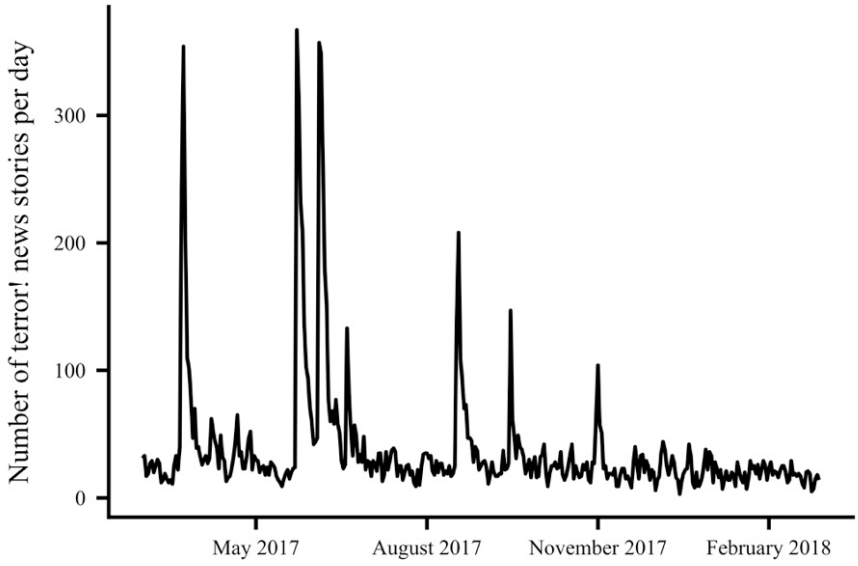


Figure 25, Number of 'terror!' news stories per day

To identify potential causalities between the number of 'terror!' news stories and the number of Islamophobic tweets sent each day, I cross-correlate the values by introducing a lag of 1 to 7 days in both directions. The results are plotted in Figure 26.

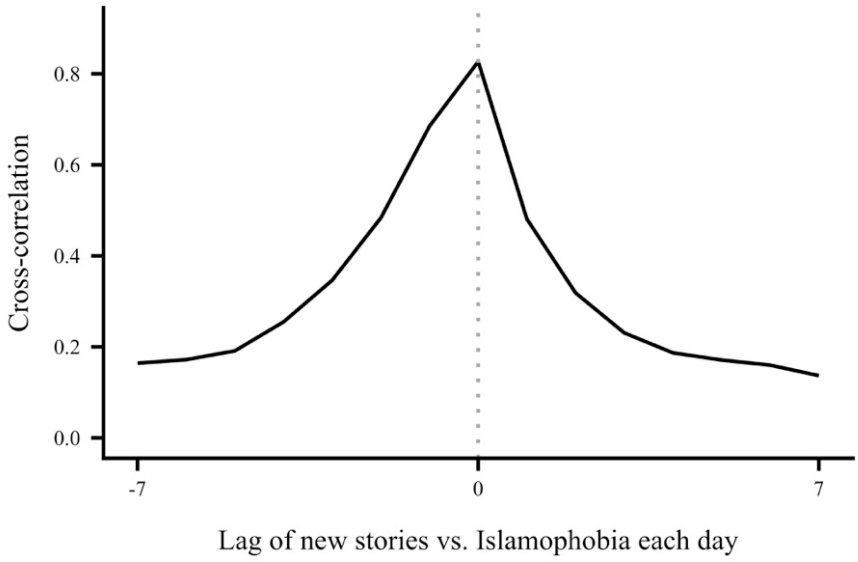


Figure 26, Number of 'terror!' news stories cross-correlated with the number of Islamophobic tweets

The cross-correlation values show that there is only a very weak direction to the relationship between news stories and Islamophobic tweeting. The 'forward' correlation curve drops off more sharply than the 'backward' correlation, which suggests that increases in the number of news stories occurs before increases in the number of Islamophobic tweets. This is also demonstrated by the mean correlation for the 1 to 7 days lag *prior* to the attack, which is 0.32, compared with the mean correlation for the 1 to 7 days *after* the attack, which is 0.24. However, overall, there is no clear relationship between the variables and the evidence is not strong enough to point to a causal relationship. The time granularity is daily, and it may be that the relationship can be identified when a shorter time period, e.g. 1 hour, is studied. However, this level of granularity is not available in the news stories dataset. It is also plausible that both the number of Islamophobic tweets and the number of 'terror!' news stories are driven by the same factor, primarily the occurrence of Islamist terrorist attacks. This is demonstrated by the correlation of 0.81 between the number killed and the number of news stories.

Overall, there is a strong relationship between (i) the number of terror-related new stories, (ii) the number of Islamophobic tweets and (iii) the number of people killed in each terrorist attack. As anticipated, news reporting on terrorism is driven strongly by the occurrence of terrorist attacks. It is not possible to identify a causal relationship between news stories and Islamophobic tweets – and it is likely that both are driven by the occurrence of Islamist terrorist attacks. As such, I do not include the number of terror-related news stories in any of the models as this is likely to confound the results and could introduce multicollinearity.

7.4 | Changes in user behaviour following terrorist attacks

The results so far indicate that the occurrence of Islamist terrorist attacks has a significant and substantial impact on the number of Islamophobic tweets, and that this effect holds across all parties. In this section, I deepen this analysis by investigating whether not only the volume of Islamophobic tweets increases but also whether *which users* are tweeting changes. I identify users who *only* send an Islamophobic tweet on the day of an attack (or in the case of London Bridge and Manchester Arena, just the day immediately after) and not at any other point during the period studied. I term these users *one-off* Islamophobic tweeters. I study a period of just one day as the previous modelling shows that this is sufficient to capture the most active period of activity when the level of Islamophobia rapidly accelerates. The number of one-off Islamophobic tweeters on a typical day is ~6 (out of the total sample size of 15,253 users), and for the terrorist attacks are as follows. For the Westminster terrorist attack there are 23 users, for Manchester Arena there are 53 users, for London Bridge there are 76 users and for Parsons Green there are 5 users. I check whether these values are significant through appropriate statistical significance tests (reported in Appendix 7.2). The first three attacks are extremely significant ($p < 0.000001$) but Parsons Green is not.

The total of 157 (23 + 53 + 76 + 5) one-off Islamophobic tweeters for the four terrorist attacks captures those who only tweet Islamophobically on the peak day following *each* attack. I therefore conduct a second analysis which accounts for users who only tweet on the peak days for all attacks *combined*. This ensures that a user who tweets Islamophobically twice during the year – but in both cases, only during terrorist attacks – is correctly identified as a user whose Islamophobia is driven by terrorism. 169 users are one-off Islamophobic tweeters during all terrorist attacks. This compares with an average of 14 one-off Islamophobic tweeters during other 4-day periods. Appropriate

statistical significance tests show that the count of 169 one-off Islamophobic tweeters is extremely significant (outlined in Appendix 7.2). The 169 one-off Islamophobes breaks down as 34 users for the BNP, 42 for the Conservatives, 56 for Labour and 37 for UKIP.

Figure 27 shows the relationship between terrorist attacks and party followership. Panel A shows the proportion of each party's followers who are a one-off Islamophobic tweeter each day, comparing terrorist attacks (in bold) with other periods. The sharp increase for all parties demonstrates the impact of terrorist attacks. Panel B shows the size of the change in the prevalence of non-Islamophobic tweeters. In all cases, the increase is considerable. BNP has the lowest prevalence of one-off Islamophobes during terrorist attacks (~0.3%) and the second lowest increase – this is expected, as followers of the BNP are more likely to repeatedly engage in Islamophobia rather than commit one-off acts. UKIP, unexpectedly, has the second greatest prevalence of one-off Islamophobes during terrorist attacks (~0.4%) and the greatest increase (~1,000%). UKIP is impacted considerably by terrorist attacks, as during the attacks a comparatively large proportion of otherwise non Islamophobic users are 'activated' as one-off Islamophobes.

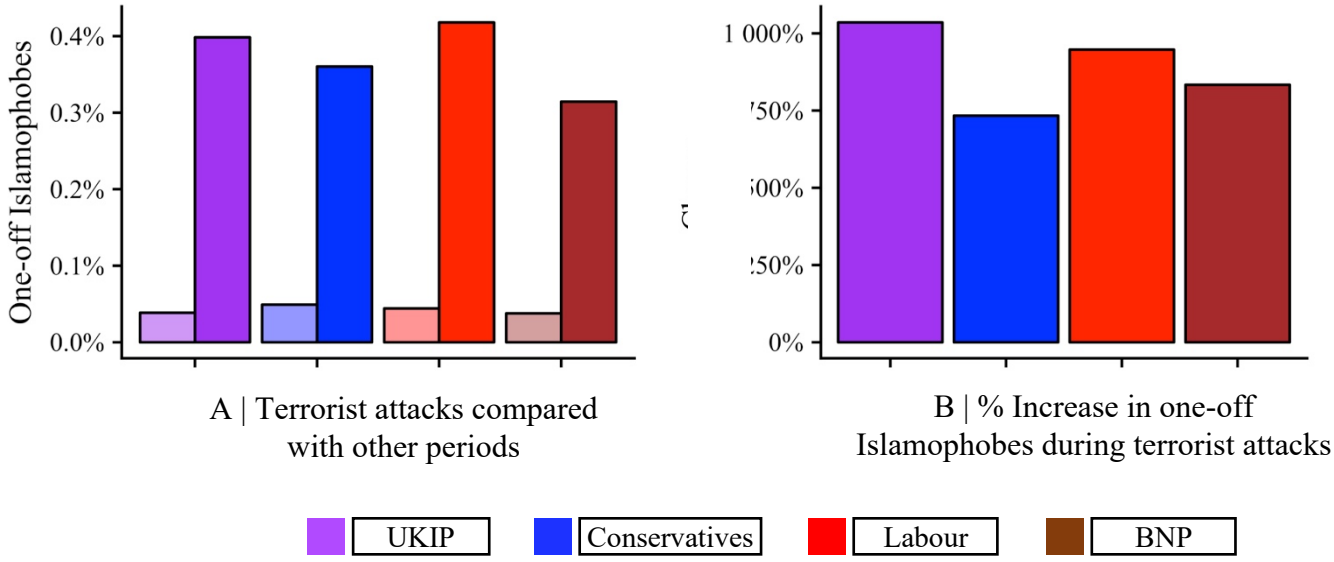


Figure 27, The relationship between terrorist attacks and the number of one-off Islamophobic tweeters, split by party

7.4.1 | Impact of the volume of Islamophobic tweets

During the periods when Islamist terrorist attacks occur a far larger volume of tweets are sent. It is plausible that the volume of tweets is related to the number of one-off Islamophobic tweeters – which could confound the observed relationship with terrorist attacks. I fit negative binomial regression models to the data to investigate this further. The models are reported in Table 33. Model 8 shows the number of one-off Islamophobic tweeters regressed against the total volume of Islamophobic tweets, calculated for each day for all parties combined. The number of tweets has a positive and statistically significant relationship with the number of one-off Islamophobes and the pseudo R-squared is high at 0.66. To evaluate the impact of Islamist terrorist attacks (above the increase in the volume of Islamophobic tweets sent each day), I create a second model (model 9) with a dummy variable for the day of the peak in Islamophobic tweeting following a terrorist attack. In this model the impact of Islamist terrorist attacks is negative (with an exponentiated coefficient of 0.928) and the pseudo R-squared is only slightly improved compared with model 8. Furthermore, the AIC of model 9 is higher and a chi-squared Analysis of Variance (ANOVA) test shows it is not a significantly better fit than model 8.

These results show that the impact of terrorist attacks on the number of new Islamophobic tweeters disappears once the increased overall volume of tweets is accounted for. To summarize: (i) Islamist terrorist attacks drive a large increase in the volume of Islamophobic tweets, (ii) there is a proportional increase in the number of one-off Islamophobic tweeters but (iii) this increase is *not* greater than other periods when the volume of Islamophobic tweets increases which means that (iv) there is not a special 'extra' increase in the number of one-off Islamophobes due to terrorist attacks – even though (v) the overall increase in the number of one-off Islamophobes is substantial. Put

simply, terrorist attacks increase the number of one-off Islamophobes but do not introduce new *dynamics* of one-off Islamophobic tweeting.

Note that whilst the coefficient for the impact of the number of tweets is very low (1.002), (i) this is multiplicative, (ii) the y-intercept is also very low (~2.4) so the multiplicative period is very long and (iii) the range of Islamophobic tweets sent each day is considerable, from 1,011 to 19,175. These factors should be considered when interpreting the model coefficients.

Statistic	MODEL 8	MODEL 9
y-intercept	2.417 *****	2.399 *****
Coefficient (Number of Tweets)	1.0002 *****	1.0002 *****
Islamist terrorist attack	/	0.928 N.S.
AIC	1618.6	1620.5
Pseudo r-squared (Cox Snell)	0.658	0.660
Pseudo r-squared (Nagelkerke)	0.662	0.662
Pseudo r-squared (Pearson)	0.715	0.718
Dispersion parameter	12.6947	12.801

Table 33, Negative binomial regression models for the number of new Islamophobic tweeters versus the number of Islamophobic tweets

7.4.2 | Changes in the distribution of Islamophobic tweets per user

To further understand *who* tweets during Islamist terrorist attacks, I measure the inequality of the distribution of Islamophobic tweets per user (for the four days of Islamist terrorist attacks combined), at the party level, by calculating the Gini coefficient. These coefficients cannot be compared with the coefficients for the number of Islamophobic tweets per user reported earlier in Table 26 as the values in this table (i) include the four peak days during terrorist attacks (and so are not a real comparison) and (ii) are not calculated over four days but a full year (365 days), which likely impacts the calculated values. To ensure a fair comparison I take four day combinations from the 90-day period

from 1st December 2017 to 28th February 2018 during which no Islamist terrorist attacks occur. There are ~2.5 million unique combinations of four days. From this, I then sample 10% of the combinations (n = 250,000) and calculate the mean Gini coefficient for each party. These coefficients are considerably lower than those reported in Table 26 and constitute a far better comparison. The Gini coefficients for each party during both periods are shown in Table 34.

Party	Gini coefficient during other 4-day combinations	Gini coefficient rank	Gini coefficient during terrorist attacks	% increase during terrorist attack
UKIP	0.585	2	0.699	19.55%
Conservatives	0.501	3	0.598	19.37%
Labour	0.486	4	0.526	8.30%
BNP	0.616	1	0.696	13.00%

Table 34, Gini coefficient of Islamophobic tweets per user for the 4 days of peak Islamophobia during Islamist terrorist attacks compared with other 4-day combinations

The Gini coefficient during terror attacks are considerably higher, with percentage increases which range from 8.3% to 19.6%. Statistical testing shows that the distribution of Gini coefficients is normal for each party, and that the values for the terror attacks are significantly different ($p < 0.001$ in all cases). This shows that, compared with other periods, during terrorist attacks a smaller proportion of users are responsible for most of the Islamophobic tweets. This means that even if most users send more tweets during terrorist attacks (which is likely), the increase in the volume of Islamophobic tweets is driven primarily by a small number of highly active Islamophobes.

The Gini coefficients during terror attacks for UKIP and the BNP are almost the same (0.699 and 0.696) – these are the highest coefficients and, as already shown, these are the parties which send the highest volume of Islamophobic tweets. This provides further support that during Islamist terrorist attacks a small number of individuals are responsible

for the large increase in the volume of Islamophobic tweets, rather than many one-off and low-volume tweeters who are activated by the attack. Labour has the smallest percentage increase (8.3%) and the absolute coefficients are also lower. Potentially, terrorist attacks have a more consistent impact on Labour, effecting all users similarly rather than only magnifying the behaviour of a small cadre of hyper active Islamophobic tweeters.

7.5 | Conclusion

This Chapter has sought to answer three research questions:

> RQ 3: To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?
>
> RQ 4: To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?
>
> RQ 5: Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?

7.5.1 | Discussion

RQ 3 | UK political parties and Islamophobic hate speech

Islamophobic behaviour is not only confined to the far right. It can be observed across all parties. There are considerable differences in the prevalence and strength of Islamophobia across followers of different parties. Interestingly, prevalence and strength are associated – parties with more Islamophobia quantitatively (i.e. a greater proportion of their tweets are Islamophobic) also have more Islamophobia qualitatively (i.e. proportionally, there is more strong than weak Islamophobia). At the same time, and contrary to my initial expectations, weak and strong Islamophobia have very similar temporal dynamics in aggregate, with a correlation coefficient of 0.939. As such, the dynamics of Islamophobia can be studied by using the binary classifier developed in Chapter 5, in which the categories are collapsed together – although it is worth noting that differences in the dynamics of weak/strong Islamophobic tweeting might still exist at the individual level (as shown in the previous chapter just for followers of the BNP). Followers of the BNP

have the highest prevalence and strength of Islamophobia, followed by UKIP and then the Conservatives and Labour.

The findings here contribute to two ongoing theoretical debates about the nature of UK political parties. First, is the ideological position of UKIP and the lack of consensus as to how it should be conceptualised and described. Throughout this Chapter, UKIP is consistently shown to be halfway between (i) the BNP and (ii) the Conservatives and Labour in terms of Islamophobic behaviour. Accordingly, I propose that it should be viewed as an ambiguous 'halfway house' between the two poles of the mainstream and the extreme. Aspects of the party, such as the dynamics of its behaviour during Islamist terrorist attacks, are akin to the BNP – but in terms of the overall volume of tweets, it is more in between the mainstream and the far right.

The second theoretical debate on UK party politics this Chapter contributes to is prejudice within mainstream parties, specifically Labour and the Conservatives. Followers of the Conservatives are almost indistinguishable from Labour (as shown above in Figures 16 and 18). This is surprising given recent media coverage of accusations of Islamophobia within the Conservatives (The Independent, 2018). Because Islamophobia manifests across UK political parties, I argue that it constitutes a *twin threat*.

The relationship between party followership and the prevalence of Islamophobia could be explained by several factors. The most likely explanation is that individuals are attracted to different parties because of their policies, discourse and attitudes towards Muslims, including the possibility that they are openly Islamophobic (as with the BNP). At the same time, it may be that their position on Muslims is only indirectly important; users are attracted to parties because of other aspects (such as their position on the welfare state, immigration or crime) but these are related to their position on Muslims. In either case, this suggests that party level differences in the prevalence of Islamophobia are

driven by the type of individuals who follow each party; Islamophobia (potentially, indirectly) drives party followership. Potentially, the opposite relationship may exist: parties exert an effect on those who follow them through exposure to mechanisms of information provision and normative pressure, thereby increasing how Islamophobic they are. This would be an example of insights from structuration theory into the dynamic between structure and agency, best typified by the quote attributed to Churchill that, 'we shape our buildings, but then our buildings shape us' – individuals chose which party they want to follow but then this subsequently exerts an effect on them. Whilst both explanations are plausible, it is most likely that differences in party-level Islamophobia are primarily due to the type of followers who are attracted to each party, as it is unlikely that the social effects exerted by Twitter followership are that strong (as discussed in the literature review in Chapter 2).

Across all parties, the vast majority of Islamophobic tweets are sent by just a few users, as indicated by (i) the long-tailed party distributions for the number of Islamophobic tweets per user and (ii) the high Gini coefficients (for all users over the whole time period, every party's coefficient is greater than 0.83). The fact that this is consistent across all parties suggests that whilst parties may differ in terms of the overall volume of Islamophobic tweets (and the proportion of those tweets which are, respectively, weak and strong), the user distributions are similar: across all parties, just a few users drive the overall volume of Islamophobia.

This provides a simple and well-evidenced answer to RQ3: Islamophobic hate speech exists and manifests amongst both mainstream and extreme party followers but differs considerably in terms of strength and volume.

RQ 4 | Islamist terrorist attacks

My analysis demonstrates that Islamist terrorist attacks have a large but temporary impact on the volume of Islamophobic tweeting. Thus, RQ 4 can be answered simply: there is strong evidence that Islamist terrorist attacks drive a very large but only temporary increase in Islamophobic hate speech amongst followers of UK political parties on Twitter. This evidence both supports and puts into question different aspects of the theory of cumulative extremism, building on the analysis in the literature review (Chapter 2).

The cycle of Islamophobic tweeting I identify aligns with the process proposed by Burnap and Williams in their study of cyberhate following the Woolwich terrorist attacks. They found that the greatest peak in hate speech occurred in the first 24 hours, followed by a 15-day period of de-escalation (Williams & Burnap, 2016). The key advance I make here is that I identify a two-phase period of de-escalation (comprising a ~4.5 day sharp de-escalation and then a longer period of less intense de-escalation lasting ~7 days) and show that, overall, the period of de-escalation is slightly shorter (approximately 11 days). My results also help to clarify other previous research. For instance, in their study of offline hate crimes following different types of political events (including terrorist attacks), King and Sutton found that 'the rate of de-escalation seems nearly as rapid as the pace of escalation, and the increases are generally short in duration; we tend to observe a spike after an event rather than a plateau' (King & Sutton, 2013, p. 888). In contrast, the results presented here (as well as in Burnap and Williams' prior study) show the de-escalation period is far longer. Noticeably, my result (showing it takes 7 days for the level of Islamophobia to return to baseline) is broadly in line with that of Byers et al., who report an 8-day de-escalation period for offline hate crime following the 9/11 terrorist attack (Byers & Jones, 2007). However, these results differ from those of Hanes and Machin (Hanes & Machin, 2014), as I did not find evidence that the baseline of Islamophobia is

higher in the aftermath of the attack – even though my data covers several months after the last attack. This is in-line with the results of other research on how peaks in online behaviour following events then decay, such as Garcia-Gavilanes' study of attention on Wikipedia following plane crashes (Garcia-Gavilanes, Mollgaard, Tsvetkova, & Yasseri, 2017). Thus, this result contributes to a growing body of research into the large but nonetheless temporary impact of certain events on social behaviour (Candia, Rodriguez-sickert, Barabási, & Hidalgo, 2019).

The escalation/de-escalation process of Islamophobia around Islamist terror attacks has implications for the theory of cumulative extremism. First, the temporary but considerable increase in Islamophobic hate speech after an Islamist terror attack supports the broad point that extremism feed off each other, which is in line with other empirical research. However, second, the fact that the baseline of hate speech does not increase considerably suggests that extremism is not accumulating over time; no increase in the overall level of hate is observed. The use of a methodologically individual research design increases the robustness of this observation. Third, the available evidence suggests that when multiple attacks happen in short succession they do not each become *more* Islamophobic but, rather, if anything there is Islamophobia fatigue (see Figure 20) – although far more evidence, taking into account a far larger number of terrorist attacks, is needed to verify this robustly. These results suggest that extremism is not so much cumulative as it is *reactive*. One form of extremism (in the form of Islamophobia) responds quickly and strongly to another extremism (i.e. Islamist terrorism) but its impact is only brief and dissipates quickly. This dataset shows little evidence that Islamophobic extremism is genuinely *growing* over time, that it is becoming more frequent or qualitatively stronger. Further research is required, but I propose that these findings can be used to defend a clarification, or sub-theoretical insight, of cumulative extremism

theory: namely, the potential for *reactive* extremisms in society. This has important implications for developing policy responses, such as providing support to victims of Islamophobia, as discussed in the final chapter.

The second contribution I make is analysing *who* tweets Islamophobically during terrorist attacks. My analysis shows that during Islamist terrorist attacks the number of one-off Islamophobic tweeters is significantly higher than other periods. For all four days of terrorist attacks combined, there are 169 one-off users compared with just an average of 14 during other 4-day combinations. This suggests there is a small cadre of users who are usually not Islamophobic but are then 'activated' to engage in Islamophobia during periods of Islamist terrorist attacks. However, I also show that this is accounted for by the larger volume of Islamophobic tweets sent during this period, as demonstrated by the analysis of models 8 and 9. Whilst terrorist attacks do attract new users to engage in one-off acts of Islamophobia, this is not above and beyond other periods when a high volume of Islamophobic tweets are sent – which suggests that the dynamics of one-off Islamophobes' behaviour are not different during terrorist attacks compared with other periods. I also show that the distribution of Islamophobic tweets per user during Islamist terrorist attacks is more unequal than in other periods (the Gini coefficient is higher). This means that the bulk of the increase in Islamophobic tweeting is due to a small number of hyper active Islamophobes who become *even more* active– rather than many low volume Islamophobic tweeters whose Islamophobia increases slightly.

This also has implications for the theory of cumulative extremism. The existence of one extremist event (the terror attack) does not have a special effect on the number of one-off Islamophobes, above the general increase in the number of Islamophobic tweets. This means that the number of individuals at risk of engaging in extremism does not increase an additional amount when another form of extremism emerges. Or, in other words,

whilst the size of the potential constituency of Islamophobic extremist actors increases following a terrorist attack, the dynamics which govern this do not change compared to other periods of time. This evidence indicates that periods of cumulative extremism (or 'reactive extremism', as I label it above) are not marked by specially greater numbers of people engaging in extremism, which further puts into question the theory.

Three factors likely explain the differences between my results and previous studies on the impact of Islamist terrorist attacks (discussed above, including (Borell, 2015; King & Sutton, 2013)): (i) social media and offline spaces have different dynamics, possibly due to the provision of social information on social media and (ii) different forms of hateful behaviour, from legal hate speech to illegal hate crime, may have different dynamics, and (iii) I have adopted a methodologically individual approach. I have tracked changes in the behavioural patterns of specific users over time – rather than conducting aggregate analyses, as in most previous research. This means that the inferences made are more robust. For instance, this study demonstrates that individuals engage in more Islamophobia rather than showing that individuals redirect their Islamophobia, which is a considerable risk with studies which use hashtags – individuals might have been just as Islamophobic prior to an attack but just have used a different hashtag or none at all.

RQ 5 | Islamist terrorist attacks and UK political parties

The results show that despite differences in terms of *volume*, the *dynamics* of behaviour around Islamist terrorist attacks are consistent across all parties. This is unexpected and highlights the far-ranging impact of Islamist terrorist attacks in UK politics, providing a simple answer to RQ 5: Islamist terrorist attacks affect the prevalence of Islamophobic hate speech very similarly across followers of different political parties on Twitter. In addition, during terrorist attacks the number of one-off Islamophobes increases across fairly consistently all parties. Note that this analysis is based solely on the binary

classifier. It is plausible that the dynamics of weak and strong Islamophobia vary, whereby – potentially – followers of mainstream parties engage in proportionally *more* strong Islamophobia during terrorist attacks. Crucially, this finding also enables me to broaden the theory of cumulative extremism. It shows that the symbiotic relationship between different extremisms not only operates at the level of groups and communities but also on individuals with different political affiliations. This highlights the complexity of modern politics and the need for a holistic approach which recognises the widespread and uneven nature of contemporary Islamophobia.

The fact that cumulative extremism operates across individuals from all parties raises a further question: *why do Islamist terrorist attacks drive an increase in Islamophobia*? Burnap and Williams suggest that it is due in part to (i) the symbolic impact of the terrorist attacks, (ii) the role of the media, as hateful tweeters 'may be fuelled by coverage in the and partly due to the amplifying impact of Twitter itself press' (Williams & Burnap, 2016) and (iii) social effects whereby users respond to normative and informational pressure to be Islamophobic, which can be either observed in content online or in the offline world. The initial results here suggest that the media does not play a considerable role and that individuals are largely driven by the symbolism of the attack itself – the peak of Islamophobia happens quickly, which means there is little time for social effects to influence behaviour. It may also be that followers of different parties are responding to different drivers; the BNP might be driven more by the symbolic impact of the attack itself whilst followers of UKIP might be driven more by exposure to social effects. Untangling these three competing hypotheses requires further investigation in future work.

The three RQs responded to in this Chapter are all directly addressed through the analysis undertaken and can be summarised as: (1) differences exist between parties in terms of

the strength and prevalence of Islamophobic tweeting, (ii) Islamist terrorist attacks drive an increase in Islamophobia but (iii) unexpectedly, they affect followers of different parties similarly. These results have considerable implications for the prevention of, and designing interventions against, Islamophobia – which is discussed in more detail in the next chapter (Chapter 8).

7.5.2 | Limitations

There are several limitations of the current work. First, the measurement process hinges on three decisions: (1) I use the binary classifier, studying Islamophobia in general, rather than separating weak and strong, (2) I measure time chronologically in seconds, specifically focusing on a 10,000 second period. As discussed in Chapter 6, this could bias the results towards periods with different *volumes* of tweets. And (3) I measure the *count* rather than the *probability* of Islamophobia. I have sought to minimize the risk of this biasing the results by conducting in-depth validity studies (reported in Appendix 7.2) to establish how these decisions impact the results. Nonetheless, it is plausible that alternative measurement decisions could lead to different dynamics being identified, thereby changing our understanding of how Islamophobia manifests. In particular, time could be measured by measuring the time periods by the *volume* of tweets which are sent rather than chronological time.

Second, there are considerable variations in the total number of Islamophobic tweets sent after each Islamist terrorist attack. This makes it difficult to interpret the y-intercept for the models, even with the inclusion of variables for the number killed and the number injured. As such, it is more useful to focus on the rate of change (given by the slope coefficients) than the absolute number of tweets.

Third, the data cleaning process removed a large number of users and tweets from the dataset. This is necessary to account for possible sources of biases, such as bots and non-English tweets. However, the removal of hyper-active users and also inclusion of 'Undetermined' language tweets could have potentially impacted the results.

Fourth, there are relatively few data points in the present work as four Islamist terrorist attacks occurred during the period. More terrorist attacks need to be studied to verify the results and assess how generalisable they are. This would require a considerably longer period of constant data collection give how rare and unanticipated terrorist attacks are.

Fifth, models 6 and 7 include variables for the number killed and number injured during attacks – but potentially this does not fully account for the considerable discrepancy between attacks in which at least one person dies and attacks in which no-one dies (specifically, Parsons Green). This could be accounted for by also including a term for *whether* there has been at least one fatality in each attack. Again, this requires a larger dataset to investigate fully.

Sixth, 15,253 users, and 11,143,987 tweets, are studied in this Chapter. This is a large quantity of data by social science standards but relatively small compared with the number of followers of each party and the total volume of data on Twitter. Due to practical constraints of data collection, I could not collect full censuses of the followers of each party. As such, caution should be taken when interpreting the findings.

7.5.3 | Extensions

The present work points to several extensions, each of which could deepen, verify and generalise the findings in future work. First, is that the time *between* terrorist attacks has not been studied. This could have one of two plausible impacts; either (i) Islamophobic apathy, whereby terrorist attacks in close succession drive less Islamophobia or (ii)

Islamophobic escalation, whereby there is an ongoing effect of successive terrorist attacks. Visual inspection of Figure 20 suggest that the second options is more likely. This could be modelled in future work by introducing a term for the number of days since the last attack. However, one constraint in studying this is that few Islamist terrorist attacks occur throughout the year.

Second, is that terrorist attacks committed by the far right (such as the Finsbury Park mosque attack on 21 June 2017) and from outside of the UK have not been modelled – although initial results suggest that these attacks also have considerable impact on the prevalence of Islamophobia. Different varieties of terrorist attack could be modelled by introducing a categorical variable for whether the attacks take place in the UK or elsewhere and by introducing a distance metric for how far the attack is from the UK (measuring either cultural, political or geographic closeness). Studying terrorist attacks from outside of the UK also points to the third extension; explicit and in-depth analysis of media coverage of Islamist terrorism. This is critical for understanding non-UK terrorist attacks as the level of media coverage differs substantially and as such the public's awareness of the attacks is most likely far lower. To fully understand the role of the media, given the short time period of Islamophobic escalation following terrorist attacks, a far more granular dataset (with hourly news coverage) is required. The initial results show that the media most likely drives Islamophobia – but this has not been robustly demonstrated yet. Furthermore, it is likely that certain types of articles from certain types of news sources (such as tabloids compared with broadsheet newspapers) have differing impact on the level of Islamophobia, which could be explored further.

Chapter 8 | Discussion and Conclusion

The aim of this thesis is:

> *To understand the nature and dynamics of Islamophobia amongst followers of UK political parties on Twitter*

In the literature review, I identified five research questions (RQ) and an additional research goal (RG). Each of the RQs and the RG have been discussed in the previous four chapters (Chapters 4, 5, 6 and 7). In summary, the outcomes for each RQ and the RG are:

> RQ 1: *What is the conceptual basis of Islamophobia?* I argue that the conceptual basis of Islamophobia is negativity and generality. These constitute orthogonal axes of Islamophobia and can be used to distinguish weak from strong varieties of Islamophobic hate speech.
>
> RQ 2: *To what extent does Islamophobic hate speech vary across followers of UK far right parties on Twitter?* The Islamophobic behaviour of followers of the BNP varies considerably. Users can be separated into different trajectories based on the temporal dynamics of their Islamophobia, including Extreme and Never Islamophobes.
>
> RQ 3: *To what extent does the prevalence and strength of Islamophobic hate speech vary across followers of different UK political parties on Twitter?* Both the prevalence and strength of Islamophobia vary across followers of different political parties. Followers of the BNP send the most and the strongest Islamophobic tweets. Followers of the Conservatives and Labour, both mainstream parties, send fewer Islamophobic tweets, and proportionally far fewer

of them are strong Islamophobic. UKIP constitutes a halfway house, which sits in between the Islamophobia of extreme and mainstream parties. Using these findings, I argue that there is an Islamophobia gap between the positions of mainstream parties and their supporters. I also argue that Islamophobia constitutes a twin threat in UK politics, comprising both the mainstream and far right.

RQ 4: *To what extent do Islamist terrorist attacks drive increases in Islamophobic hate speech amongst followers of UK political parties on Twitter?* Islamist terrorist attacks are a key driver of Islamophobic hate speech. Around Islamist terrorist attacks, Islamophobia follows a discernible pattern of escalation, peak, de-escalation and return to baseline.

RQ 5: *Do Islamist terrorist attacks have the same effect on the prevalence of Islamophobic hate speech across followers of different political parties on Twitter?* Islamist terrorist attacks affect followers of all parties. Whilst the overall prevalence and magnitude of Islamophobia varies considerably across parties (see RQ 3), the same pattern of Islamophobic escalation and de-escalation can be observed. I argue that this can be used to extend Eatwell's theory of cumulative extremism to beyond groups to also include extremist individuals within non-extremist groups and parties. I also argue that cumulative extremism could be re-framed as *reactive* extremism, as the overall level of extremism (in the form of hate speech) appears to not increase over time.

RG: *To create a machine learning classifier for Islamophobic hate speech which is closely informed by theoretical work on the concept of Islamophobia*. The machine learning classifier developed in Chapter 5 is closely informed by the conceptual work in Chapter 4. It distinguishes between weak and strong varieties

of Islamophobia, and has high accuracy given the complexity of the task. As such, it fulfils this additional research goal.

In this chapter I discuss and synthesize the findings of this thesis and consider future work and extensions. The reported results each constitute a noteworthy advance on previous scholarship in this area, and are relevant for policymakers, government and activists.

In the first section, I discuss the nature of Islamophobia on Twitter within UK politics I consider three sources of heterogeneity within Islamophobic tweeting: (i) the different ways in which Islamophobia manifests, including both weak and strong varieties, (ii) user dynamics, including the unequal distribution of Islamophobic tweets per user and (iii) temporal dynamics, specifically the role of Islamist terror attacks. At the end of the first section, I draw on work in environmental studies and complexity theory to outline useful analogies for Islamophobic tweeting: the *wind system* and *hurricane*. I suggest that characterising Islamophobic tweeting in this way effectively captures and sheds light the complex, dynamic, and devastating nature of the phenomenon. The section concludes with reflections on how this analogy can be applied in other contexts.

In the second section, I consider the dynamics of party followership in relation to Islamophobia and the implications of the present work for understanding UK party politics more broadly. In the third section, I critically reflect on the role of social media and its relationship with Islamophobia. In the fourth section, I discuss the policy implications of this thesis. Here, I focus on five areas: (i) defining Islamophobia, (ii) monitoring and predicting Islamophobia, (iii) providing support to victims, (iv) countering Islamophobia and (v) processes of radicalization. In the final section, I examine the thesis' limitations. Overall, I argue that the overarching research aim has been realised.

8.1 | Islamophobic hate speech on Twitter is highly heterogeneous

The empirical work undertaken in this thesis points to several key insights about the nature of Islamophobia amongst followers of UK political parties on Twitter. These insights are relevant for understanding Islamophobia more broadly: (1) on social media and (2) in UK society and politics more broadly. The results show there is no such thing as a typical Islamophobic Twitter follower. Instead, Islamophobia manifests unevenly. In particular, there are three sources of heterogeneity which are important for understanding Islamophobic hate speech: (i) how it manifests (ii) how users engage in different strengths and volumes of Islamophobia, and (iii) the strong temporal aspect. I then propose these results can be integrated into a unified meta-theoretical argument: the *wind system* of Islamophobia.

8.1.1 | Islamophobia manifests in varied ways

The key finding from Chapter 4 (in which I engage in qualitative and conceptual analysis of Islamophobic tweets), is that, even within just speech, there is a huge variety of ways in which Islamophobia can be expressed. Based on the empirical analysis of Islamophobic tweets, I propose a definition of Islamophobia which focuses on two orthogonal conceptual axes: negativity and generality. In principle, the philosophical and qualitative work undertaken here can be applied elsewhere. However, the extent to which this conceptual argument can be applied in other settings and to other types of Islamophobia is constrained by the dataset on which it is based, which has two key characteristics. First, it consists of social media posts (tweets). Second, the tweets are sent by followers of far right accounts. Nonetheless, it can be applied to other social media posts (such as Facebook comments and Instagram photo descriptions) and tweets sent by other users, (such as non-political users, e.g. celebrities). The framework can also be

generalised further to non-verbal forms of Islamophobia. Consider, for instance, physical assault. This is a violent act, and as such can be considered an affective manifestation of deep *negativity* against the targeted victim. Insofar as it is targeted against a Muslim because of their identity, it constitutes a *general* act. This line of reasoning requires further conceptual analysis, as well as engagement with victims of different types of Islamophobia. Nonetheless, the key point here is that (1) even just in terms of speech, Islamophobia manifests in different ways – and (2) nonetheless, in principle, the framework developed in this thesis can be used to analyse it systematically.

In Chapter 4, I used the two conceptual axes of Islamophobia (negativity and generality) to identify different *strengths* of Islamophobic hate speech: strong and weak. This marks an advance on previous empirical research, which typically adopts a binary perspective on Islamophobia. The empirical results from Chapters 6 and 7 show that the prevalence of weak and strong Islamophobia varies considerably; in general, strong Islamophobia is far less prevalent than weak Islamophobia. However, there are conflicting dynamics in how Islamophobia manifests. Chapter 6 shows that, in terms of users, the weak/strong distinction is critical to capturing different patterns of Islamophobic behaviour over time, in particular the escalating, de-escalating and casual trajectories. Without this distinction, these user trajectories could not be identified. In contrast, Chapter 7 shows that, in terms of the temporality of Islamophobic tweeting, the weak/strong distinction is less important. Weak and strong Islamophobic tweets are highly correlated over time, and as such in this chapter I collapsed the two classes together. Overall, the results in Chapters 6 and 7 strengthen the conceptual and descriptive analysis in Chapters 4 and 5 by showing that Islamophobia manifests in many different ways.

8.1.2 | A small number of users are responsible for most Islamophobia

The findings from Chapters 6 and 7 both provide evidence of substantial user level variations in Islamophobia. User dynamics are important to study as they provide insight into who is responsible for Islamophobia, and as such who should be targeted and supported by interventions.

Chapter 6 shows that users have very different behavioural dynamics, with some frequently sending Islamophobic tweets (the 'Extreme' Islamophobes), others not sending any Islamophobic tweets (the 'Never' Islamophobes) and a large number sitting in between (the 'Causal' Islamophobes, and those on the escalation and de-escalation trajectories). The users in each of these trajectories engage in distinct behavioural patterns and as such should be analysed and studied separately. The findings in Chapter 7 support the results of Chapter 6 and demonstrate considerable variation in how many Islamophobic tweets each user sends. The Gini coefficient for the number of Islamophobic tweets per user for the four parties ranges from 0.831 to 0.883 (Chapter 7, Table 4) – which shows considerable inequality, demonstrating that a small number of users are responsible for the vast majority of Islamophobic tweets.

In the future, the typology of users I developed in Chapter 6 for followers of the far right BNP could be extended to the followers of other mainstream political parties. I would anticipate that the prevalence of each trajectory is very different compared with the BNP. For instance, I anticipate that for mainstream parties, the number of Extreme Islamophobes is far lower and the number of Never Islamophobes is far higher. There could also be interesting cross-party shifts in behaviour. For instance, users who are part of the far right but never engage in Islamophobia might start following mainstream parties, and vice versa – followers of mainstream parties who engage in Islamophobia

could start following far right parties. Studying this would give greater insight into the link between Islamophobic tweeting and party followership. However, it would require a more detailed dataset about party followership patterns and, to ensure full coverage of the UK political spectrum, would necessitate the inclusion of more parties, such as the Liberal Democrats and the Greens.

The relationship between Islamist terrorist attacks and qualitatively different types of users is partially addressed in section 7.4 of Chapter 7, where I examine the inequality of Islamophobic tweeting during terrorist attacks, as well as the number of one-off Islamophobes during attacks. During terrorist attacks, the number of one-off Islamophobes increases, but only in line with the increase in the volume of Islamophobic tweets sent (when more tweets are sent, there tends to be more one-off Islamophobes). There is not an additional proportional increase in the number of one-off Islamophobes. However, I advise caution in interpreting these results, not the least as this requires further investigation as infrequent 'everyday' Islamophobes who do not recognise, or problematize, their own prejudices – and who could be at risk of radicalising into perpetual Islamophobes in the future – ought to be a key concern for society. All too often, only extreme and perpetual forms of Islamophobia are recognised as harmful and taken seriously. Overall, the results show there is no such thing as a typical Islamophobe, whether that is amongst followers of the far right or followers of mainstream parties.

8.1.3 | Islamophobia varies considerably over time

Both Chapters 6 and 7 show the temporal dynamics of Islamophobic tweeting – that is, how Islamophobic behaviour changes over time. In future work, the temporal analyses in Chapters 6 and 7 could be combined. In particular, the analysis of terrorist attacks could be combined with the analysis of user trajectories. For instance, in Chapter 6 I identify

one escalating trajectory and *two* de-escalating trajectories (one minor, the other major). This is potentially due to the fact that the three Islamist terrorist events occurred at the start of the period; some of the 'de-escalation' observed may be because users are not 'activated' in the later period as no attacks occur. These dynamics could be analysed in future work through more complex modelling, either by extending the latent Markov chain model I used, or by using an alternative method, such as growth curve mixture modelling. Potentially, there may be far more casual Islamophobes than the results in Chapter 6 suggest (note that this is already the most prevalent trajectory, accounting for 31.66% of users). A further area of investigation is the small number of users who *escalate*. They are quite unusual in that they were not 'activated' by the Islamist terrorist attacks at the start of the period but then engaged in considerable Islamophobic tweeting at the end. The behaviour of these users requires in-depth and detailed investigation to better understand their escalation pathways.

A key point illustrated by studying the temporal dynamics of Islamophobia is that studies of radicalization and users' behaviour must not operate in a vacuum: Islamophobes, across the political spectrum, respond to the world around them. Even this study of a single online platform using observational digital trace data can identify that users' Islamophobia is driven considerably by the occurrence of external events. The occurrence of such events – which includes not just terrorist attacks but also other political events of significance, such as elections – must be considered when modelling user trajectories and pathways towards extremism. This issue is due to receive significant attention in 2019 and beyond, for example, in research on the impact of political events on hate speech based at Cardiff University (Cardiff University, 2018).

8.1.4 | Islamophobia is not a *wall* but a *wind system*

In the literature review, I discussed Awan's characterisation of certain parts of social media, in particular spaces dominated by the far right, as 'walls of hate' (Awan, 2016). This position typifies the predominant approach in characterisations of the far right, both online and offline, as perpetually and committedly Islamophobic (Biggs & Knauss, 2012; Goodwin, 2013b; Lee, 2016). The substantive findings of this thesis make an important challenge against Awan's 'wall' analogy. The results show that there are, indeed, walls of hate – but the walls are not everywhere. Even amongst followers of the BNP, 1,843 out of 6,406 users studied in Chapter 6 never send an Islamophobic tweet and 2,028 of them (31.66%) engage in Islamophobia casually. Most of these walls are also impermanent – they are suddenly, and explosively, thrown up when an Islamist terrorist attack occurs but then quickly disintegrate. This points to the inadequacy of the 'wall' (a permanent and static structure) as an appropriate analogy for online Islamophobia. This clarification does not in any way diminish the severity and harm of Islamophobic tweeting on social media. Rather, it shows that Islamophobia is a complex system that manifests dynamically.

I propose that to effectively theorize Islamophobia, a new term is needed which, rather than implying stasis, suggests dynamism, unpredictability and flux. This is not only an exercise in language but also understanding; as Tierney et al. put it, 'metaphors matter'. The choice of words to describe a phenomenon is entangled in the discursive construction of that phenomenon, and as such how it is understood, interpreted and responded to (Tierney, Bevc, & Kuligowski, 2006). Drawing upon terminology from extreme event analysis in environmental studies, I argue that Islamophobia on social media amongst UK political parties should be viewed as analogous with a *wind system*, and peaks of Islamophobic behaviour as *hurricanes* (Harris et al., 2018; Herring, Hoerling, Kossin,

Peterson, & Stott, 2015; Meehl et al., 2000). This analogy can also be extended to understand other forms of hate, such as misogyny and homophobia.

Extreme events in environmental research can be defined as 'hazards or events that generate impacts on our social, ecological, and/or technical systems' (McPhillips et al., 2018, p. 441). This definition can be applied in the context of political behaviour on social media: it succinctly captures after an Islamist terror attack. Furthermore, Harris et al. state that, 'Climatological extreme events are — by definition — rare, low-frequency, intense events' (Harris et al., 2018, p. 579). They elaborate how extreme events are worth studying because 'ecosystems are vulnerable to state change' which, in turn, can 'caus[e] complex and catastrophic responses' (Harris et al., 2018, p. 585). This description is a well-suited analogy for understanding the impact of Islamist terror attacks which, drawing together the results of Chapters 6 and 7, most likely induce a state change whereby users transition from a none-to-weak or a weak-to-strong Islamophobic tweeting state. Online hate comprises a complex system with potential for chaotic developments, and unpredictable actions and events.

I contend that peaks in Islamophobia can be viewed as a type of *social* extreme event – specifically, as a hurricane. A hurricane is a large swirling storm with very fast winds (by definition, winds in a hurricane travel faster than 74mph). Hurricanes form over warm seas and then move inland where, according to NASA, 'they push a *wall* of ocean water ashore' – which can have hugely devastating consequences for the landmasses targeted (NASA, 2018). This process captures the short sharp escalation in the volume of Islamophobia during an Islamist terrorist attack which, as I argued before, can be characterised as a temporary wall of hate. In Chapter 7, I identified that after the peak of Islamophobia – as with the peak speed of the hurricane – there is an extended period of de-escalation during which the level of Islamophobia (or, equally, wind speed) reduces

to the point where the impact of the hurricane is no longer discernible from everyday fluctuations (Fitzpatrick, 2005). Hurricanes, in both natural and societal contexts, follow broadly similar dynamics of escalation and de-escalation.

Characterising Islamophobia as a wind system works on several levels. First, there is always a constant but relatively low level of Islamophobia (the 'baseline' level discussed in Chapter 7) – this echoes how there is always a low level of wind turbulence. Second, in parallel with the huge damage caused by a hurricane (Neil Adger, Hughes, Folke, Carpenter, & Rockström, 2005), the impact of large peaks in Islamophobia is devastating on targeted groups (Tell Mama, 2017). Third, the victims of a hurricane are in no way responsible for the damage it causes, they are simply caught in the storm; as an analogy with Islamophobia, the hurricane helps articulate the dynamics of not only destruction but also agency. Victims of Islamophobia are not responsible for, or causes of, Islamophobia. Finally, at present, there is a lack of understanding into the nature and causes of hurricanes. Donnelly and Woodruff argue that, 'the processes that control the formation, intensity and track of hurricanes are poorly understood' (Donnelly & Woodruff, 2007, p. 465). The same can be said of Islamophobia on social media, where many of the properties and aspects of Islamophobia are not yet well understood and the field, compared with other areas of political science, is still nascent. In this sense, Islamophobic hate speech on social media can be understood as a complex system with many different imbricated parts and emergent properties, which makes understanding dynamic processes contained within it very difficult.

One potential limitation of this analogy is that it implies a lack of agency on the part of the Islamophobic tweet*er* – that is, the individual who sends the Islamophobic tweets. In popular culture, hurricanes are viewed as unstoppable natural events. However, this is not reflected in the latest scientific research, which emphasizes the role of anthropogenic

change in driving extreme environment events (Herring et al., 2015; Meehl et al., 2000; Webster, Holland, Curry, & Chang, 2005). As Harris et al. note, 'scientific attribution of individual extreme weather events to anthropogenic climate change is increasing' (Harris et al., 2018, p. 579). This is an important qualification; it suggests that both the overall prevalence of extreme events is attributable to human behaviour (that is, the *aggregate* number can be explained) and also *individual* events can be traced back to human behaviour. As such, characterising peaks in Islamophobia as a hurricane is not to claim that they are uncontrollable or inscrutable but, rather, the direct results of human behaviour and human decisions. Indeed, the power of this analogy is that it helps highlight how many different factors in a complex system lead to Islamophobic hurricanes not that there are no causes. Investigating these causes should be a concern of future research, including the attitudes of individuals, the information streams they have access to, the news media, the socio-technical affordances of the platforms, various forms of normative pressure and also their individual emotional response to particular events (e.g. Islamist terrorist attacks).

In the future, the wind system and hurricane analogy could be developed further by drawing on the large body of research into how hurricanes should be responded to and managed. The Saffir-Simpson hurricane wind scale is used to characterise the strength and magnitude of hurricanes (Kantha, 2006). An equivalent scale could be useful for providing policymakers and communities to capture the intensity of an Islamophobic hurricane – though there is a risk that such a scale might be reductive and as such would need to be implemented carefully and through dialogue with Muslim communities. In environmental research on extreme events Herring et al. argue that researchers should not just monitor 'event magnitude and likelihood' but should also evaluate, 'societal

resilience, vulnerability, and preparedness' (Herring et al., 2015, p. 7). We need to know not only when hurricanes will hit but also how prepared we are for them.

A similar framework to those used for nature hurricanes could be used to also capture the resilience, vulnerability and preparedness of social media platforms, Muslim communities and wider society to withstanding the impact of Islamophobic hurricanes. This would help guide decisions regarding the allocation of Government resources. Work in predicting extreme weather events could be used to meaningfully assist with predicting Islamophobic hurricanes, such as the work of Anastasiades and McSharry on the utility of predicting *distributions* of values rather than single values (Anastasiades & Mcsharry, 2014)

The idea that Islamophobic hate speech can be seen as a wind system can be generalised to other social media contexts, platforms and actors, thereby taking the analogy beyond the confines of just followers of UK political parties on Twitter. The overall prevalence of Islamophobia is likely to be similar to that of the followers of the Conservatives and Labour; a low baseline but nonetheless (i) very large spikes at certain points in time and (ii) with certain users sending a large volume of Islamophobic tweets. This, therefore, is a line of reasoning which is relevant more broadly for the study of Islamophobia and, potentially, also other forms of hate speech. In principle, the dynamics identified in this thesis may be very similar across different types of hate – even though it is likely that Islamist terrorist attacks are only specific to Islamophobia, other hates, such as misogyny, may be driven by equivalent news and political events.

8.2 | Islamophobia comprises a twin threat in UK politics

Chapter 7 highlights three main points: (1) followers of far right parties engage in the most Islamophobia, (2) nonetheless, there is still considerable Islamophobia amongst followers of mainstream parties and (3) UKIP constitutes a *halfway house*, situated in between the mainstream and the far right. For these reasons, I contend that there is a *twin threat* of Islamophobia in UK politics, whereby both the mainstream and the far right send Islamophobic tweets. Crucially, this does not imply that the threat posed by the two sides is the same ('twin' does not denote cloned, identical nature and behaviour, but rather fundamental similarity). The far right is more likely to engage in direct, overt and highly aggressive Islamophobia whilst the mainstream engages in a lower volume of weaker Islamophobia. The threat posed by the mainstream is often ignored or under-appreciated – even though it may well be more insidious precisely because it is not problematized in contemporary political discourse.

This concept of the *twin threat* in UK party politics also has utility when situated in the complex/environmental systems analytic I outlined in the previous section. Islamophobia within UK politics on Twitter is analogous with a single, interconnected wind system; the same swirling mass of weather, with some pockets in near stillness and other pockets in raging storm. Islamophobia differs across UK politics; strong gales, lighter winds and breezes can metaphorically convey the force of tweeting from the BNP to UKIP to the Conservatives and Labour. But there is always the potential for change, unevenness and disruption. For instance, the tempestuous force of tweets from the most Islamophobic followers of Labour is far stronger than some followers of the BNP. Most importantly, the same tailwinds that drive Islamophobia within the far right also drive Islamophobia

elsewhere; followers of all the political parties studied in this thesis are affected by Islamist terror attacks.

The twin threat of Islamophobia points to a dissonance in how political parties seek to present themselves and the practices of their supporters. In the literature review, I discussed the 'policy gap' between how parties officially view immigrants and create policies addressing migration, and the views of their supporters and the wider public. Given that all mainstream parties in the UK state their opposition to prejudice and support of liberal values, the results here indicate a similar *Islamophobia gap* between them and their followers. Even amongst followers of Labour, there is a small number of users who engage in strong Islamophobic behaviour. Investigating this result further will require detailed, systematic analysis of both the discursive practices of the parties (i.e. through their leaders' speeches, manifestos, party websites, and social media posts) and their followers in other contexts, both online and offline. These results could also be used to interrogate the extent to which political parties really are anti-prejudicial and to what extent they, whilst formally disavowing prejudice and bigotry, nonetheless engage in subtle and indirect forms (Leruth & Taylor-Gooby, 2018).

The wind system of Islamophobia, comprising a twin threat in UK politics, has implications for how we understand political parties in general, in particular the far right. Historically, the far right has primarily been associated with fascism and post-fascism (Copsey, 1994, 2007; Goodwin, 2011; Richardson & Wodak, 2008). However, since the 20th century, it has undergone several changes – only some niche far right subcultures such as skinheads, white supremacists and neo-Nazis are explicitly anti-democratic, and most 'comply by the minimal procedural rules of parliamentary democracy' (Castelli Gattinara & Pirro, 2018, p. 4). In contrast, *prejudice* is a hegemonic element within, if

not the defining feature of, the ideology of the contemporary far right (Mudde, 2002, 2007b, 2017). Indeed, the explicit articulation of prejudices (whether anti-Islamic, anti-Roma or anti-Immigrant) is what gives the far right 'issue space' within political discourse (Cole, 2005). Articulating prejudice makes far right parties a distinctive option, ensuring they appeal to a certain part of the electorate (Golder, 2016; Lucassen & Lubbers, 2012; Veugelers & Magnan, 2005).

Issues relating to prejudice also play a pivotal role in the ideology and discourse of mainstream parties, who often focus on issues such as social integration and multiculturalism, and non-prejudicial nationalism (Bulmer and Solomos, 2015). De Cleen and Stavrakakis argue that mainstream political discourse is dominated by two primary issues: (i) political representation (such as populism) in the form of debates about democracy and power and (ii) to prejudice and identity, through discourses relating to nationalism, nativism and belonging (Cleen & Stavrakakis, 2017). These issues operate within the far right in different ways to how they operate in the mainstream. For instance, empirical work has shown that supporters of Donald Trump's US presidential campaign in 2016, and voters for the BNP in the UK during the 2010s, can be distinguished from voters of other parties by their prejudicial attitudes and, in particular, concerns about intergroup conflict (Ford & Goodwin, 2010; Oliver & Rahn, 2016).

The ideological centrality of prejudice and identity within both far right and mainstream politics has certain implications. Primarily, it means that the characterisation of UKIP vis-a-vis Islamophobia as a *halfway house* between the mainstream and the extreme can be used to re-evaluate received understandings of the party as a whole. It provides evidence to see UKIP as politically halfway – and with the concerning potential to fully join the far right. If UKIP is halfway between the mainstream and the extreme and

capable of receiving ~4 million votes in the 2015 general election (and still 600,000 at the 2017 election), it raises questions about the fragmented and divisive nature of UK politics, and the potential closeness of the mainstream and far right. That said, evidence of prejudicial behaviour, such as that provided in Chapter 6, should not be the sole criterion used to characterise parties, but rather used alongside other dimensions of party politics, including descriptive and causal factors (such as the socio-demographics and geographic location of their voters) to compare and situate them. Nonetheless, it is a useful lens which has thus far been considered insufficiently.

This analysis also reopens the debate about the role of the far right in UK politics, and it's supposed 'failure' (Goodwin, 2013a; Ignazi, 2003). My findings suggest that there are important behavioural affinities between the far right and the mainstream. Noticeably, as I discussed in Chapter 7, processes of *cumulative extremism* operate even amongst followers of mainstream parties, such that they become motivated to engage in extremist Islamophobic acts following Islamist terror attacks. This provides complementary evidence in support of the thesis put forward by Margetts et al. that there is considerable 'latent support' for the far right (John & Margetts, 2009; Margetts et al., 2004) although, rather than latent support, I find evidence of Islamophobic behavioural affinity.

The arguments advanced in this section have implications for how we view the political system. Different political parties are not categorically distinct and separate entities but, rather, are on a single spectrum. The far right should not be viewed as entirely Islamophobic and the mainstream should not be viewed as entirely non-Islamophobic. The reality of party followership is far more complex. Whilst the far right and the mainstream can still meaningfully be viewed as different, not least as the prevalence and strength of Islamophobia expressed by their followers differs hugely, this is a difference

in *degree* rather than in *kind*. This could be developed further in future work by examining users who follow multiple political parties and evaluating how their Islamophobia evolves over time. For instance, it is plausible that during Islamist terrorist attacks, far right parties attract more followers. Another area of future study, alongside the analysis of offline party supporters, is whether these results can be replicated in other online settings, such as on Facebook and Reddit.

8.3 | Social media

In this section I reflect critically on (i) the extent to which the results here can be generalised to other settings, including other social media platforms and offline contexts and (ii) their implications for how we characterise social media and understand its role UK in society.

8.3.1 | Generalising beyond Twitter

Many researchers in the fields of computational social science and complex systems contend that online research can provide insight into more general aspects of society. As Conte et al. put it in their 'computational manifesto', 'Information and communication technologies can greatly enhance the possibility to uncover the laws of the society' (Conte et al., 2012, p. 327). Similarly, in a study of online memory, Garcia-Gavilanes argue that 'Wikipedia traffic data reliably reflect the Internet users' behaviour in general' (Garcia-Gavilanes et al., 2017, p. 1). These arguments seem plausible – indeed, the online is a constitutive part of society. Yet the universality of mechanisms identified from online research can be overstated. Social media is highly unrepresentative of the offline world and Internet users in general. There are considerable inequalities in terms of internet access, attitudes, skills, and usage across different demographics (Friemel, 2016; van Deursen & van Dijk, 2014). Internet users are consistently younger, better educated and wealthier than the average population, and Twitter users are even more young, educated and wealthy than Internet users in general (Blank, 2017). A large number of users have been studied in this thesis. However, this does not overcome the fact that social media users are unrepresentative, despite the hopes of some researchers (Ruths & Pfeffer, 2014). As such, social media platforms cannot be used as a proxy for studying offline or online

behaviour in general, even though findings about online dynamics might be useful for explorative research and developing hypotheses.

Online users have the choice of many different platforms to engage with. Research suggests that different platforms are used for different purposes, and that users adopt different identities and behavioural patterns in different online spaces (Miller & Horst, 2012). Platform choice is often motivated by attitudinal and personality traits. For instance, evidence suggests that users of Instagram and Facebook are more narcissistic than users on Twitter (Correa, Hinsley, & Zúñiga, 2010; Davenport, Bergman, Bergman, & Fearrington, 2014; Maruf, Meshkat, Ali, & Mahmud, 2015; Phua, Venus, & Jay, 2017; Sheldon & Bryant, 2016). This means it is difficult to generalize the findings of this research to other platforms, which not only have different socio technical affordances but also attract different types of users. Because users' identities and behaviours vary across platforms, it is plausible that Twitter, which is a broadcaster platform often used for political talk, may contain more Islamophobia than other platforms such as Facebook or Reddit. In particular, Reddit has been associated with more conversational and deliberative forms of communication rather than polarised broadcasts (Sowles et al., 2018). On the other hand, niche social media platforms such as 4chan have been shown to harbour a large number of very extreme Islamophobes and other prejudicial users, which may also limit the applicability of these findings (Hine et al., 2016). The fact that Facebook has higher privacy settings, and peoples' online identities there are more closely entangled with their offline ones, may also mean that individuals are less willingly to engage in socially criticised behaviour, such as Islamophobia (Hollenbaugh & Ferris, 2014; Waterloo, Baumgartner, Peter, & Valkenburg, 2018). For these reasons, I advise caution in generalising the findings of this thesis to other social media platforms.

The present work consists of a single platform study. It has enabled the creation of deep and nuanced knowledge about the nature of hate speech on Twitter. Many researchers argue that Twitter is over-used in social scientific research because it is relatively easy to collect. Cihon and Yasseri draw attention to how Twitter-based research often 'fails to use standardized methods that permit interpretation beyond individual studies' (Cihon & Yasseri, 2016, p. 1). Accordingly, whilst the results presented here could be used to understand behavioural dynamics in other contexts, this requires further validation.

Future research could address the problem of generality by studying multiple platforms simultaneously. This will help to overcome the limitations of studying individual profiles on a single platform. This was discussed in the conclusion of Chapter 6, including the fact that users may switch from one platform to another, from one identity to an alter or from the online world to the offline – and in all cases, users could move in either direction. Fully understanding how Islamophobia operates across different platforms requires comprehensive research connecting users' separate identities, and even their intra-platform alters. However, this is technically difficult when using observational data rather than self-reported data such as surveys (Zafarani & Liu, 2013). It is also ethically problematic as individuals may not consent to having their separate online identities combined – and behaviours which they want to keep separate conjoined.

8.3.2 | The 'wild west' of social media

Social media has been described as a toxic 'wild west' by both academics (Tench & Jones, 2015) and politicians, such as US senator Mark Warner (CNet, 2018). The Competition Commissioner of the EU claimed that 2018 is the year in which technology 'has become darker and more muddy' (BBC, 2018e). Indeed, the potential of social media to enable harmful behaviour was a key starting point of this thesis, as discussed in the

Introduction. But are such descriptions warranted? Primarily, the 'wild west' label relates to the huge variety of harmful behaviours which are reported on and encouraged on social media (ranging from self-harm to trolling (Gabriel, 2013; Jakubowicz, 2017)), as well as the difficulty regulating and monitoring such spaces due to myriad technical, legal and political issues (Belli & Zingales, 2017).

The current project has only focused on one type of harmful behaviour (Islamophobic hate speech) and cannot comment directly on others, although the broad framework and findings are likely relevant. The findings suggest that there is a considerable volume of Islamophobic hate speech on Twitter. However, there are two important caveats to this. First, the users I have studied are likely to be considerably more Islamophobic than most. Second, the time period I am studying includes many terror attacks which, as the results show, drive a large increase in Islamophobia. As such, the average prevalence of Islamophobia reported in Chapters 6 and 7, provides only limited insight into Islamophobia in general. Third, I have not studied counter-speech and anti-fascist groups challenging Islamophobia. Thus, the picture given here is partial and only shows a small part of the overall story of harmful and hateful behaviour on Twitter. These actions might also provide more insight into the observed dynamics, such as the trajectories of Islamophobia.

Social media has come under increasing criticism for its potentially manipulative and divisive role in society. This is partly a response to growing recognition of its importance. Islamophobic hate speech is an undeniably harmful and divisive behaviour. However, some suggest that in monitoring and countering it, there is also a risk of creating new ethical and social problems, such as restricting freedom of expression or invading users' privacy. This is through the perception that machine learning tools to identify Islamophobic tweets are invasive. Whilst Islamophobic hate speech might be harmful

and unpleasant, it is largely legal and only in some cases does it impinge on platforms' policies. To ensure that this research does not further contribute to the erosion of trust in social media – which, fundamentally, is a tool which enables individuals to connect and share their experiences, and thus can be mobilized for many pro-social purposes – it is crucial that the findings and methodologies developed here are not used for draconian purposes or to stifle freedom of speech and association. It should also be recognised that in many online contexts it is not only an issue of free speech but also 'who speaks': censorship and constraints on freedom of expression impact who feels comfortable and welcome in online spaces. It also impacts, often unevenly, who is constrained in speaking. These issues should be carefully considered when using any form of automated technology to monitor and moderate online behaviour.

8.4 | Policy: what can we do?

This thesis contributes to governmental and wider policy work in five areas: (i) defining Islamophobia, (ii) monitoring and predicting Islamophobia, (iii) providing support to victims, (iv) countering Islamophobia, and (v) understanding processes of radicalization.

First, this thesis contributes to the UK Government's efforts to define Islamophobia. Until very recently there was no widely accepted definition, even though, as MEND puts it, and as discussed at the start of Chapter 4, this is 'timely and essential' work, which will 'help policymakers better to understand and respond to the problem of anti-Muslim prejudice' (Ingham-Barrow, 2018, p. 9). The definition developed in this work is not only useful for robust empirical science but could also help policymakers in countering Islamophobia and provide support to victims. Indeed, during the All Party Parliamentary Group on British Muslims' evidence gathering phase for their work on defining Islamophobia, I contributed preliminary findings from this PhD, which were used substantially in their final report (All Party Parliamentary Group on British Muslims, 2018).

Second, this thesis contributes to efforts to monitor Islamophobic content on social media. This is a key concern of Government and platforms, particularly in cases where the content is illegal. For instance, many of the largest social media platforms (including Twitter, Snapchat and Facebook) have signed up to the EU Commission's directive to respond to all flagged hate speech within either 7 days or, for violent extremist and terrorist content, just 24 hours. This is a huge bureaucratic and technical challenge with considerable financial implications (Reuters, 2018). As I discussed in the conclusion to Chapter 5, the multi-class and binary classifiers developed here are specific to the training data and context. Applying the classifier to content produced in other contexts would require further validation and testing. Nonetheless, in principle, the classification

methodology which I have developed could be re-implemented with training data suitable for other use cases. Furthermore, the work in Chapter 6, in which I identify an 'Extremist' trajectory of Islamophobia for followers of the BNP who are perpetually Islamophobic, could contribute to Government's work monitoring extremist users.

Policymakers and Government are interested in not only understanding and theorizing the dynamics of Islamophobia but also predicting them, thereby providing potentially real-time insight into when Islamophobia will peak and decline in volume. The difference between prediction and explanation has been much discussed in the academic literature (Shmueli, 2009; Yarkoni & Westfall, 2017). It has been described as the difference between machine learning and statistics. In machine learning the goal is to maximize performance, often through using impenetrable black box algorithms, whilst in statistics the goal is to understand and measure the impact of different variables (Boulesteix & Schmid, 2014). In many disciplines, machine learning might be able to predict individual users' behaviour rather than just aggregate behaviours (Bzdok & Meyer-lindenberg, 2018). However, the method used in Chapter 6, latent Markov modelling, is ill-suited to predicting individuals' behaviour but can be used to predict the *distribution* of users in different Islamophobic states. The methods used in Chapter 7 could be used to predict individuals' behaviour. However, the models would need to be far more advanced and complex to really capture user differences in behaviour.

The third area of policymaking is providing support to victims of Islamophobia. During Islamist terrorist attacks, there is a large but temporary increase in Islamophobia. Responding to this as quickly and effectively as possible should be a priority of Government during these periods. One strategy could be for the Government, and Muslim community groups, sending counter speech messages (Ernst et al., 2017; Sponholz, 2016)

and might hugely limit the quantity and intensity of harmful content observed by Muslims and others. These could either be prepared in advance or sent by digital activists. Alternatively, as the results indicate that the period of maximum Islamophobia is very brief (the escalation period following the attack lasts for only approximately 11 hours), it could even be possible for social media platforms to temporarily adjust the content that potentially vulnerable groups see on their timelines. The analysis in Chapter 7 shows that that just a few users are responsible for the vast majority of Islamophobia. It could be possible to flag such users as perpetual and constant Islamophobes, and enable vulnerable users to opt to not see their content.

The fourth area of policymaking is countering Islamophobia. This discussion is largely speculative, as social policy and intervention strategies require detailed empirical investigation in their own right. Indeed, there is a great risk that approaches which are inadequately tested and trialled will have unintended negative effects. As Wojcieszak found with a study of the online white supremacist forum Stormfront, 'Although the goal was to encourage neo-Nazi sympathizers to reconsider their predilections, this study suggests that such and similar actions may backfire' (Wojcieszak, 2010, p. 651). This point feeds into debates as to whether bans (whether temporary or permanent) are effective. On the one hand, a study of the impact of banning hateful subreddits in Reddit found that 'bans work' as other subreddits do 'not inherit the problem' (Chandrasekharan et al., 2017, p. 1). Some users leave Reddit whilst others modify their behaviour. On the other hand, research also shows that legal action against far right leaders for their prejudicial behaviour can actually increase support for them (Spanje & Vreese, 2015). Even if bans 'work' to remove hateful content each platform this does not necessarily ameliorate the problem for society, and there are considerable concerns that they drive

hateful and extremist individuals 'underground' where they might become even more extreme (Cowls & Brown, 2015).

Different types of users, engaging in different types of Islamophobic behaviour, may warrant different responses. For instance, causal causal Islamophobes do not engage in regular and ongoing patterns of Islamophobic behaviour. As such, these users might respond best to lighter touch interventions, which can be achieved through automated techniques such as chat bots (Munger, 2017). In contrast, 'extreme' Islamophobes, who may send illegal tweets need more heavy handed and fast moving interventions. 'Escalating' Islamophobes must also be addressed with particular concern, especially at early stages of escalation when few Islamophobic acts have been committed. Identifying them early on and deterring them could be crucial for ethically moderating online spaces.

The fifth area of policymaking which this thesis contributes to is the link between the online and the offline in terms of radicalization and extremist behaviour. Initial evidence suggests that online behaviour, in particular hate speech, is linked with offline hate. For instance, through a qualitative analysis Awan and Zempi contend that, 'there is a continuity of anti-Muslim hostility in both the virtual and the physical world' (Awan & Zempi, 2016). Muller and Schwarz find that changes in anti-refugee sentiment on the Alternative fur Deutschland Facebook page predicts violent crimes against refugees in certain German municipalities, and conclude that 'social media can act as a propagation mechanism between online hate speech and real-life incidents' (Müller & Schwarz, 2017). There is further evidence that online and offline prejudicial behaviour are linked (Gill, Corner, Thornton, & Conway, 2015; Peddell, Eyre, McManus, & Bonworth, 2016; Szmania & Fincher, 2017). Gruenewald et al. show that in the USA many far right terrorists downloaded extremist literature from the Internet, and argue that use of the Internet 'may be increasing as a tool for recruitment, sharing of tactics and attack

planning' (Gruenewald, Chermak, & Freilich, 2013, p. 80). This thesis has not directly addressed the extent to which online and offline Islamophobia are connected; this is difficult to assess due to the limited data available and the important ethical considerations which constrain analysis of individuals across the online/offline divide, as well as across different platforms. Nonetheless, the work undertaken here contributes to such efforts by identifying users which are highly Islamophobic and time periods in which Islamophobia peaks. These users will be highly appropriate focuses of research into the online/offline connection.

8.5 | Thesis limitations

There are several limitations to the work undertaken in this thesis which could constrain the robustness and validity of research contributions and as such require further consideration. Research constraints are also discussed in each of the empirical chapters; the issues discussed here relate to the thesis as a whole.

First, are ethical considerations. These have been a central priority throughout the design and implementation of the work and in the presentation of results. In all chapters, I have anonymized and minimised the potential for harm to users whose data has been collected. I have presented results in aggregate and ensured privacy by only sharing information that is directly relevant to the main arguments, and not reporting any person-identifying information. The research has kept in line with the ethical principles outlined in Chapter 2 and has closely adhered to the guidelines set out by the Departmental Research Ethics Committee of the Oxford Internet Institute.

Second, is the role of bots. In this thesis, I have removed hyper-active accounts (defined as those who tweet 40 times or more per day) from the dataset. This has minimized the chance that bots unduly bias the results through their high volume and atypical behaviours. However, since the research was first designed, the role of bots on social media has been researched more extensively and has been identified as a very concerning aspect of online political behaviour and discourse (Howard, Woolley, Calo, & Howard, 2018). As such, whilst I have sought to mitigate the impact of bots, a better strategy might have been to include bots in the research design by explicitly modelling their role and impact.

Third, is the focus on followers of political parties, as discussed in Chapter 2 and tested through the small pilot study reported in Chapter 3. Viewing followers as a constitutive

element of political parties is an important aspect of this work, which builds on prior theoretical arguments by Margetts and others into the nature of modern political parties and the importance of social media actors (Margetts, 2006; Vaccari & Valeriani, 2016). Nonetheless, more work is needed in future studies to understand the constitutive role of social media followers within political parties. The work undertaken in this thesis provides a useful starting point for this future research avenue.

Fourth, I have adopted an *individualistic* ontology to the study of Islamophobic hate speech. For instance, in Chapter 4, I justify the inclusion of hatred against both Muslims (people) and Islam (religion) within my definition of Islamophobia on the basis that even anti-Islamism inflicts harm against actual individuals; drawing on the work of Parfitt, I argue that it is a type of 'person-affecting' Islamophobia (Parfitt, 1987). Individualism is widespread within liberal Western discourses but is less common within other cultures, which place far greater emphasis on group identity and community belonging (Asad, Butler, & Mahmood, 2009; Castles & Davidson, 2000). Thus, many Muslim NGOs and charities would not necessarily view the clear-cut distinction between Muslims and Islam as intrinsically important – each is necessarily imbricated with the other. Similarly, I make a clear-cut distinction between Islamophobia (*qua* religious identity) and other facets of identity (such as race, nationality, legal immigration status and gender). Again, this can be contrasted with non-Western notions of identity, which tend to view such aspects as inherently intertwined. For instance, the APPG on British Muslims, which has liaised closely with British Muslim communities, holds that Islamophobia is 'a form of racism' (All Party Parliamentary Group on British Muslims, 2018). In future work, the different modalities and manifestations of Islamophobia could be taken into account more fully by considering the intersectional nature of Islamophobia, such as how it is linked to

the *visibility* and *performance* of Muslimness, which, in turn, is often linked to gender and class dynamics (Zempi & Chakraborti, 2015).

One way of achieving this is to engage closely with Muslim communities and the victims of Islamophobia. Indeed, a key limitation of the framework I have developed to characterize Islamophobia is that it is based on the philosophical analysis of a privileged white male researcher (i.e., me). Through engagement with Muslim communities via the All Party Parliamentary Group on British Muslim's community 'listening' process, I have learnt that Muslims emphasize some facets of Islamophobia as being the most important. It might be possible to refine the existing framework by reviewing it with people who have been targeted by Islamophobic behaviour, and exploring how it relates to other facets of identity, such as gender and class. In particular, it could be possible to make further fine-grained distinctions about Islamophobia, such as between targeted/non-targeted, violent/non-violent and swearing/non-swearing. At present, these aspects are only implicitly considered in my framework in that they each impact the degree of negativity which is expressed. Put simply, to fully understand different manifestations of Islamophobia it is crucial that the voices and stories of Muslims are put centre stage, rather than the concerns of removed, non-Muslim academic researchers. In this thesis, the focus has been on the *articulation* of Islamophobic hate speech rather than its *impact* on victims – which could be addressed in future work.

Finally, this thesis aims to mix qualitative and quantitative methods within a complementary computational social science research design (Blok & Pedersen, 2014). This mixed approach has proven fruitful in framing my use of computational methods to address pressing social research questions, and my use of conceptual social research in informing the development of the supervised machine learning classifier (see Chapters 4 and 5). Integrating qualitative and quantitative methods is a challenge for all researchers,

as is ensuring that quantitative methods are not overly reductive. The goal of producing complementary research has been productive in mitigating the risk of reductiveness. It has facilitated robust, theory-driven social science using computational methods – rather than exploratory data-driven work which is not anchored in theoretical debates. In doing so, this PhD contributes to ongoing academic discussions around the role of 'big data' computational analyses by demonstrating the benefits of a well-integrated data science research design.

8.6 | Impact

I have sought to ensure my work has impact in three ways: (i) sharing code, (ii) publishing papers and (iii) engaging with non-academics. I have developed all of the code used in the thesis. In some cases, such as with the Twitter data collection, I have built on existing scripts (such as (Hale, 2014), and in other cases I have used publicly available coding solutions from Stack Overflow. The scripts have been shared on online at `https://github.com/bvidgen`. This methodological transparency enables other researchers to reproduce my findings and will also encourage them to develop and advance my research and its impact (Mesirov, 2010; Open Science, 2015). Whilst the code used in this project can be shared, the data cannot be shared as this could risk jeopardizing the anonymity of users. Because only the code is shared and not the data, this project is near the bottom of the 'spectrum of reproducibility' for computational science outlined by Peng (Peng, 2011). Nonetheless, I have sought to maximize the amount that is shared at all times and will continue to make my work as open as possible.

The findings of this work will be published in appropriate journals:

1. Conceptual analysis of Islamophobia (Chapter 4). This will most likely be published in a qualitative or critical journal of politics, such as the *Journal of Political Ideologies, Race and Class* or *Ethnic and Racial Studies*.
2. Weak and strong Islamophobic hate speech classifier (Chapter 5). A pre-print of this article is available on Arxiv (Vidgen & Yasseri, 2018a). This will be submitted to a computer science conference or data science conference, such as the *International Conference on Natural Language Processing and Information Retrieval* (2019, 3rd edition), available at http://www.nlpir.net, *Empirical Methods in Natural Language Processing & International Joint Conference on Natural Language Processing* (EMNLP) 2019, available at https://www.emnlp-ijcnlp2019.org, or the *Association for Computational Linguistics (ACL)* 2019, available at http://www.acl2019.org/EN/call-for-papers.xhtml.

3. Trajectories of Islamophobic behaviour within the far right (Chapter 6). I aim to publish this in a social data science or computational social science journal with a track record in publishing political science research, such as: *EPJ Data Science*, *Royal Society Open Science* and *PLoS ONE*.
4. The role of Islamist terrorist events in driving Islamophobia (Chapter 7). I aim to publish this in one of the journals from point 3.
5. Differences in the strength and prevalence of Islamophobic hate speech across followers of different political parties (Chapter 7). I seek to publish this in a political science journal, such as: *Political Quarterly*, *Perspectives on Politics* and *Politics*.

Non-academic engagement is important for ensuring that rigorous academic work has a meaningful and positive impact on society. It also helps to justify the allocation of public resources by demonstrating the social value of research (Perkmann et al., 2015). I have sought to engage with non-academic audiences. First, through three blogs aimed at the general public. One on p-values and the use of statistics in social science research, published in *Policy & Internet* (Yasseri & Vidgen, 2016), another on the nature of the online far right, published in *Open Democracy* (Vidgen, 2017), and a third on the classification work in Chapter 5, published in *The Conversation* (Vidgen & Yasseri, 2018b). Second, I am in dialogue with representatives from two of the biggest Muslim charitable organisations working to counter Islamophobia, MEND and Tell MAMA. I hope to build on my existing relationship with these charities to conduct new research which centres on the experiences of victims of Islamophobia and ensures their concerns are responded to. Engaging with charities will also facilitate the sharing of reciprocal insights and feedback on research, and inform their policy proposals. Third, I have engaged with UK Government: I have co-authored an unreleased research paper on far right extremism online and also contributed to the All-Party Parliamentary Group on British Muslims' report defining Islamophobia (All Party Parliamentary Group on British Muslims, 2018).

8.7 | Conclusion

This thesis makes a timely and constructive contribution to ongoing scholarly and policymaking discussions regarding the dynamics of Islamophobic hate speech within both UK politics and on social media. Through this Conclusion, I have sought to systematize and synthesize the key findings from the thesis, which has brought together and extended scholarship from overlapping areas of research: (i) Islamophobia, (ii) party politics and (iii) social media.

I have made three main contributions. The first contribution is conceptual: I have argued that Islamophobia can be conceptualised in terms of negativity and generality. The second contribution is methodological: I have developed a supervised machine learning classifier which is closely informed by the conceptual work undertaken in this thesis. The third contribution is theoretical: I have made several substantive findings in this thesis. These relate to the *twin threat* of Islamophobia, the characterisation of UKIP as a *halfway house* in-between mainstream and far right parties, the *heterogeneity* of the far right, the potential existence of an *Islamophobia gap* between mainstream parties and their supporters, the *large but temporary impact* of Islamist terrorist attacks in driving Islamophobia across the political spectrum. I have used these findings to extend the theory of cumulative extremism to include individuals as well as groups and questioned whether academic attention should be re-focused to what I call 'reactive' extremism, given that the overall prevalence of Islamophobia appears to not increase over time.

Drawing these findings together, I have argued that Islamophobic hate speech amongst followers of UK political parties on Twitter can be characterised as a *wind system* which contains Islamophobic *hurricanes*. This analogy highlights the huge devastation that Islamophobia can cause, its chaotic nature, and the complexity of the dynamics which underpin it.

End of Volume I

Volume II

Tweeting Islamophobia: Islamophobic hate speech amongst followers of UK political parties on Twitter

Bertram Vidgen

Wolfson College, University of Oxford

Appendices

Appendix 3.1 | Data collection frequency

The steps taken to test the impact of different frequencies of data collection (weekly vs. daily) are reported. First, the method for this study is described. Second, the two collection frequencies (weekly and daily) are compared. Third, the impact of the collection frequencies is discussed in relation to how they impact the daily volume of tweets collected. Fourth, the presence of bots is discussed and, fifth, the impact of reduced tweet collection on the detection of Islamophobia is discussed (using the multi-class classifier developed in Chapter 4). Finally, sixth, the results of this study are summarised.

3.1.1 | Method

For a two-week period, from Thursday 9th August to Thursday 23rd August 2018, I collect tweets for 1,000 users on both a daily and weekly basis, in both cases using Twitter's pagination function. The users are sampled randomly from the followers of the BNP's account (@bnp) on Wednesday 8th August 2018.

3.1.2 | Comparison of data collection methods

Out of 1,000 sampled users 657 users tweet during the two-week period (the remaining 343 either do not tweet or have their accounts set to private). All users appear in both datasets. However, there is a discrepancy of 1,828 tweets (4.95%) between the two datasets; the daily collection method collects 38,739 tweets whilst the weekly collection method collects only 36,911.

For 85.7% of users both data collection methods return the same number of tweets, for a further 12.6% the discrepancy is less than 5% and for just 1.7% is the discrepancy greater than 5%. The top 10 discrepancies are shown in Table 3-1. Whilst the weekly data

collection method collects tweets for all users – and in the vast majority of cases, collects all of their tweets – for some high-volume users the weekly method collects considerably fewer tweets.

Rank position	Tweets collected by daily method	Tweets collected by weekly method	Difference (count)	Difference (percentage)
1st	817	364	453	55.4%
2nd	751	365	386	51.4%
3rd	720	405	315	43.8%
4th	433	271	162	37.4%
5th	552	407	145	26.3%
6th	462	357	105	22.7%
7th	488	379	109	22.3%
8th	5	4	1	20%
9th	10	9	1	10%
10th	83	76	7	8.4%

Table 3-1, Top ten users with the greatest discrepancy in number of tweets based on data collection method

3.1.3 | Impact of collection methods on daily volumes of tweets

For more fine-grained insight, I compare the daily volume of tweets for each collection method. I find that in 13 out of 14 days the daily collection outperforms the weekly collection. The greatest discrepancy is 11.9% – or 315 tweets. Importantly, as shown by Figure 3-1, the general trend of the volume of tweets is the same with both collection methods.

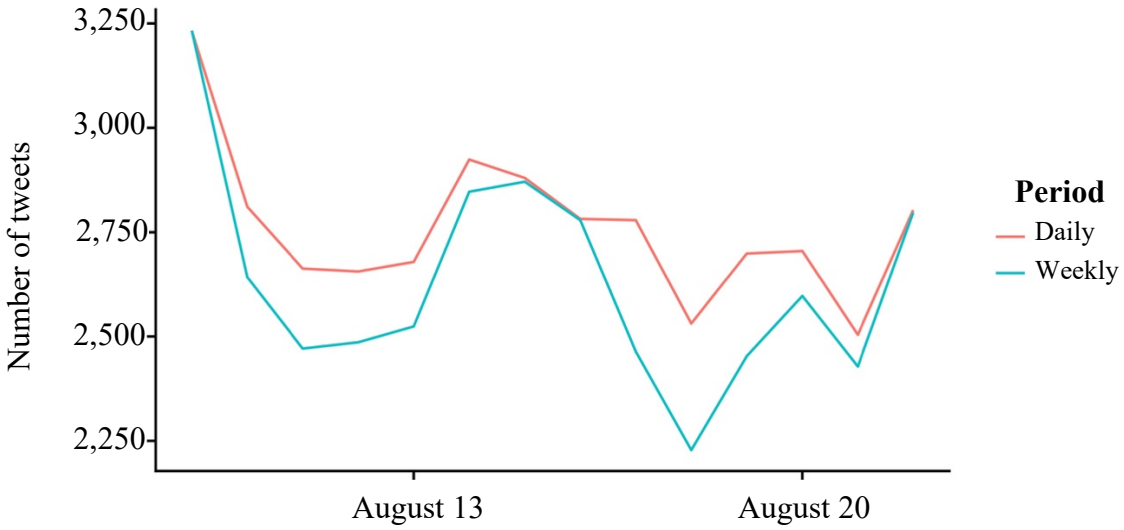


Figure 3-1, Daily volume of tweets collected via the daily and weekly collection methods

3.1.4 | Presence of bots

I check whether there is any relationship between the volume of tweets collected (and missed), and whether the accounts are bots. Accounts are assigned a probability of being a bot using the 'Botometer' application from the University of Indiana's Truthy Project (Davis, Ferrara et al. 2016). There are limitations to their bot detection methodology and the assigned probabilities should be treated with caution, particularly on a reasonably small dataset such as this – but nonetheless it has been successfully used in prior research (Monsted, Sapiezynski et al. 2017).

The average probability of a user in the sample being a bot is 0.186. For the 11 users where the discrepancy between the amount of data collected is over 5%, the average probability of being a bot is higher at 0.410. The probability of a missed tweet coming from a bot is 0.462. Overall, this suggests that the under-recording of the weekly method is more likely to affect bots than regular users. My ability to detect bots is not affected by the volume of tweets that I collect for each user as the bot classification process is

independent of the data collected. Thus, this result will not jeopardize any future analysis involving bots.

3.1.5 | Islamophobic hate speech

Using the multi-class Islamophobia classifier outlined in Chapter 5, I measure the prevalence of Islamophobic tweets over the two-week period. The distribution for both collection methods is very similar. In both cases, 'None' Islamophobic is the most prevalent, accounting for 89.6% of tweets in the Daily frequency and 88.6% in the Weekly frequency. Weak Islamophobia accounts for 8% and 8.4%, and Strong Islamophobia accounts for 2.5% and 3%. These results indicate that the different collection periods do not make a material difference to the level of Islamophobia that is recorded. The Weekly collection frequency is a very close approximation of the results from the Daily collection frequency. The results of this analysis are shown in Figure 3-2.

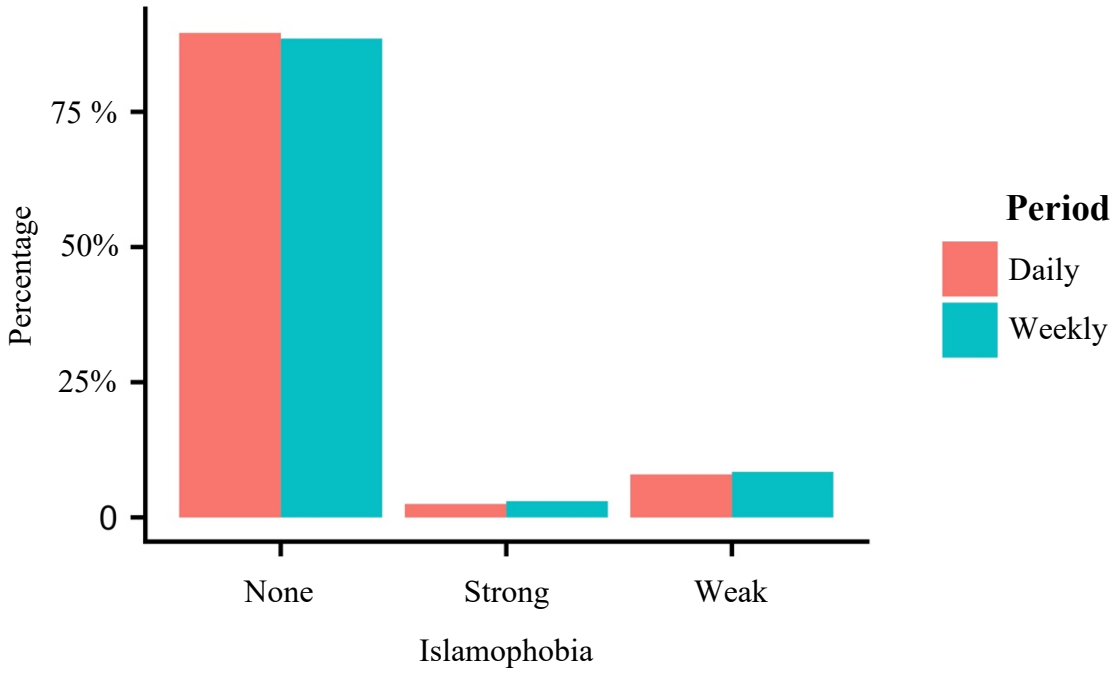


Figure 3-2, Prevalence of Islamophobia in tweets collected via the daily and weekly collection methods

I then analyse whether the capture of Islamophobic behaviour is influenced by the day on which data collection takes place. For each day I collect the number of Islamophobic tweets. On each day, the daily collection method has better coverage of Islamophobic tweets, as indicated by the prior analyses. The size of the difference ranges from 3 to 47 tweets with a mean of 16. Importantly, there does not appear to be a systematic bias whereby one specific day (such as the day before new weekly data collection is implemented) consistently under-reports the level of Islamophobia. Noticeably, compared with Figure 3-1, it appears that data coverage of Islamophobic tweeting for the weekly method (compared with the daily method) is better than coverage of tweets in general. This is shown in Figure 3-3. For this to be fully verified, data would need to be collected for a longer period. Nonetheless, these initial results suggest that the weekly reporting method does not introduce considerable biases which would jeopardize the robustness of the dataset.

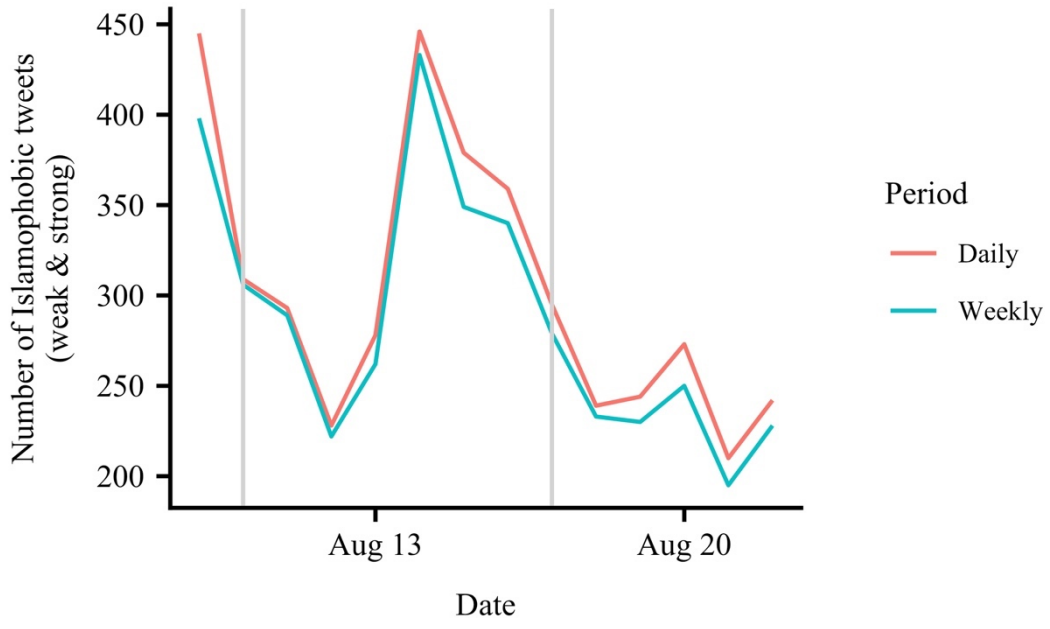


Figure 3-3, Prevalence of Islamophobia over time in tweets collected via the daily and weekly collection methods

Appendix 5.1 | Annotated dataset

The creation of an annotated dataset for the Islamophobia classification task is reported and discussed. First, the identity and background of the annotators is discussed. Second, the results of a trial preliminary study are reported. Third, the sampling method and results of the final annotation study are reported. Fourth, the annotation guidelines are provided.

5.1.1 | Annotators

As discussed in Chapter 5, annotators' identity and background is increasingly recognised as an important consideration. In the present work, three annotators are used to annotate tweets in both the main and preliminary studies. One is the author of the present work, a white middle class cis-gendered heterosexual male. The second is a PhD student from Turkey who studies left-wing politics at a university in the UK. He is a practicing Muslim, presents as Arab and is a middle class cis-gendered heterosexual male. The third annotator is a PhD student from Greek Cyprus who studies populism at a university in Italy. She is a middle-class cis-gendered heterosexual female. All three annotators are studying politics at PhD level and have a good understanding of both UK politics and research on prejudice. Whilst the annotators are all cis-gendered, middle-class and heterosexual they reflect a mix of national backgrounds (the UK, Greek Cyprus and Turkey), religious affiliations (Muslim and non-religious) and genders (two male and one female).

5.1.2 | Preliminary study

Initially, a preliminary study of 200 tweets was annotated by all three annotators using an early draft of the annotation guidelines. The purpose of this study is solely to evaluate both the annotation guidelines and the annotators; the annotations are not used to train

the classifier. The dataset consists of 100 randomly sampled tweets from far right accounts, as analysed in Chapter 3 and 100 randomly sampled tweets from followers of the BNP, as analysed in Chapter 7.

Inter-rater agreement measures how consistent different annotators are. To check inter-rater agreement I calculate percentage agreement, Fleiss' kappa and Krippendorf's alpha for all three annotators. The latter two methods are robust to chance agreement and are widely used in text annotation studies (Mchugh, 2013; McHugh, 2012). Overall inter-rater reliability is moderate, as shown in Table 5-1. I report Fleiss' kappa of 0.619, Krippendorf's alpha of 0.836 and percentage agreement of 86.1%. Kappa values can be interpreted such that 0.01-0.20 indicates none to slight agreement, 0.21-0.40 indicates fair agreement, 0.41-0.60 indicates moderate, 0.61-0.80 indicates substantial and 0.81-1.00 indicates almost perfect agreement. Only values which indicate greater than 'moderate' agreement should be used for research (Landis & Koch, 1977; McHugh, 2012, p. 279). Krippendorf's alpha values can be interpreted such that less than 0.67 indicates insufficient agreement, 0.67 to 0.80 indicate tentative agreement and values greater than 0.80 indicate definite agreement (Hallgren, 2012).

Measure	**Score**
Percentage agreement	86.1%
Fleiss' kappa (for three annotators)	0.619
Krippendorf's alpha (for three annotators with ordinal data)	0.836

Table 5-1, Inter-rater reliability scores for preliminary annotation study

Further analysis shows that the annotators have strong agreement on whether a tweet is Islamophobic or not but have considerably weaker agreement when distinguishing between 'weak' and 'strong' Islamophobia. Out of 175 tweets which are labelled 'Not Islamophobic' by at least one annotator, all three annotators agree on 161 of the tweets

(agreement of 92%). In contrast, of the 39 tweets labelled either strongly or weakly Islamophobic by at least one annotator, all three annotators agree on just 11 of the tweets (agreement of 28%). This is also reflected in the scores for category-wise Fleiss' kappa, which calculates agreement for each level of a variable (in this case, 'Strong Islamophobia', 'Weak Islamophobia' and 'Not Islamophobic'). I report agreement levels of 0.89 for 'None', 0.59 for 'Weak' and 0.76 for 'Strong', as shown in Table 5-2. This indicates that annotators only have strong agreement when annotating 'Not Islamophobic'; distinguishing between the *degree* of Islamophobia is a considerable annotation challenge in need of further investigation.

Measure	Score
Fleiss' kappa category-wise score for 'Not Islamophobic'	0.831
Fleiss' kappa category-wise score for 'Weak Islamophobia'	0.313
Fleiss' kappa category-wise score for 'Strong Islamophobia'	0.520

Table 5-2, Category-wise Fleiss kappa scores for preliminary annotation study

After the preliminary test study was completed, the results were discussed by all three annotators via Skype and points of disagreement explored in-detail. The annotation guidelines were then updated and more examples provided. Particular attention was paid to the difference between strong and weak Islamophobia.

5.1.3 | Full annotation study (4,000 tweets)

4,000 tweets are annotated to create a training set to train the classifier. To ensure the classifier can be applied robustly across the corpus all tweets in the present work, the annotated dataset is sampled from tweets analysed in both Chapters 6 and 7. The sources of tweets used to create the annotated dataset are shown in Table 5-3 below.

Source	Number of tweets
Far right seed accounts	1,000
Followers of the BNP	500
Followers of Britain First	500
Followers of the Conservatives	500
Followers of UKIP	500
Keyword search within the entire dataset of tweets (produced by followers of the BNP, Britain First, UKIP, Conservatives, Labour and followers of multiple parties)[25]	1,000
TOTAL	4,000

Table 5-3, Sources of tweets for full annotation study

5.1.3.1 | Inter-rater agreement for full study

Measure	Score
Percentage agreement	89.9%
Fleiss' kappa (for three annotators)	0.837
Krippendorf's alpha (for three annotators with ordinal data)	0.895

Table 5-4, Inter-rater reliability scores for full annotation study

Table 5-4 shows that Inter-rater agreement scores are very high across all three measures, indicating strong overall agreement. The greatest improvement is in the kappa and alpha values, which are the measures most robust to chance agreement. Whilst percentage agreement has increased by only 3.9 percentage points from the test annotation study to 89.9%, Fleiss' kappa has increased substantially by 0.218 from 0.619 to 0.837. This indicates that annotators not only agree on the 'Not Islamophobia' category but also in

[25] The keyword search is specific to Islam and Muslims: "Muslim", "Islam", "Mosque", "Halal", "Mecca", "Hajj", "Koran", "Quran", "Mohammed", "Burqa", "Burkha".

how to distinguish between Weak and Strong manifestations. This is reflected in Table 5-5 which shows the category-wise Fleiss kappa scores. Across all three levels, annotation is strong. The value for 'Weak Islamophobia' is lowest (0.737), which reflects the difficulty of identifying weaker expressions of Islamophobia; often they are the most ambiguous and can overlap with both 'Not Islamophobic' articulations and articulations of 'Strong Islamophobia'.

Measure	Score
Fleiss' kappa category-wise score for 'Not Islamophobic'	0.87
Fleiss' kappa category-wise score for 'Weak Islamophobia'	0.737
Fleiss' kappa category-wise score for 'Strong Islamophobia'	0.907

Table 5-5, Category-wise Fleiss kappa scores for full annotation study

5.1.3.2 | Intra-rater reliability in the full study

Intra-rater reliability measures how internally consistent annotators are. This is an important measure as often annotators' evaluations shift over time, usually due to either fatigue or better understanding of the dataset and annotation guidelines. Low intra-rater reliability is concerning because it indicates that either the data or the task is ambiguous or that annotators lack motivation. This considerably increases the likelihood of noise in the annotation results, reducing its robustness.

On completion of the annotation task, annotators are presented with a hundred tweets sampled randomly from the first one thousand tweets they annotated (note that all three annotators are presented with the same hundred tweets). Intra-rater reliability scores are then calculated. Both unweighted and weighted kappa are calculated; weighted kappa is higher for all three annotators, which indicates that annotators are more likely to disagree by a single category rather than confusing 'Not Islamophobic' and 'Strong

Islamophobia'. For all three annotators the scores are high (e.g. Cohen's unweighted kappa ranges from 0.883 to 0.933), which indicates very strong agreement, as shown in Table 5-6. Furthermore, for the category-wise kappa values the lowest score is 0.827, which shows that intra-rater agreement is strong for all three annotators across all categories. Overall, these results suggest that raters' annotations are internally consistent over time and as such can be trusted.

Measure	Score for annotator one	Score for annotator two	Score for annotator three
Percentage agreement	95%	97%	96%
Cohen's kappa (unweighted)	0.883	0.933	0.906
Cohen's kappa (weighted)	0.912	0.949	0.931
Krippendorff's alpha (for ordinal data)	0.917	0.942	0.938
Fleiss' kappa category-wise score for 'Not Islamophobic'	0.899	0.926	0.923
Fleiss' kappa category-wise score for 'Weak Islamophobia'	0.827	0.901	0.851
Fleiss' kappa category-wise score for 'Strong Islamophobia'	0.942	1.00	0.947

Table 5-6, Inter-rater reliability scores for the three annotators in the full annotation study

5.1.3.3 | Balanced classes in the annotated dataset

Of the 4,000 annotated tweets, all three annotators agree on 3,596 tweets and disagree about 404 (10.1%). Tweets are assigned to classes based on the majority decision. For instance, if two annotators annotate a tweet as 'Not Islamophobic' and one annotator annotates it as 'Weak Islamophobia' then it is assigned to the 'Not Islamophobic class'. All 404 tweets which do not have unanimous agreement are reviewed by the present's work author to sense check their assignments – but no tweets are moved to a different class.

In the final dataset (using majority decisions) 3,106 tweets are classed as 'Not Islamophobic', 484 tweets are classed as 'Weak Islamophobia', 410 tweets are classed as 'Strong Islamophobia'. The training set needs roughly equal classes, and as such the number of 'Not Islamophobic' tweets is sampled to 447 instances (the difference between the number of tweets in the other two classes). This creates a final dataset of 1,341 tweets.

5.1.4 | Annotation guidelines

Overview

The empirical focus of the present work is Islamophobic hate speech on social media. Building on previous academic work, Islamophobic hate speech is defined as:

"Any content which is produced or shared which expresses indiscriminate negativity against Islam or Muslims."

We then distinguish between weak and strong manifestations of Islamophobia. Strong Islamophobia is defined as:

"Speech which explicitly expresses negativity against Muslims."

Weak Islamophobia is defined as:

"Speech which implicitly expresses negativity against Muslims."

Identifying Islamophobia

Both Muslims and Islam are included within our definition as targets of Islamophobia. This is because anti-Islam negativity is often a proxy for negativity against Muslims. So, if you are trying to identify whether the 'topic' of Muslims appears in the dataset, bear these aspects in mind:

- Any reference to Muslims or Islam, or closely associated artefacts, events and practices (such as mosques, the Qu'ran, Mecca, the Hajj) means that we are potentially looking at Islamophobia.
- In particular, any explicit references to Muslims and Islam, or to Muslims *qua* group means that we are dealing with some form of 'generality'. That is, just to

use the label 'Muslims' means that someone is making some sort of general statement (however implicitly). In practice, this usually means any reference to Muslims as a group or to Islam in sweeping generalised terms. BUT! Remember that just because someone is being general, it does not mean that they are necessarily being Islamophobic – they might be expressing neutral or even positive sentiments about Muslims.

Weak and Strong Islamophobia

A key innovation in the present work – and, sadly, a very difficult task for you – is to distinguish between weak and strong varieties of Islamophobic hate speech. Please note that these terms refer only to an analytical distinction, rather than to the morality or 'impact' of tweets – weak Islamophobic tweets may still cause victims considerable harm and should be treated as seriously as strong Islamophobia.

Strong Islamophobia can vary, and includes:

- Expressing explicitly negative *views*, such as describing Muslims as barbarians
- Calling for prejudicial *actions*, such as demanding that Muslims are forcibly banned from the UK
- Expressing negative *emotions* about Muslims, such as anger and distrust, which are often articulated through the use of profanities

Examples of strong Islamophobic tweets include:

1. "Muslim men groom and rape children"
2. "Muslim mothers want to practice FGM ni the UK!"
3. "Typical, another bloody Muslim just blew himself up. LOSER"
4. "Fuck alllll Muslims"
5. "Muslim invasion, they're going to take over the UK"

6. "Top European Lawyer says that Muslims don't obey rule of law and should not be allowed to remain in Europe whilst posing a threat"
7. "The Police target Muslims because they're a problem, new #evidence"
8. "Huge rally atm against Loughborough Mosque – let's take back our country"

In example 6 the speaker is supposedly reporting someone else's claims (the 'top lawyer' that is referenced) – but nonetheless it is still the speaker who is engaging in Islamophobia as s/he is the one who has shared the content. Note also that in determining whether the tweet is Islamophobic, the 'truth' of the claims is not evaluated. Even if a claim is supported by supposed evidence, Islamophobia can still be expressed. In any given context, 'truth' is always contested, and there is no neutral objective position from which to judge the epistemology validity of any claim (B. J. Allen, 2017). Thus, whilst intuitively it seems like many Islamophobic tweets contain falsehoods, this is not the conceptual basis on which we decide whether or not they are Islamophobic.

Weak Islamophobia

Weak Islamophobia is distinguished from strong based on whether the negativity is implicit or explicit. There are two main types of weak negativity. First, is emphasizing perceived differences between Muslims and other members of society, such as attributing to Muslims strange or unusual practices. Such content excludes and marginalizes Muslims in an insidious fashion; Muslims are not explicitly targeted and attacked but, rather, their incompatibility is highlighted. This can be seen as implicitly negative as perceived differences are not celebrated but problematized. Examples include:

1. "Muslims are just different!"
2. "Muslim food smells so weird"
3. "Wearing a Burkha doesn't feel very #UK"

The second form of weak negativity is to take the tropes associated with strong negativity (such as claiming that Muslims are terrorists, barbarians or uneducated) and

to ostensibly link them to only a small subset of Muslims (e.g. to just one individual terrorist or Muslims only living in one small geographical area, such as Rotherham) – and by doing so to implicitly forge a connection between the negative trope and *all* Muslims, By using the term 'Muslims' or 'Islam', even with caveats to heighten the specificity (such as 'this Muslim terrorist' or 'Muslim Men in Rotherham'), an *implicit* connection is established with all Muslims. The key point here is that discourses about paedophiles, terrorists or FGM practitioners can often be articulated without the need to reference Muslim identity. Examples of this type of weak Islamophobia are provided below. In all of the cases, the speaker appears to be commenting on a specific case but still implicitly creates an association with the negative trope and all Muslims.

1. "Muslim terrorists attack London Bridge"
2. "Muslim radicals in the desert kill Christian hostage"
3. "Muslim pedos are sick"

Annotation process

You will be provided with a csv file with your name in the file name. This file will contain the list of tweets, each tweets' ID and some relevant metadata. Two columns will be of interest to you: 'strength' and 'comments'. Strength is where you enter your annotation. Enter '0' if there is no Islamophobia at all, '1' if weak Islamophobia is expressed and '2' if strong Islamophobia is expressed. Use the 'comments' section to explain your annotation, (if needed) flag any issues and to draw attention to any interesting features of the tweets.

Data

You will be presented with 4,000 tweets. We have:

- Removed URLs from the tweets as these do not contain any semantic content and can make reading the remaining content in the tweets more difficult.
- Removed emojis from the tweets as these can be difficult to display and may result in annotators viewing different content. From our test studies we do not believe that this will make a substantive difference to your annotations – negativity against Muslims is Islamophobic, irrespective of whether it is preceded or followed by a smiley face.
- Not provided any links to photos and other forms of media as these will not be used to train our classifier. From our test studies we believe that this content is unlikely to make a considerable difference to your annotations.

We advise caution; if you think that viewing the additional media might render a tweet Islamophobic but the text content of the tweet – *by itself* – is not Islamophobic then do **not** label it as Islamophobic. Base your annotations solely on the content you are presented with.

Final advice

- Be as literal as possible in applying the guidelines; do not over think it.
- Please take into account context! The tweets you will see are produced by far right Twitter accounts. Use your common sense to work out the nature of the tweets.
- In the UK many different groups may be victims of prejudice (likely targets include immigrants, refugees, people who are gay and people who are transgender0. Unless you think that they are being targeted as a proxy for Muslims (as with misdirected Islamophobia) do NOT include them in your annotation.
- Hashtags are equivalent to other forms of speech and should be analysed as such and taken literally – "#BanIslam" can be considered equivalent to writing "Islam should be banned".
- If you are unsure whether a tweet is either not Islamophobic or weakly Islamophobic it is best to mark it as weak Islamophobic and flag the annotation in the 'comments' section. We will then revisit the annotation you have provided.
- Overall, we anticipate that most tweets will not be Islamophobic of any sort – so do not worry if you annotate many '0' labels.

If you have any concerns or queries then please refer back to this guide and the examples provided. Do not hesitate to contact me.

Appendix 5.2 | Input feature selection

5.2.1 | Input features

- Sentiment, derived using the open-source 'SentimentAnalysis' package in R (Feuerriegel & Proellochs, 2018). The package implements several dictionary-based approaches to detect sentiment. After testing, I opt to use Mohammad and Turrney's crowd-sourced and ethical lexicon which categorises text into eight categories of emotion: trust, fear, sadness, surprise, anger, disgust, joy and anticipation, as well as positivity and negativity (Mohammad & Turney, 2013).
- Polarity, also derived using the SentimentAnalysis package in R. Polarity is measured using a dictionary of positive and negative words, which I then convert into two different scores: the net sentiment (positive + negative) and the absolute sentiment (positive + the absolute value of negative).
- Count of swear words, identified using the Office of Communication's report into offensive language on television (Ipsos MORI, 2016).
- Count of exclamation marks, count of question marks, and count of the total number of all punctuation marks
- Whether any Muslim names are mentioned (binary input), identified using a dictionary derived from the Wikipedia page, 'Arabic names'. Accessed on Monday 2nd July and available at: `https://en.wikipedia.org/wiki/Arabic_name`
- Whether any Mosques are mentioned (binary input), identified using a dictionary derived from the Wikipedia page, 'List of Mosques in the United Kingdom'. Accessed on Monday 2nd July and available at: `https://en.wikipedia.org/wiki/List_of_mosques_in_the_United_Kingdom`

- Whether any Terror attacks are mentioned (binary input), identified using a dictionary derived from the Wikipedia page, 'List of Islamist Terror attacks'. Accessed on Monday 2nd July and available at: `https://en.wikipedia.org/wiki/List_of_Islamist_terrorist_attacks`
- Count of parts of speech, identified using the 'SpaCyR' package in R (Ken Benoit & Matsuo, 2018). SpaCyR is a wrapper to the SpaCy NLP system, which uses a statistical approach to tag parts of speech. It categorises text into grammatical categories. In the dataset, I find the following parts of speech: proposition, verb, adjective, noun, participle, number, symbol, adverb, conjunction, determiner, space, pronoun, and interjection.
- Count of named entities, also identified using the 'SpaCyR' package in R. Entities are pre-defined categories which can be automatically extracted from language, such as locations, monetary values, and dates, and are similarly tagged by SpaCyR with a statistical approach. In the dataset, I find the following named entities: events, facilities, geo-political entity, law, location, organisation, person, product, nationalities.

5.2.2 | Model 7 testing

For the multi-class classifier I use the word embeddings model (model 5 in Chapter 5) as a starting point and then extend it by introducing additional variables. I test one to seven additional input features and find that a specific combination of six additional input features maximises accuracy. The accuracy of the final multi-class classifier is 74.60%. Table 5-7 summarizes the input variables that maximize accuracy at each round.

Number of variables	Variable	Accuracy	Improvement on embeddings
0	Word Embeddings model alone	72.48%[26]	/
1	+ Count of mentions of Mosques	73.38%	0.90
2	+ Count of mentions of Mosques + count of part of speech: 'determiner'	73.94%	1.46
3	+ Count of mentions of Mosques + presence of HTML + count of sentiment: 'fear'	73.95%	1.47
4	+ Count of mentions of Mosques + presence of HMTL + gross calculation of polarity + count of part of speech: 'determiner'	74.38%	1.9
5	+ Count of mentions of Mosques + presence of RT + count of named entity recognition: 'facilities' + count of named entity recognition: 'organisation' + count of named entity recognition: 'location'	74.55%	2.07
6	+ Count of mentions of Mosques + presence of HTML + presence of RT + count of part of speech: 'conjunction' + count of named entity recognition: 'location' + count of named entity recognition: 'organisation'	74.60%	2.12
7	+ Count of mentions of Mosques + presence of HTML + count of exclamation marks + count of part of speech: 'adverb' + count of part of speech: 'conjunction' + count of part of speech: 'determiner' + count of named entity recognition: 'location'	74.59%	2.11

Table 5-7, Best performing models with additional variables

[26] The accuracy of this word embeddings model alone is higher than the model shown in Chapter 5 (72.17%) because this model uses word vectors from a s model trained on the full corpus of tweets, rather than a sample.

Appendix 6.1 | Information about the BNP dataset

For the 6,611 users who are considered active persistent followers of the BNP before bots are removed, Table 6-1 shows the number of tweets in each language. 'English' is the dominant language, accounting for 7,860,423 out of the 10,229,137 tweets (76.8%). The second most prominent language is 'Undetermined', which accounts for 618,952 tweets (6%). The other most popular languages are from Europe, including Spanish, French, Dutch, German, Portuguese and Italian.

Language	Number of tweets		Language	Number of tweets
en	7,860,423		th	1,982
und	618,952		cs	1,676
es	430,741		ps	1,451
fr	404,360		eu	1,427
nl	137,025		fa	1,255
ja	112,350		lt	1,203
de	104,969		ko	1,005
pl	100,047		uk	927
it	94,652		is	833
el	74,225		ca	722
sv	48,475		vi	623
pt	38,554		bg	619
ru	22,765		sl	598
hi	21,993		ta	296
in	15,813		ne	98
ar	13,891		ml	98
tl	12,513		mr	97
tr	11,749		gu	33
fi	11,626		ckb	13
sr	11,445		sd	13
ur	8,128		bo	6
lv	8,087		te	6
da	7,284		my	5
bn	7,000		hy	4
et	6,502		ka	3
ro	5,431		or	2
ht	5,070		si	2
no	4,327		am	1
iw	4,186		pa	0
kn	4,066		km	0
cy	3,242		lo	0
zh	2,239		ug	0
hu	2,009			

Table 6-1, The number of tweets produced by active persistent followers of the BNP in each language

Appendix 6.2 | User typology of Islamophobia

For a (hypothetical) user who sends ten tweets, the random probabilities would be as follows:

1. The user tweets only none Islamophobically: 0.172
    a. Probability of a tweet being none Islamophobic = 0.839
    b. 0.839^10
    c. 0.172
2. The user tweets only weak Islamophobically: 0.000000000214
    a. Probability of a tweet being weak Islamophobic = 0.108
    b. 0.108^10
    c. 0.000000000214
3. The user tweets only strong Islamophobically: 0.000000000000185
    a. Probability of a tweet being strong Islamophobic = 0.053
    b. 0.053^10
    c. 0.000000000000185
4. The user tweets only none and weak Islamophobically: 0.406
    a. 1 – (probability of the user sending *at least* one strong tweet + probability of user sending only weak tweets + probability of user sending only none tweets)
    b. 1 – (0.422 + 0.000000000214 + 0.172)
    c. 1 – 0.594
    d. 0.406
5. The user tweets only none and strong Islamophobically: 0.147
    a. 1 – (probability of the user sending *at least* one weak tweet + probability of user sending only none tweets + probability of user sending only strong tweets)
    b. 1 – (0.680 + 0.172 + 0.000000000000185)
    c. 1 – 0.853
    d. 0.147
6. The user tweets only weak and strong Islamophobically: 0.0000000117

a. 1 – (probability of the user sending *at least* one none tweet + probability of user sending only weak tweets + probability of user sending only strong tweets)
b. 1 – (~1 + 0.000000000214 + 0.000000000000185)
c. 1 – ~1
d. 0.0000000117

7. The user tweets none, weak and strong Islamophobically: 0.275
   a. 1 – sum(all other options)
   b. 1 – (0.172, 0.000000000214 + 0.000000000000185 + 0.406 + 0.147 + 0.0000000117)
   c. 1 – 0.725
   d. 0.275

Type	**Probability**
Only none Islamophobic	0.172
Only weak Islamophobic	0.000000000214
Only strong Islamophobic	0.000000000000185
Only none and weak Islamophobic	0.406
Only none and strong Islamophobic	0.147
Only weak and strong Islamophobic	0.0000000117
None, weak and strong Islamophobic	0.275

Table 6-2, Summary of probabilities for each type of Islamophobic behaviour

Appendix 6.3 | Latent Markov model details

6.3.1 | Measuring time

Alongside the strategy used to split time total time (*T*) into separate bins (each of length *t*) outlined in Chapter 6 (where time is scaled by the overall volume of tweets), three alternative strategies can also be used.

First, time can be measured linearly based on a regular fixed interval. For instance, this could include studying each users' weekly tweeting behaviour (dividing the period of data into 51 equal-sized *t* periods) or daily tweeting behaviour (dividing it into 365 equal-sized *t* periods). Very fine-grained periods, such as a minute or second, could also be used. However, this level of precision is most likely not necessary with Twitter data given that most users only tweet a few times per day. The main problem with this strategy is that the volume of tweets sent by each user varies over time such that any two *t* periods (irrespective of how long *t* covers – whether it is a day or a week) are not comparable. Using a fixed linear time period could easily lead to biases whereby periods with a higher volume of tweets exhibit systematically different tweeting behaviours to periods which contain a lower volume of tweets. It is also likely that there are many time periods in which the majority of users do not send any tweets and as such have no value recorded for them.

Second, time can be measured by fixing time intervals based on a fixed number of tweets which each user sends. In this approach, each time interval *t* is fixed at a certain number of tweets. The number of tweets which constitutes a single period *t* is the same across all users but in principle can vary considerably, from just one tweet (i.e. each tweet is analysed separately) to many more (e.g. several hundred). This can be implemented by summing the total number of tweets for each user. Then, divide the total number of tweets

per into fixed periods of length *t*. For instance, if *t* is set to 10 tweets then a user who sends 200 tweets would have 20 recorded time periods. There are two problems with this approach. First, users send different numbers of tweets. This means that they have different lengths of behaviour; whilst *t* is fixed, *T* is free to vary for each user. Second, users tweet at different times. This means that the t_1 for one user may not correspond to the same actual time period as t_1 for another user. Thus, even though this strategy is simple, both conceptually and practically, these two limitations mean that it is inherently unsuitable for comparing values for different users.

Third, time can be measured by dividing the total volume of tweets for each user into a fixed number of *T* periods of varying *t* length. This is the inverse of the second approach: the number of periods is fixed but the number of tweets within each period varies. For instance, if *T* is set to 10 then for a user who sends 100 tweets, *t* would equal 10 tweets. For a user who sends 800 tweets then *t* would equal 80 tweets. This approach is better suited than the second approach as it ensures that users have fewer empty time periods. Provided that *T* is set so that it is close to the minimum number of tweets sent by users than most users will have a value for every *t.* In effect, this compensates for the varying volume of tweets sent by each user; users with more tweets simply have more tweets per interval rather than more intervals. However, this strategy has the same substantial drawback as the second one in that time periods for each user are unlikely to be similar. One user may tweet ten times during the period at evenly spaced intervals of four weeks; another user might also tweet ten times at evenly spaced intervals but all in one day; yet another user might tweet ten times during the period in a bursty uneven manner, with varying time periods between each tweet. This strategy does not account for this, which makes it difficult to compare different users.

6.3.2 | Measuring Islamophobia

Two alternatives could be used to the strategy adopted for measuring Islamophobia, outlined in Chapter 6. First, is to take the most prevalent type of behaviour within each time period *t* (the mode). This is an intuitive way to represent user behaviour but is likely to result in many periods classified as none Islamophobic and very few classified as strong Islamophobic (as this only accounts for 5.3% of all tweets). For instance, it is plausible that in a given period a user could send several strong Islamophobic tweets but many more none Islamophobic tweets and so be recorded as 'none' – few users are so hateful that the most frequent type of tweet they send is strongly Islamophobic. This strategy is therefore likely to perform poorly at capturing users' Islamophobic behaviour.

Second, is to use a more complex strategy, such as a hybrid in which users have to tweet a certain number or percentage of tweets in the higher categories of Islamophobia during each period *t* for them to be considered representative of that period. However, this would make it very difficult to interpret the output of the model and is unneeded.

6.3.3 | Number of latent states

I fit the number of latent states (K) in the LM model by comparing the Aikake Information Criterion (AIC) and Bayesian Information Criterion (BIC) of models with 1 to 6 latent states. For a given model, n is the number of instances (the sample size), k is the number of estimated parameters, $\hat{L}$ is the maximum value of the likelihood function and ln is the natural logarithm. The likelihood function estimates how likely a specified model is given the observed data; this allows one to compare how likely it is that models with different parameters generated the observed data. For both AIC and BIC, models with the lowest values are optimal. Note that AIC and BIC can only be used to compare models fitted on the same dataset and cannot be used to compare models fitted on different datasets.

AIC is given as:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Eq. 6-1

BIC is given as:

$$\text{BIC} = \ln(n)\,k - 2\ln(\hat{L})$$

Eq. 6-2

To ensure that the number of time periods does not unduly affect the trajectories which users are assigned to, I test for time periods (*T*) of length 10, 25, 50 and 100. Testing is implemented in the R package 'LMest' through the search.model.LM() function. The results are shown in Table 6-3. The optimal number of latent states increases as the number of time periods increases, despite the penalties imposed by both AIC and BIC. For T = 10, 3 latent states are optimal according to AIC and 4 according to BIC. I take these values as the upper limit on the number of latent states, given previous work which suggests that they can overestimate the optimal number (M. J. Green, 2014). The choice of three latent states supports the theoretical interpretation of the user behaviour in the data; namely, that there are 3 states of activity; none Islamophobic tweeting, weak Islamophobic tweeting and strong Islamophobic tweeting. Accordingly, I set the number of latent states to 3 in the LM model for T = 10.

Time period (*T*)	**Metric**	
	AIC	**BIC**
10	3	4
25	5	4
50	5	5
100	7	7

Table 6-3, Optimal number of latent states for each time period *T*

6.3.4 | Length of time period *T*

To evaluate the impact of the time period T, I fit models with the same number of user trajectories (5) but varying numbers of time periods, covering 10, 25, 50 and 100. Note that the total number of users in the LM model is 4,563 rather than 6,406, as the 'Never Islamophobes' are not included. I visually inspect the user trajectories within each model to determine which qualitative label ('Escalating Islamophobes' etc.) is appropriate. Table 6-4 summarises the number of users assigned to each trajectory. The results indicate that varying the time period only has a small impact on which trajectories are assigned to. The observed variations are somewhat expected given that the user trajectories for each model also vary – it is plausible that, because of this, liminal users are assigned to different trajectories. In-depth analysis (not presented here) shows that, as expected, users are consistently assigned to the 'Extreme Islamophobes', 'Escalating Islamophobes' and "De-escalating Islamophobes' trajectories. The primary source of variation is between 'Perpetual Islamophobes' and 'Casual Islamophobes' as these categories overlap the most. This robustness check suggests that, in the future, more work could be undertaken to further segment and separate these trajectories.

User trajectories	Mean	Standard deviation	T = 10	T = 25	T = 50	T = 100
Escalating Islamophobes	406	29	382 (-24)	382	420	440
Perpetual Islamophobes	700	190	864 (+164)	864	524	548
De-escalating Islamophobes	324	16	313 (-11)	313	346	326
Casual Islamophobes	2,160	153	2,028 (-132)	2,028	2,307	2,278

Extreme Islamophobes	973	5	976 (+3)	976	966	971

Table 6-4, Number of users assigned to each trajectory for models with varying time periods

6.3.5 | Number of clusters

I cluster the user behavioural trajectories, based on their latent states in the LM model, into *typified* user trajectories using the k-modes clustering algorithm (Huang, 1998), which is an extension of the widely used k-means algorithm, and used in previous research using LM modelling (Druce, McBeth, et al., 2017). K-modes partitions N response vectors into K clusters based on the modal values for each item in the vector sequence – in this case, a separate modal value is calculated for each of the ten time periods. Then, based on the distance between each response vector, they are clustered together. This method is particularly suitable in this case as k-modes gives the same penalty for all values which are not equal to the mode; as there are only three levels for each item, this assumption is unlikely to introduce considerable biases (it poses greater problems when each item can take on many values). Although it does not explicitly model time, k-modes takes into account the ordering of values and as such is suitable for this application.

K-modes finds a local optimum, and as such the output can vary depending on how it is randomly initialized. As such, I repeat the analysis five times using different random initializations and find that the results are similar in all cases. The results of fitting the number of clusters is shown in Figure 6-1 for one of the random initializations. Using the 'elbow method' (a visual inspection of the second derivative, i.e. the point at which the rate of improvement in fit decelerates) I find the optimum number is between 5 and 10. I inspect the user trajectories created for each number of clusters and conclude that five

typified user trajectories best represents the data. The issue here is of balancing generalisability with explanation; a model with three user trajectories splits users into clusters based on whether their dominant latent state is either none, weak or strong – which is useful but not particularly nuanced. In contrast, a model with ten user trajectories splits users into clusters with very complex behavioural patterns, which cannot be easily described and differentiated. I opt for five as this produces typified user trajectories which are mutually exclusive and collectively exhaustive. Fitting for six, seven and eight trajectories produces models with duplicate trajectories.

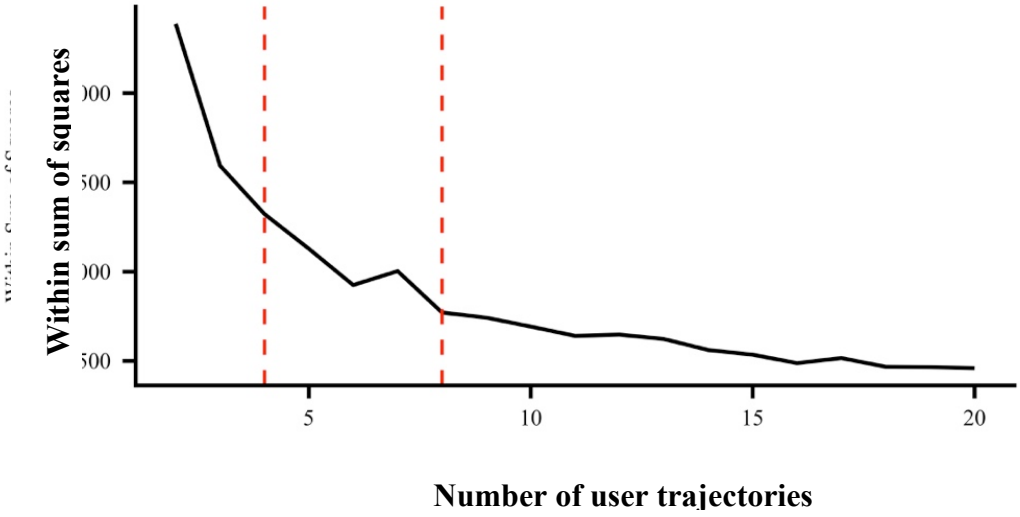


Figure 6-1, Model fitting for the number of user trajectories

6.3.6 | Typified user trajectories

For the typified user trajectories I plot three figures, just one of which (Figure 6-2) is in Chapter 6. Figure 6-2 shows the probabilities for users in each typified trajectory of being in each of the three latent states. This is the most theoretically informative figure as it shows the underlying latent states which govern users' actual manifest behaviour. Figure 6-3 shows the probabilities for users in each typified trajectory of exhibiting each of the three types of behaviour. This is calculated by taking the probabilities assigned to the

latent states and multiplying them out by the behavioural probabilities associated with each latent state (shown in the main chapter). As anticipated, there is more variety in these values. Interestingly, both 'Casual Islamophobes' and 'Extreme Islamophobes' show considerably more weak Islamophobic behaviour in this figure. Figure 6-4 shows the empirical prevalence of each type of behaviour. To calculate these values, I take a count of users' behaviour in each trajectory at each time period and then convert this into a percentage. The three figures show the same overall pattern for each typified user trajectory, which indicates that the six trajectories show here capture meaningful differences in how users behave.

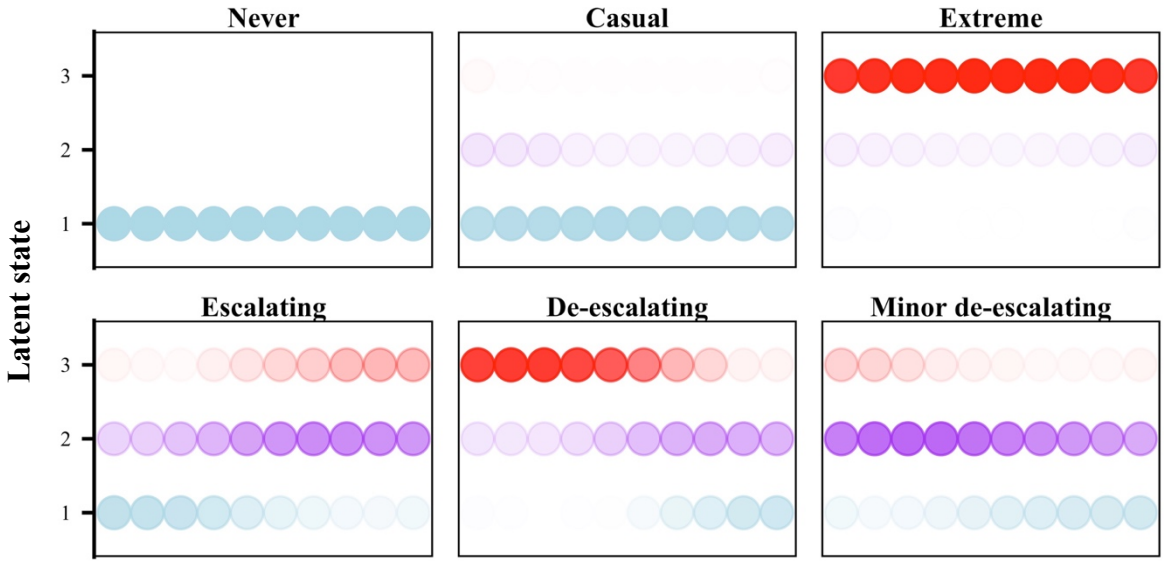


Figure 6-2, Probabilities for users in each typified trajectory being in each of the three latent states (T = 10), *shown in the main chapter*

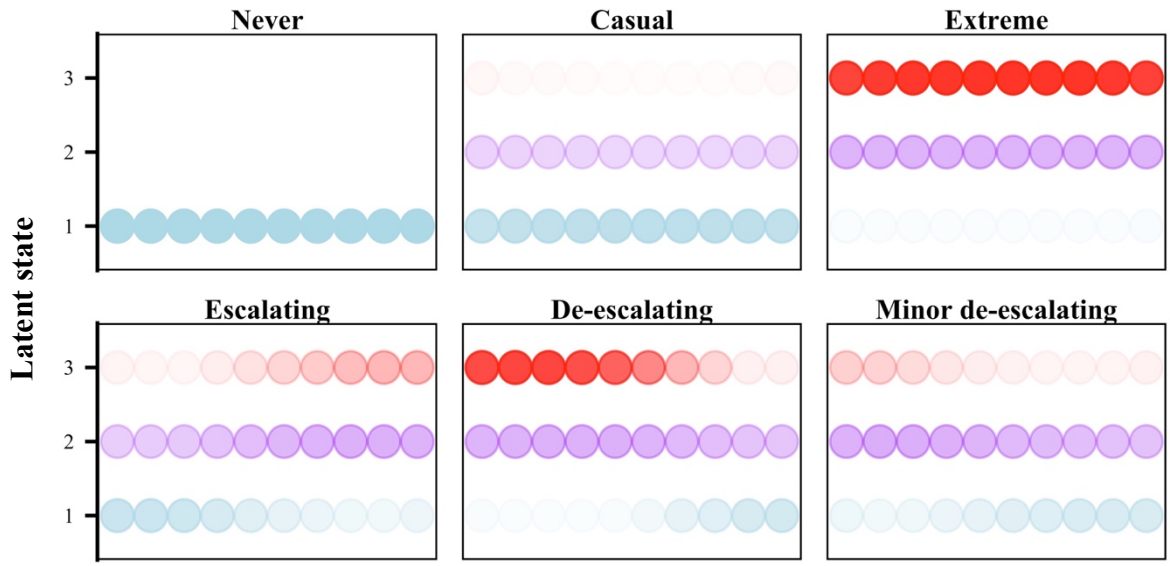


Figure 6-3, Probabilities of Islamophobic behaviour for users in each typified trajectory at each time point (T = 10)

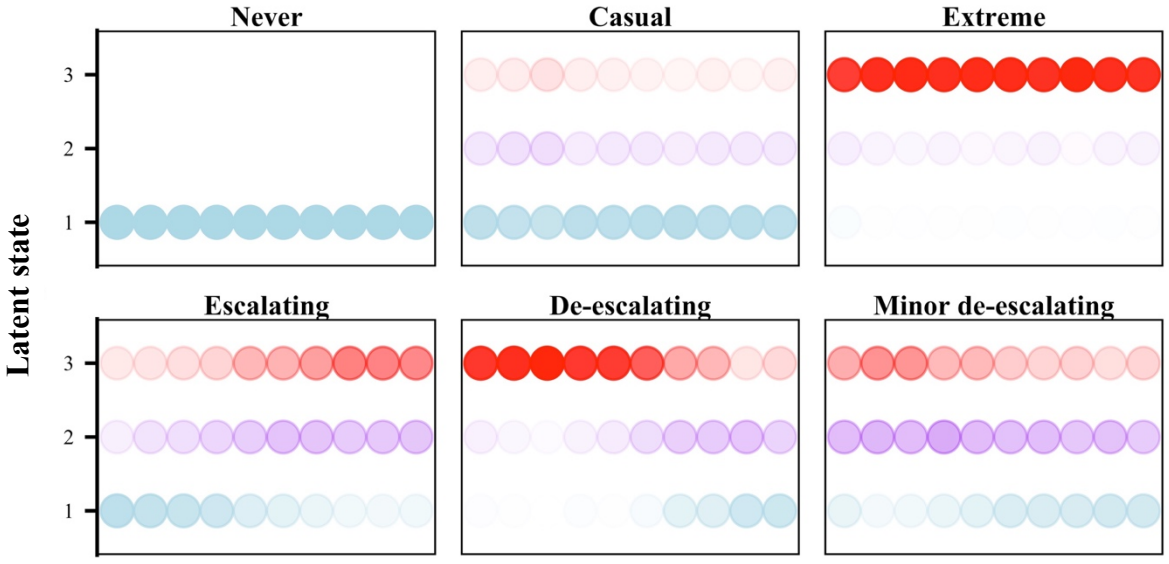


Figure 6-4, Empirical prevalence of Islamophobic behaviour for users in each typified trajectory at each time point (T = 10)

6.3.7 | Prediction with less data

Table 6-5 shows the performance of the LM models at predicting the aggregate behaviour of users for each future time period. Only the results for time period 10 are shown in Chapter 6.

Latent model	Behaviour	Time period 10	Time period 9	Time period 8	Time period 7
Actual values	None	2,400	2,447	2,309	2,381
	Weak	679	724	660	704
	Strong	1,484	1,392	1,594	1,478
LM_6	None	2,242	2,236	2,230	2,223
	Weak	685	678	671	662
	Strong	1,636	1,649	1,662	1,678
LM_7	None	2,348	2,350	2,354	
	Weak	675	666	657	
	Strong	1,540	1,547	1,552	
LM_8	None	2,353	2,631		
	Weak	669	781		
	Strong	1,541	1,151		
LM_9	None	2,429			
	Weak	665			
	Strong	1,469			

Table 6-5, Performance of models LM_6, LM_7, LM_8 and LM_9 versus the actual values for each time period

Appendix 7.1 | Political party followers

7.1.1 | Distribution of users' tweets which are Islamophobic

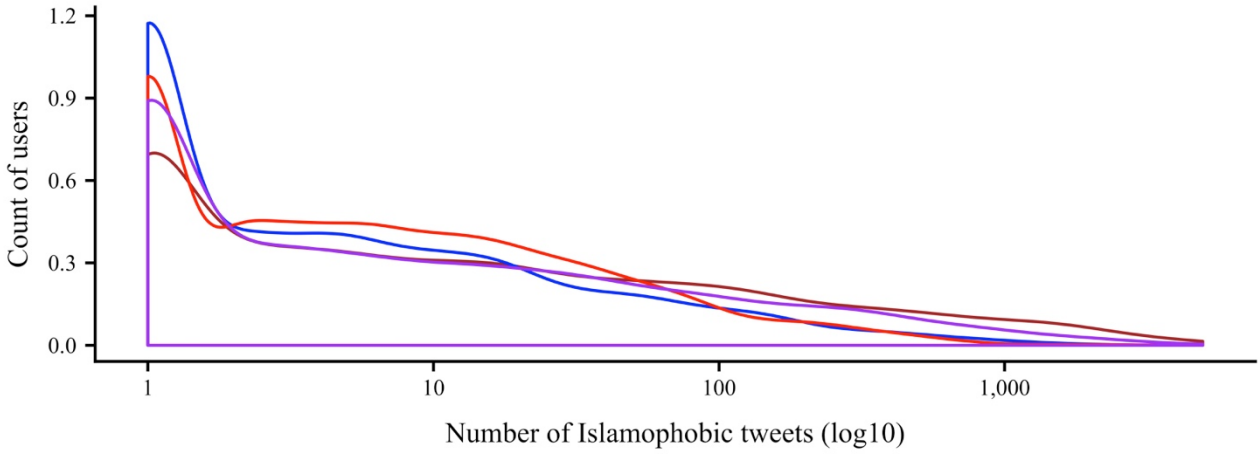


Figure 7-1, Density plot for the number of Islamophobic tweets for each user, split by party – zero values accounted for by adding 1 to each users' count

7.1.2 | Fixed effect linear regression model, model fitting

In this section I report on the residuals plot for the fixed effect OLS regression model with time granularity of 1,000,000 seconds. Figure 7-2 shows the distribution of residuals, which is well-approximated by a normal distribution. This is also demonstrated by the normal QQ-plot in Figure 7-3. Figure 7-4 shows the residuals plot. The residuals are broadly homoscedastic with only a few variations; the large number of data points in the plot (n = 280,312) should be considered when examining this panel. These plots indicate that the two key assumptions of linear regression (homoscedasticity and normal distribution of residuals) are met, which justifies the use of this method.

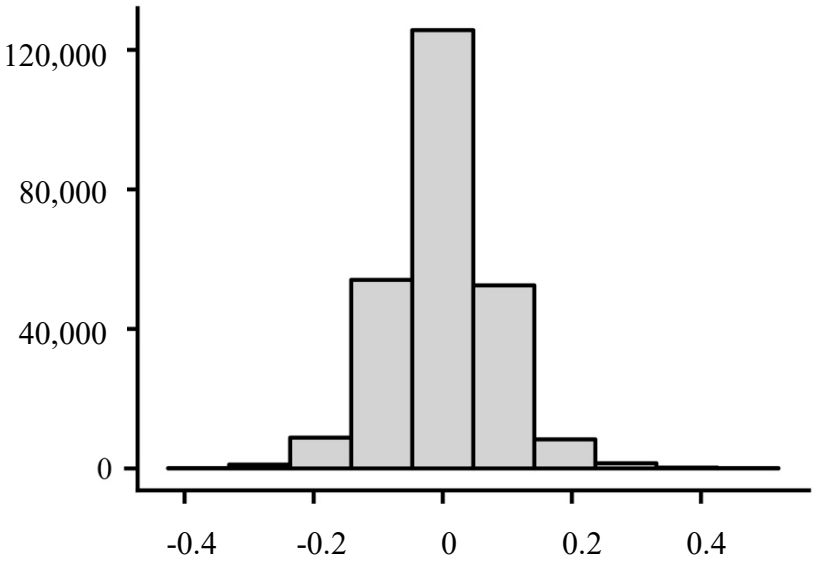


Figure 7-2, Distribution of residuals in fixed effect linear regression model with time granularity of 1,000,000 seconds

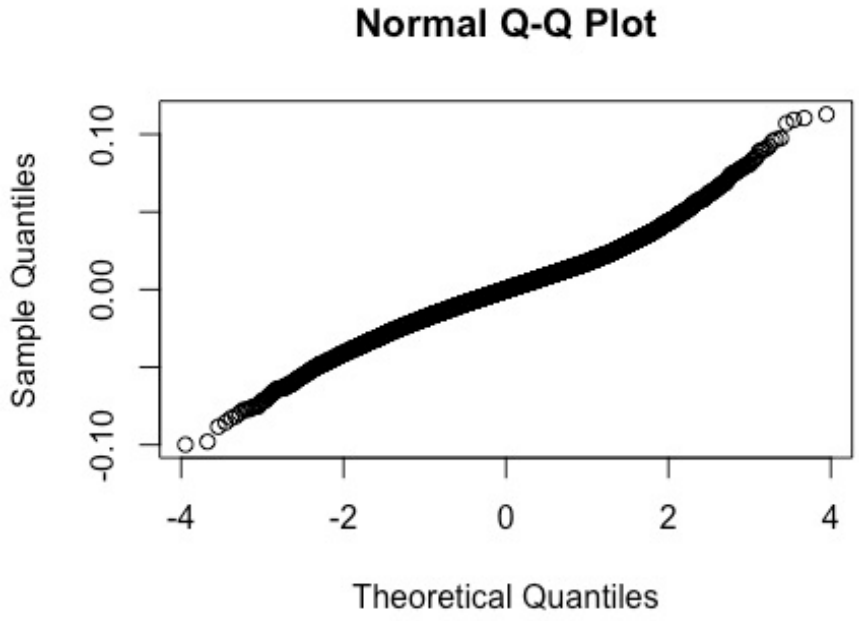


Figure 7-3, Normal QQ plot of residuals in fixed effect linear regression model with time granularity of 1,000,000 seconds

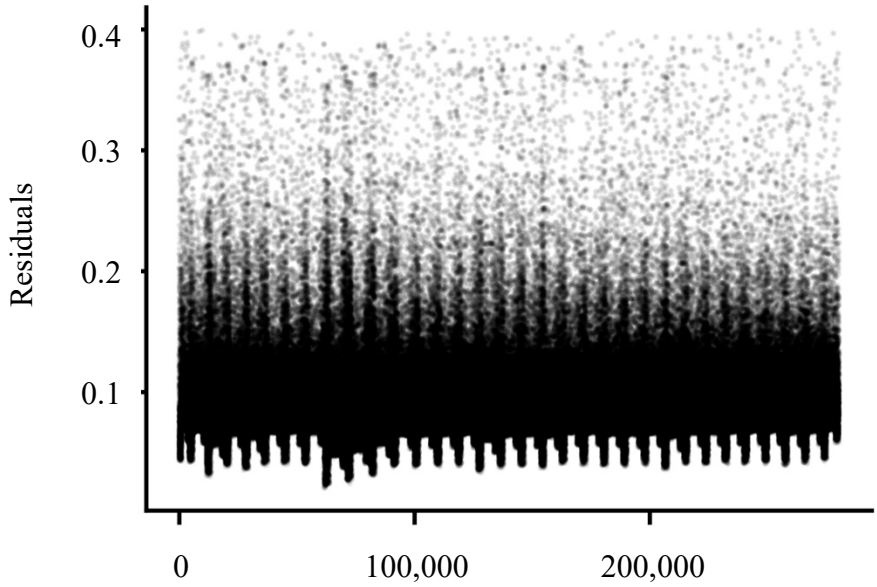


Figure 7-4, Distribution of residuals in fixed effect linear regression model with time granularity of 1,000,000 seconds

Appendix 7.2 | Further information on terrorist attacks analysis

7.2.1 | Terror attacks

Table 7-3 shows all terror attacks in the UK around the period studied. The rows in grey pertain to attacks which either fall out of the period studied or are not ideologically Islamist. The data is taken from the Wikipedia page, 'List of terrorist incidents in Great Britain'[27] and verified by reviewing 'The report of the Independent reviewer of terrorism, legislation on the operation of the terrorism acts 2000 and 2006' (Hill, 2018). In addition, a noteworthy potential terror attack occurred in June 2017 (and was widely reported on in February 2018), when a far right activist was prevented from attacking a gay pride event. This attack is not included in Table 7-3 as it did not actually occur.

Date	Name	Ideology	Number Killed	Number Injured
2016 June 16	Murder of Jo Cox	far right	1	3
2017 March 22	Westminster	Islam	4	50
2017 May 22	Manchester	Islam	22	139
2017 June 3	London Bridge	Islam	8	48
2017 June 19	Finsbury Park	far right	1	0
2017 September 15	Parsons Green	Islam	0	30

Table 7-3, Terrorist attacks committed in the UK, covering 2016 to 2018

Table 7-4 shows Islamist terrorist attacks committed during 2017 and 2018 in Europe and the USA. The rows in grey relate to attacks which fall outside of the period studied. The data is taken from the Wikipedia page, 'Islamic terrorist events in Europe'[28] and

[27] Wikipedia page, 'List of terrorist incidents in Great Britain', available at: https://en.wikipedia.org/wiki/List_of_terrorist_incidents_in_Great_Britain#2010s, accessed on 2nd November 2018

[28] Wikipedia page, 'Islamic terrorist events in Europe', available at: https://en.wikipedia.org/wiki/Islamic_terrorism_in_Europe, accessed on 2nd November 2018

'Terrorism in the United States'.[29] It is verified by reviewing Europol's 2017 and 2018 reports, 'European Union Terrorism Situation and Trend Report' (Europol, 2017, 2018).

Date	Details	# Killed	# Injured	Country
2017 January 1	Istanbul shooting	39	70	Turkey
2017 April 3	Suicide underground	15	64	Russia
2017 April 7	Hijacked truck	5	14	Sweden
2017 April 20	Police shooting	1	3	France
2017 June 6	Algerian PhD student	0	1	France (Paris)
2017 June 19	Gendarmerie	0	0	France (Paris)
2017 June 20	Central Station	0	0	Belgium
2017 July 28	Asylum seeker attack	1	6	Germany
2017 August 9	Soldiers attack	0	6	France
2017 August 17	Ramblas	16	152	Spain
2017 August 18	ISIS attack	2	8	Finland
2017 August 25	machete attack	0	1	Belgium
2017 October 1	Knife attack	2	0	France
2017 October 11	New York truck attack	8	11	USA
2018 May 23	Carcassone	4	15	France
2018 May 12	Paris	1	4	France
2018 May 29	Liege	4	4	Belgium
2018 August 31	Amsterdam	0	2	Netherlands

Table 7-4, Islamist terrorist attacks committed in Europe and the USA, during 2017 and 2018

[29] Wikipedia page, 'Terrorism in the United States', available at: `https://en.wikipedia.org/wiki/Terrorism_in_the_United_States`, accessed on 2nd November 2018

7.2.2 | UK Islamist terror attacks

In this section, I provide greater detail on each of the four terror attacks. All four of the attacks follow broadly similar dynamics, although they are considerably less pronounced for Parsons Green. Additional detail is provided on the Westminster terror attack to contextualise the results.

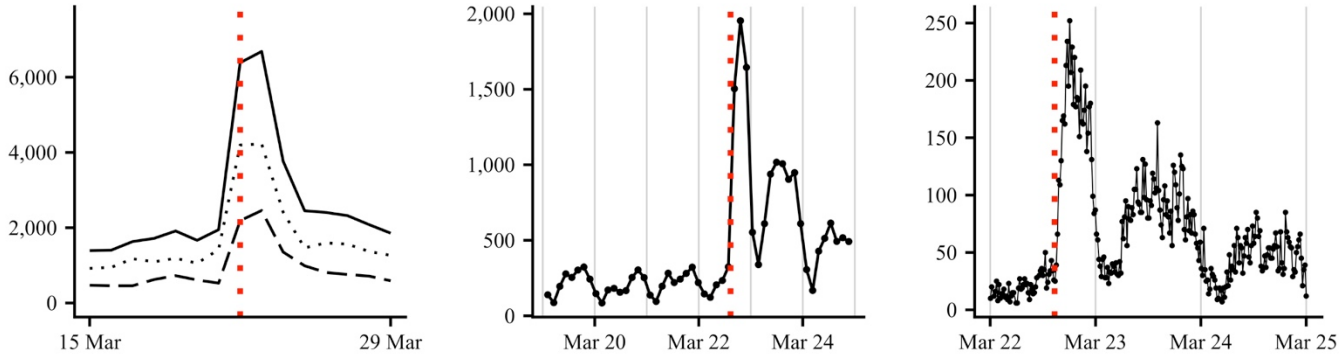
7.2.2.1 | Westminster terror attack (22nd March 2017)

Figure 7-5 shows the temporal dynamics of Islamophobic tweeting, aggregated for all parties, around the attack. The time of the attack is shown in red.[30] Panel A shows the volume of Islamophobic tweets sent each day, split into weak (dotted line on top) and strong (dashed line on the bottom). As expected, given the correlation coefficient between weak and strong over the entire period of 0.939 (reported above), weak and strong follow very similar patterns. As such, for the other two panels I only show the total volume of Islamophobia. Panel B shows a reduced time period of 6 days (the day of the attack, three prior days and two subsequent). The number of tweets is measured every 10,000 seconds (~3 hours). Panel C shows the day of the attack and the two subsequent days. The number of tweets is measured every 1,000 seconds. As expected, the variability increases as the time granularity decreases. Nonetheless, it is striking that even in Panel C the number of Islamophobic tweets is high at all times on the day of the attack.

There is a large peak in Islamophobic tweeting immediately following the attack. The following day there is also an increase in Islamophobia compared to the period prior the attack but by the next day (attack + 2 days) the level of Islamophobia has reduced

[30] The time of the attack is taken from news reports, taken from the Nexis database, and then verified by checking the dedicated Wikipedia page.

considerably. This requires further quantitative investigation, but it provides initial evidence that the impact of terrorist events is intense but short-lived. Apart from the large peak in Islamophobic tweeting, a striking aspect of these plots (particularly panel B), is the circadian rhythm – the natural 24-hour pattern of activity and rest which living



A | Strength of Islamophobia B | Islamophobic tweets C | Islamophobic tweets

creatures follow, and has been shown to influence many aspects of online behaviour (Yasseri, Sumi, & Kertesz, 2012).

Figure 7-5, The temporal dynamics of Islamophobic tweeting around the Westminster terrorist attack, aggregated for all parties

Terrorist attacks are likely to drive an increase in Islamophobia. Figure 7-6 shows the temporal dynamics of Islamophobic tweeting, split by party. The number of Islamophobic tweets is measured at 1,000 second intervals (~16 minutes). The plot shows that, broadly, the volume of tweets sent by the parties is line with their ranked volumes; BNP sends the most, followed by UKIP and then Conservatives and Labour exhibit similar, and far more muted, levels of Islamophobic tweeting. It is surprising the extent to which Conservatives and Labour closely match each other – pointing to close links in the behavioural patterns of the followers of these parties.

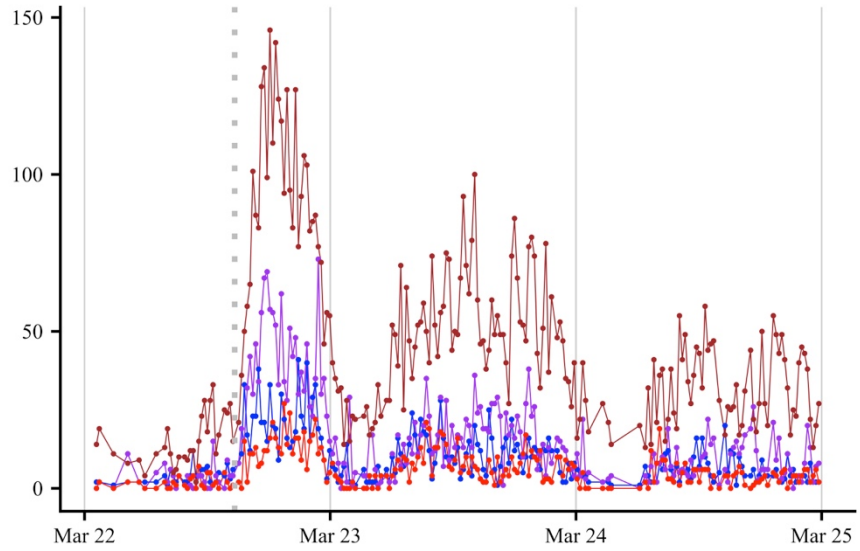


Figure 7-6, The temporal dynamics of Islamophobic tweeting around the Westminster terrorist attack, split by party

7.2.2.2 | Manchester arena (22nd May 2017)

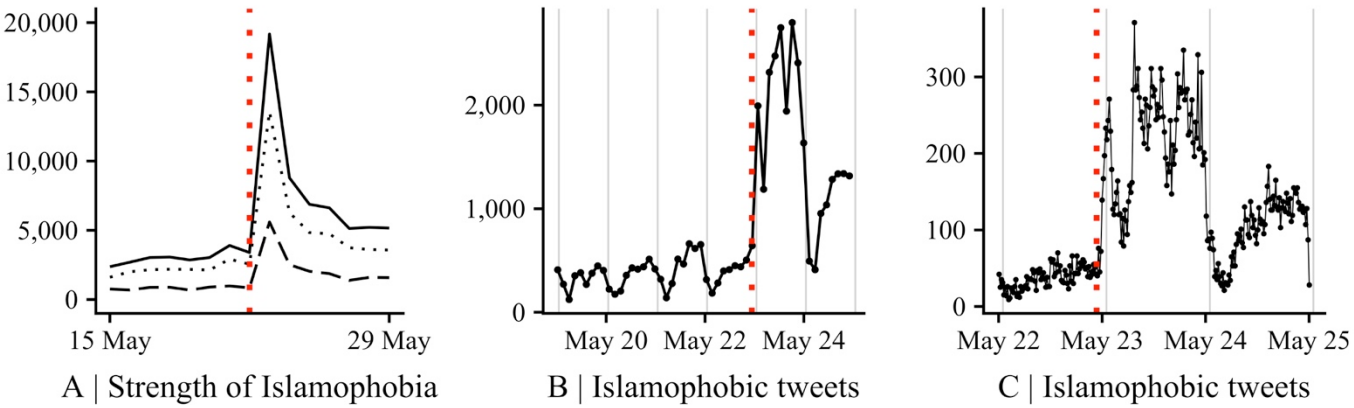


Figure 7-7, The temporal dynamics of Islamophobic tweeting around the Manchester Arena attack, aggregated for all parties

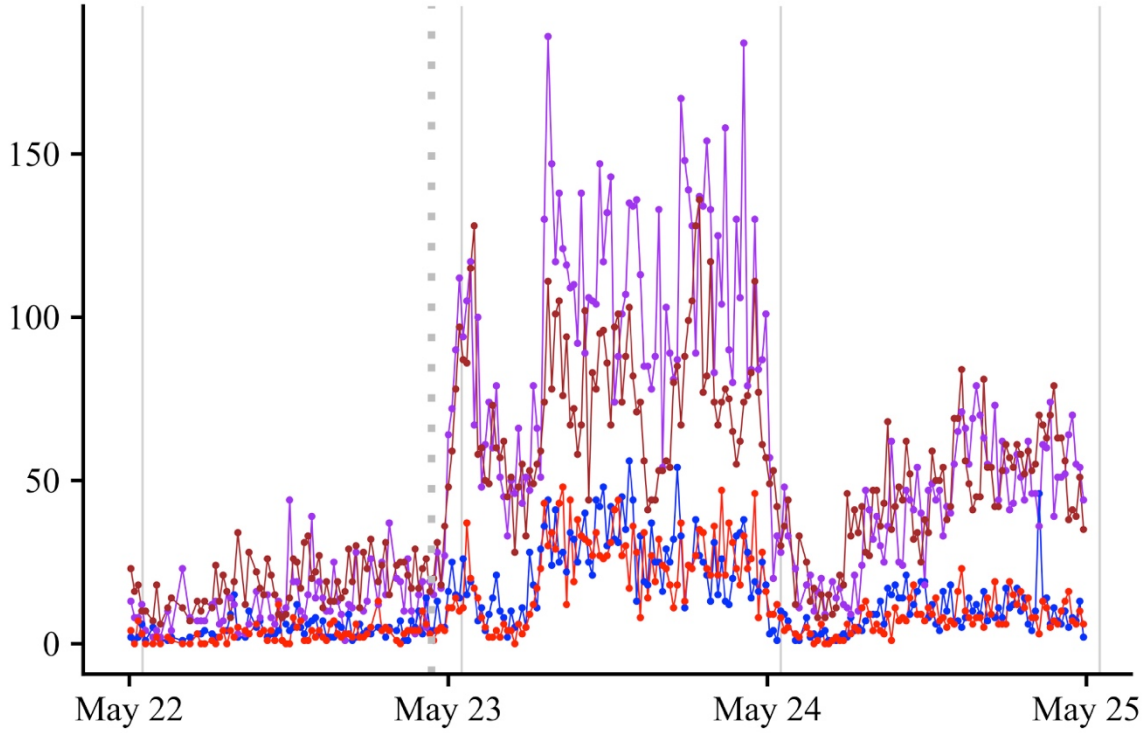


Figure 7-8, The temporal dynamics of Islamophobic tweeting around the Manchester Arena attack, split by party

7.2.2.3 | London Bridge (3rd June 2017)

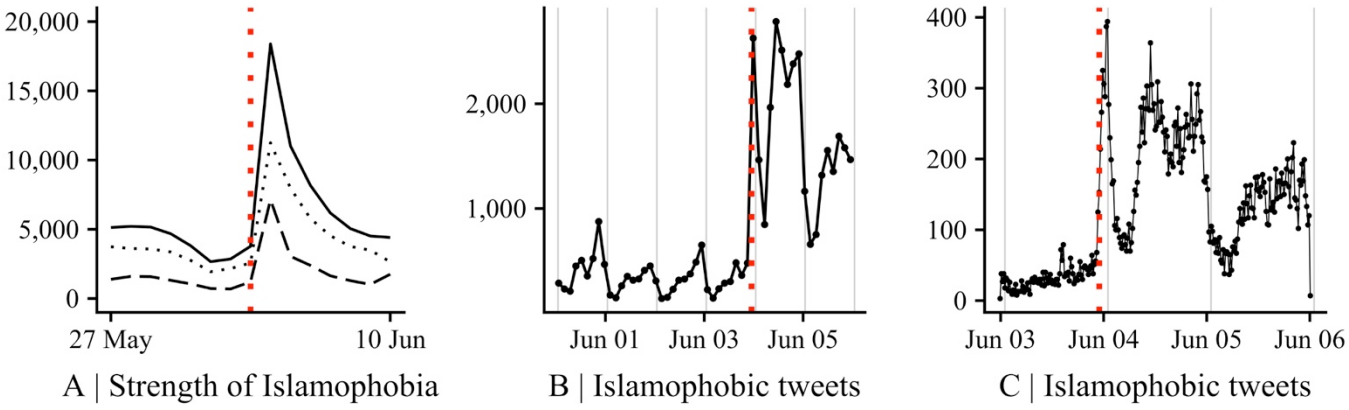


Figure 7-9, The temporal dynamics of Islamophobic tweeting around the London Bridge attack, aggregated for all parties

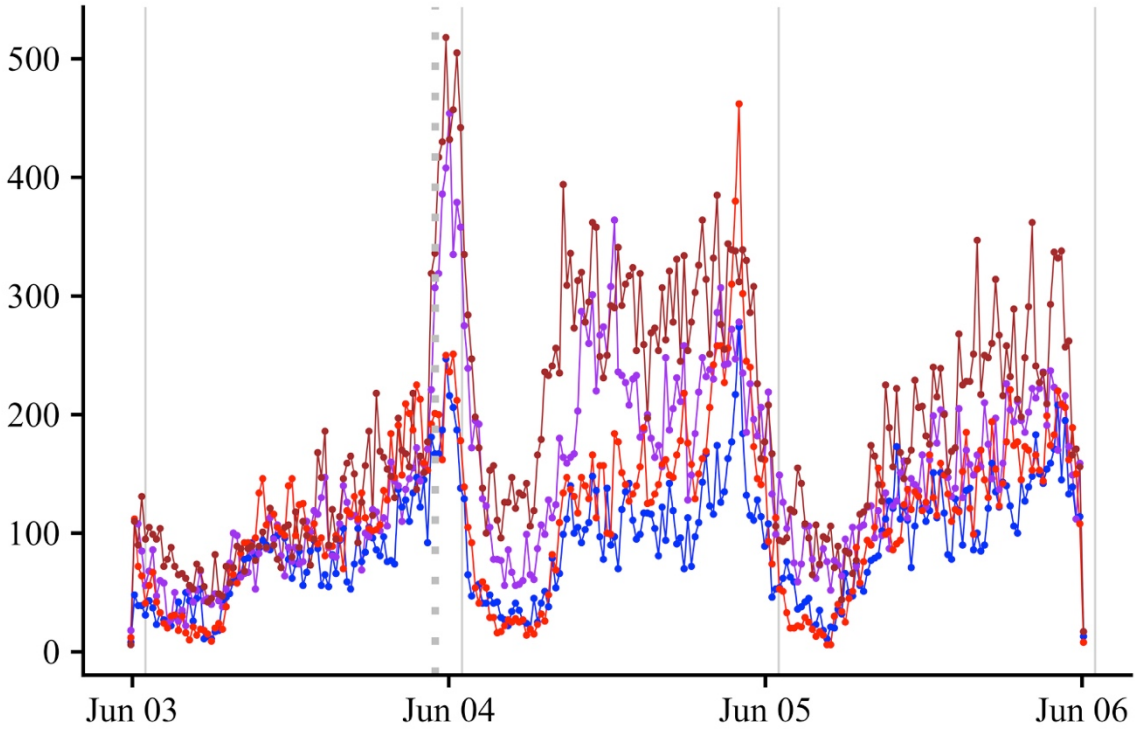


Figure 7-10, The temporal dynamics of Islamophobic tweeting around the London Bridge attack, split by party

7.2.2.4 | Parsons Green (15th September 2017)

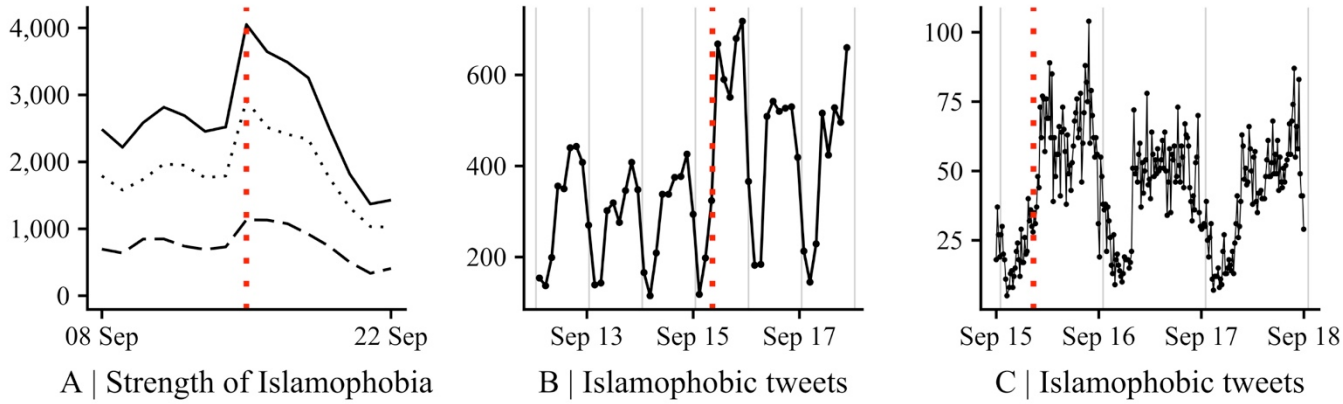


Figure 7-11, The temporal dynamics of Islamophobic tweeting around the Parsons Green attack, aggregated for all parties

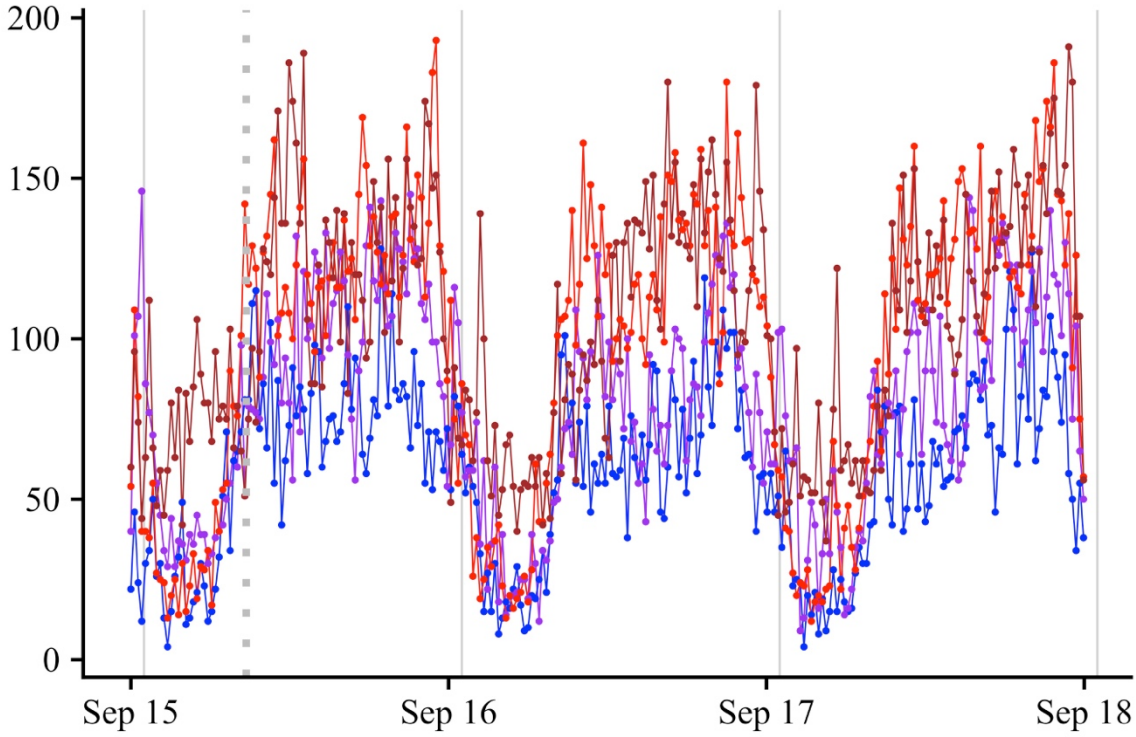


Figure 7-12, The temporal dynamics of Islamophobic tweeting around the Parsons Green attack, split by party

7.2.3 | Details on segmented regression

7.2.3.1 | OLS segmented regression model with time granularity of 10,000 seconds

The OLS segmented regression model differs from the negative binomial model in two important ways. First, the coefficients can be interpreted straight from the output; the second slope has a coefficient of 0.0598 which means that for every 10,000 second time interval which passes the probability of a user sending an Islamophobic tweet increases by 0.0598. Second, is that the breakpoints are 'hard' in this model and the slope in between each of the breakpoints is linear. Note that the results of this model are very similar to the results of the negative binomial segmented regression models below. The model is plotted in Figure 7-13.

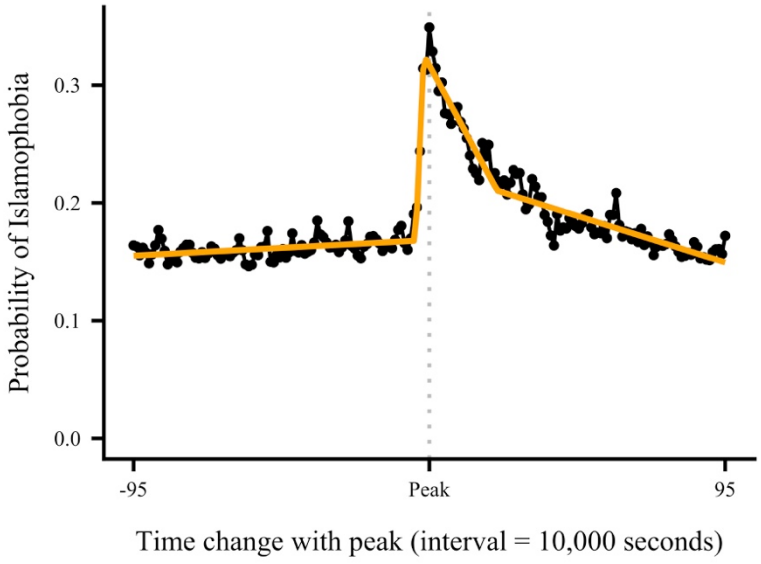


Figure 7-13, OLS segmented regression for all four attacks with time granularity of 10,000 seconds

Statistic	Model 7-1 OLS model (10,000 seconds)
Estimated breakpoints	-10, 0, 30
y-intercept	0.1685 *****
Slope 1	0.00014 *
Slope 2	0.0598 *****
Slope 3	-0.00487 *****
Slope 4	0.00083 *****
Change vs. slope 1	0.05966
Change vs. slope 2	-0.06468
Change vs. slope 3	0.00404
Breakpoint 1	-4.4306 *****
Breakpoint 2	-1.7940 *****
Breakpoint 3	21.8123 *****
Convergence	10 iterations
Multiple R-squared	0.5916
Adjusted r-squared	0.5878

Table 7-5, Summary of OLS segmented regression model, all four attacks

7.2.3.2 | Negative binomial segmented regression models with different time granularities

The results of fitting negative binomial segmented regression models for different time granularities is shown in Table 7-6. Plots for each of the models are shown below. Figures 7-14, 7-15 and 7-16 show lines from the fitted values in the segmented regression models on top of the average count of Islamophobic tweets across all 4 attacks.

The results indicate that the choice of three breakpoints is well-justified, and that the temporal dynamics hold a cross different time granularities. However, with a time granularity of 100,000 seconds there are only 19 recorded time intervals. This is too few data points to fit 3 break points in the model, and only 1 break point can be used. As Figure 7-16 shows, this model performs poorly at capturing temporal changes in the prevalence of Islamophobia. For this reason, I focus on more granular time periods in the

analysis in the main chapter. Figure 7-14 shows the regression curve for time granularity of 1,000 seconds. The curve does not capture the full extent of the spike. Further, increasing the number of break points from 3 to either 4 or 5 does not substantially increase the curve's coverage of the spike. This suggests that extremely granular time periods (in this case, just ~16 minutes) cannot be easily modelled. Figure 7-15 best represents the data.

Statistic	**Model 7-2** **NB model (1,000 seconds)**	**Model 7-3** **NB model (10,000 seconds)**	**Model 7-4** **NB model (100,000 seconds)**
Estimated breakpoints	-100, 0, 300	-10, 0, 30	0
y-intercept	30 *****	331 *****	3,310 *****
Slope 1	1.00019 *****	1.002 ***	1.124 *****
Slope 2	1.0368 *****	1.524 *****	0.890 *****
Slope 3	0.996 *****	0.965 *****	/
Slope 4	0.999 *****	0.993 *****	/
Change vs. slope 1	1.0366	1.520	0.792
Change vs. slope 2	0.9604	0.633	/
Change vs. slope 3	1.0035	1.029	/
Breakpoint 1	-45.666 *****	-7.207 *****	0.375 *****
Breakpoint 2	0.998 *****	-3.698 *****	
Breakpoint 3	272.199 *****	29.556 *****	/
Convergence	6 iterations	65 iterations	75 iterations
Pseudo r-squared (Cox Snell)	0.486	0.529	0.436
Pseudo r-squared (Nagelkerke)	0.486	0.529	0.436
Pseudo r-squared (Pearson)	0.408	0.42	0.339
Dispersion parameter	3.1087	3.6483	4.8508

Table 7-6, Summary of negative binomial segmented regression models with different time granularities, all four attacks

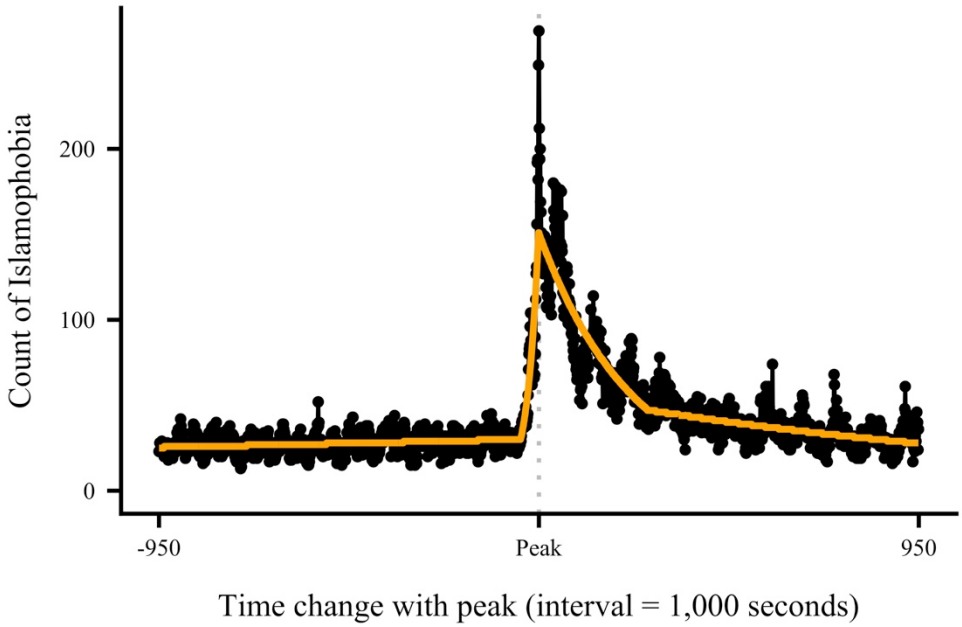


Figure 7-14, Negative binomial segmented regression for all four attacks with time granularity of 1,000 seconds

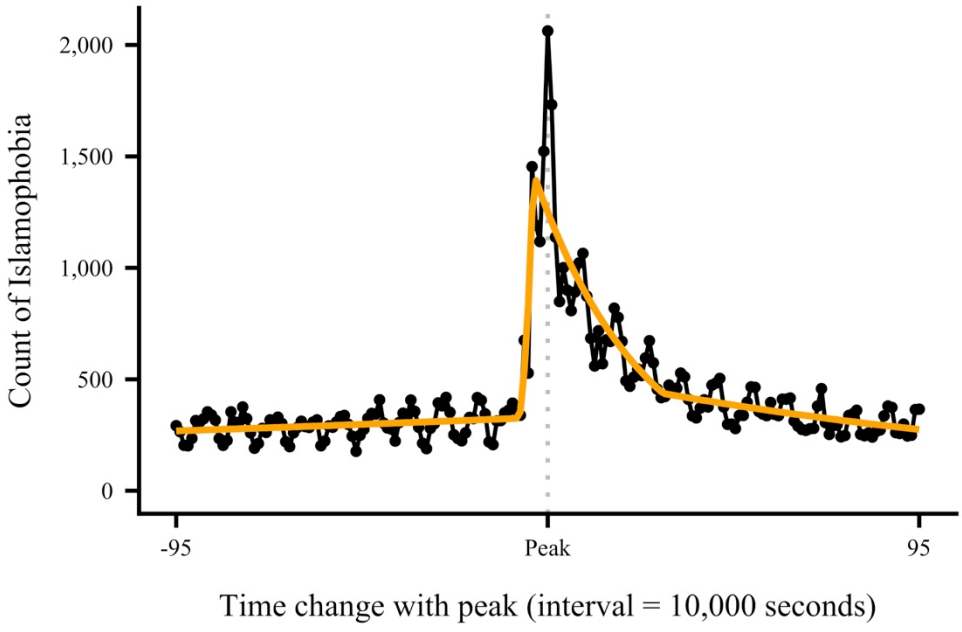


Figure 7-15, Negative binomial segmented regression for all four attacks with time granularity of 10,000 seconds (*presented in Chapter 7*)

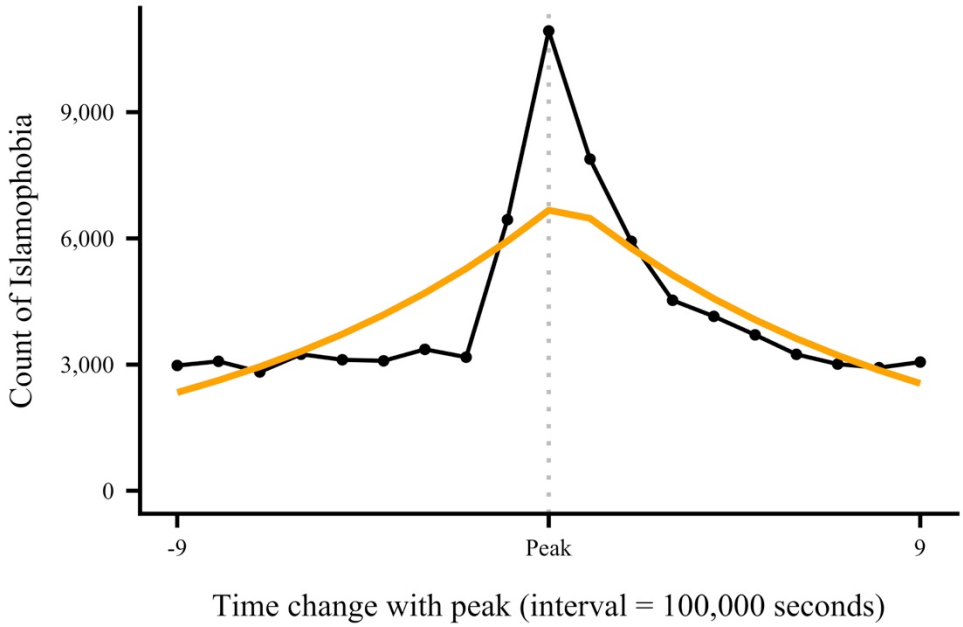


Figure 7-16, Negative binomial segmented regression for all four attacks with time granularity of 100,000 seconds

7.2.3.3 | Time: simulation of data for the negative binomial segmented regression model

In the negative binomial segmented regression models above I simulate the count of tweets for the London Bridge attack at each time interval *prior* to the attack. This because it overlaps with the aftermath of the Manchester attack, thereby biasing the results. I simulate the data by using the mean count and standard deviation of Islamophobic tweets in the day immediately prior to the London Bridge attack – this time period accounts for circadian patterns in tweeting. This is discussed in Chapter 7. To check this modelling decision, I also fit a negative binomial segmented regression on an averaged dataset with the time series prior to the London Bridge attack removed. The results of this model are shown in Table 7-7, alongside the results of the non-averaged simulated model (shown in Table 7-6 above).

There are two key similarities in the models. Crucially, the slope estimates are nearly identical in all cases, which suggests that both models capture the same *dynamics* of Islamophobic tweeting. This is also reflected by the fact that the breakpoint estimates are very similar too. There are also two differences. First, the pseudo R-squared values are far lower in the non-averaged model, which reflects the large variety in the count of Islamophobic tweets which are sent after each attack. Second, the y-intercept is also higher in the non-averaged model, which is a product of both the data simulation and in particular the mean value used, which is higher than the mean count of Islamophobic tweets before both the Parsons Bridge and Westminster attacks. This suggests that simulating the data is a reasonable decision as the dynamics of Islamophobia are similar with the averaged model. As such, due to its greater interpretability and fidelity to the data I use the non-averaged simulated model for the Chapter.

Statistic	Model 7-5 NB model (10,000 seconds)	Model 7-6 Averaged NB model (10,000 seconds)
Estimated breakpoints	-10, 0, 30	-10, 0, 30
y-intercept	331 *****	238 *****
Slope 1	1.002 *	1.003 *****
Slope 2	1.524 *****	1.513 *****
Slope 3	0.965 *****	0.965 *****
Slope 4	0.993 *****	0.993 *****
Change vs. slope 1	1.520	1.509
Change vs. slope 2	0.633	0.638
Change vs. slope 3	1.029	1.029
Breakpoint 1	-7.207 *****	-7.413 *****
Breakpoint 2	-3.698 *****	-3.368 *****
Breakpoint 3	29.556 *****	29.548 *****
Convergence	65 iterations	127 iterations
Pseudo r-squared (Cox Snell)	0.529	0.984
Pseudo r-squared (Nagelkerke)	0.529	0.984
Pseudo r-squared (Pearson)	0.42	0.825
Dispersion parameter	3.6483	18.776

Table 7-7, Summary of averaged and non-averaged negative binomial segmented regression models

7.2.3.4 | Negative binomial segmented regression without the Parsons Green attack

Qualitative analysis of Parsons Green shows that it exhibits very different temporal dynamics to the other three attacks. Here, I report on the negative binomial model with a time granularity of 10,000 seconds without the Parsons Green attack. Figure 7-17, and the regression coefficients in Table 7-8, are very similar to the model with all four attacks included (reported above), and most crucially the changes in the slopes at each breakpoint are similar and the signs are the same. The only important difference is that the maximum number of Islamophobic tweets is greater (as shown in the plots) – but the y-intercept in both models is very similar. The pseudo R-squared is higher in the model with the Parsons Green attack removed, which reflects the extent to which that attack differs from the others. Overall, these results indicate that including the Parsons Green attack does not have a material impact on the overall interpretation of the output in Chapter 7.

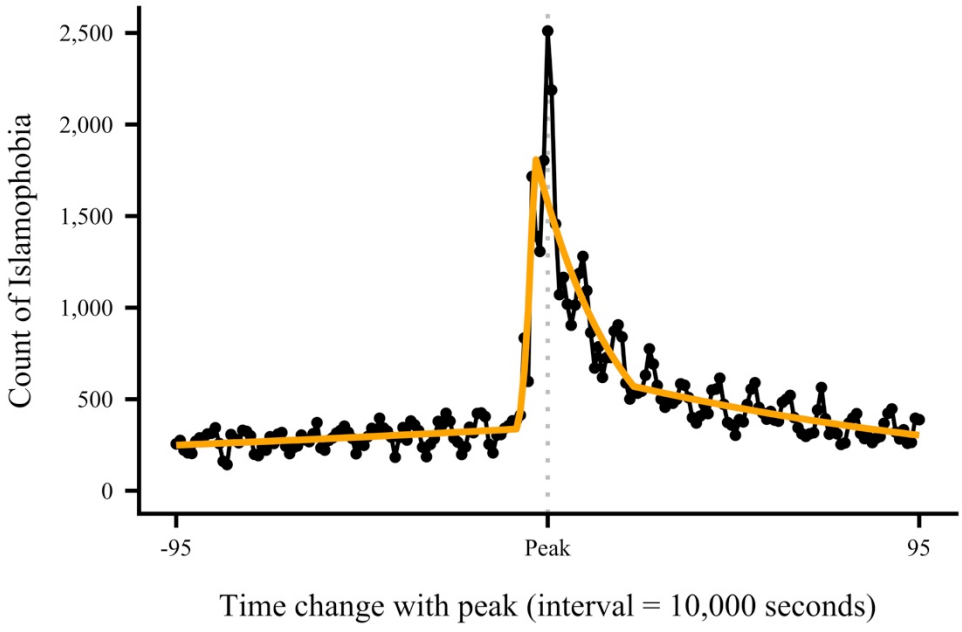


Figure 7-17, Negative binomial segmented regression model with the Parsons Green attack removed, time granularity of 10,000 seconds

Statistic	Model 7-7 NB model (10,000 seconds) All four attacks	Model 7-8 NB model (10,00 seconds) Parsons Green attack removed
Estimated breakpoints	-10, 0, 30	-10, 0, 30
y-intercept	331 *****	346 *****
Slope 1	1.002 *	1.0035 *****
Slope 2	1.524 *****	1.496 *****
Slope 3	0.965 *****	0.9548 *****
Slope 4	0.993 *****	0.991 *****
Change vs. slope 1	1.520	1.490
Change vs. slope 2	0.633	0.637
Change vs. slope 3	1.029	1.038
Breakpoint 1	-7.207 *****	-7.636 *****
Breakpoint 2	-3.698 *****	-3.434 *****
Breakpoint 3	29.556 *****	21.999 *****
Convergence	65 iterations	64 iterations
Pseudo r-squared (Cox Snell)	0.529	0.668
Pseudo r-squared (Nagelkerke)	0.529	0.668
Pseudo r-squared (Pearson)	0.42	0.538
Dispersion parameter	3.6483	4.1331

Table 7-8, Summary of negative binomial segmented regression models with the Parsons Green attack removed, time granularity of 10,000 seconds

7.2.3.5 | Negative binomial segmented regression model fitting

In this section I report on the residuals for the negative binomial segmented regression for all four attacks with time granularity of 10,000 seconds. Figure 7-18 shows the residual plots and Figure 7-19 shows the distribution of residuals. The residuals plot shows that the segmented regression models struggle to fully capture the big spike in Islamophobia immediately following the terrorist attacks.

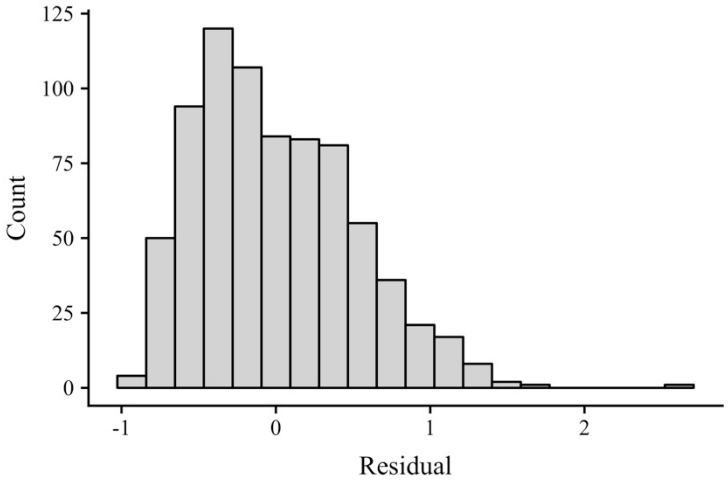


Figure 7-18, Distribution of residuals in negative binomial segmented regression model for all four attacks with time granularity of 10,000 seconds

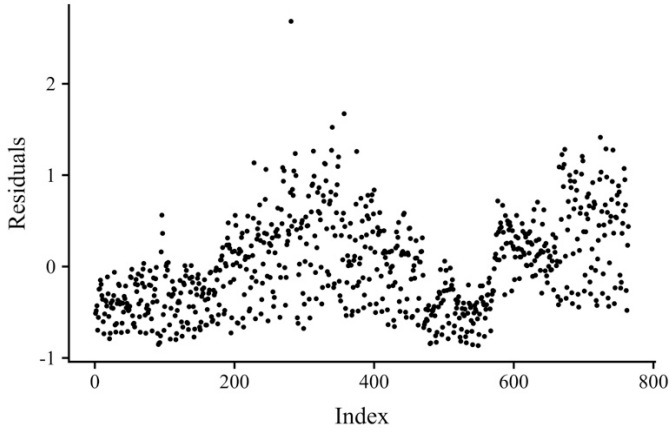


Figure 7-19, Residuals plot of negative binomial segmented regression model for all four attacks with time granularity of 10,000 seconds

7.2.3.6 | Negative binomial segmented regression with different numbers of breakpoints

In this section I fit negative binomial segmented regression models with 3 to 6 breakpoints. The purpose of this is to demonstrate how fit improves with more breakpoints but at the risk of reducing model parsimony and generalisability. The models are shown in Figure 7-20 and indicate that there is a clear risk of overfitting with more breakpoints – with little improvement in fit. In particular, 5 and 6 breakpoints show potentially spurious fits; the shape of the lines are very similar to the models with both 3 and 4 breakpoints – but with considerably greater complexity.

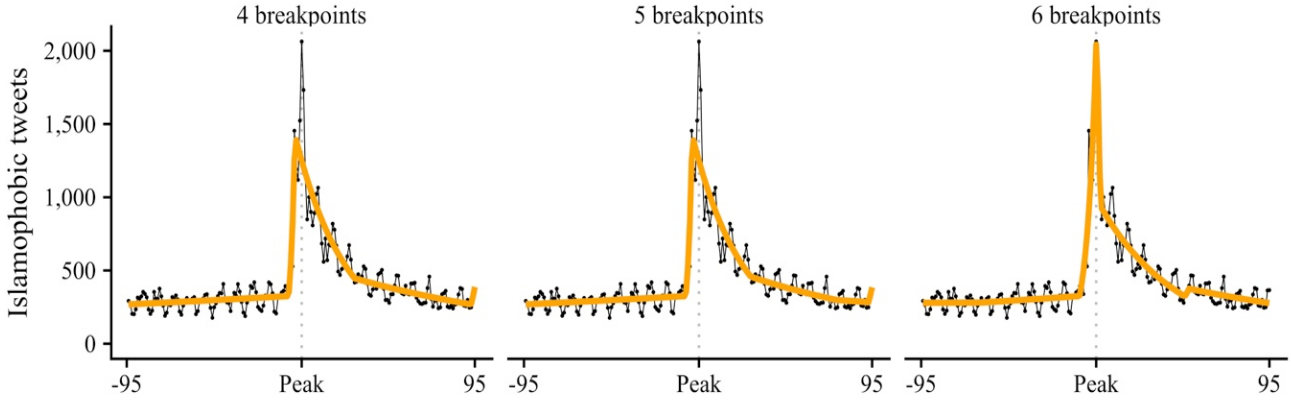


Figure 7-20, Negative binomial segmented regression models with different numbers of breakpoints

7.2.4 | Party differences

7.2.4.1 | Set-scale graph for party differences in Islamophobia following terrorist attacks

Figure 7-21 has a set-scale to highlight differences in the overall volume of Islamophobic tweets.

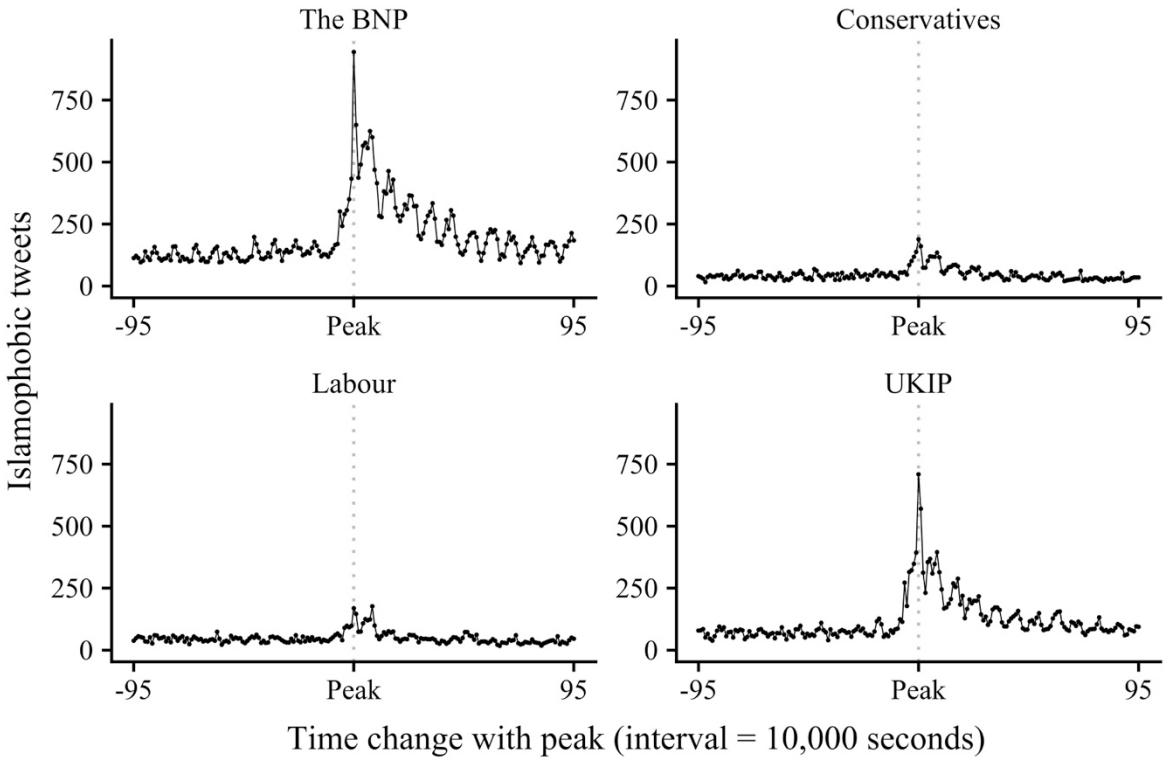


Figure 7-21, Party differences in the number of Islamophobic tweets around Islamist terrorist attacks (set-scale)

7.2.4.2 | Negative binomial segmented regression model with party differences, fitting

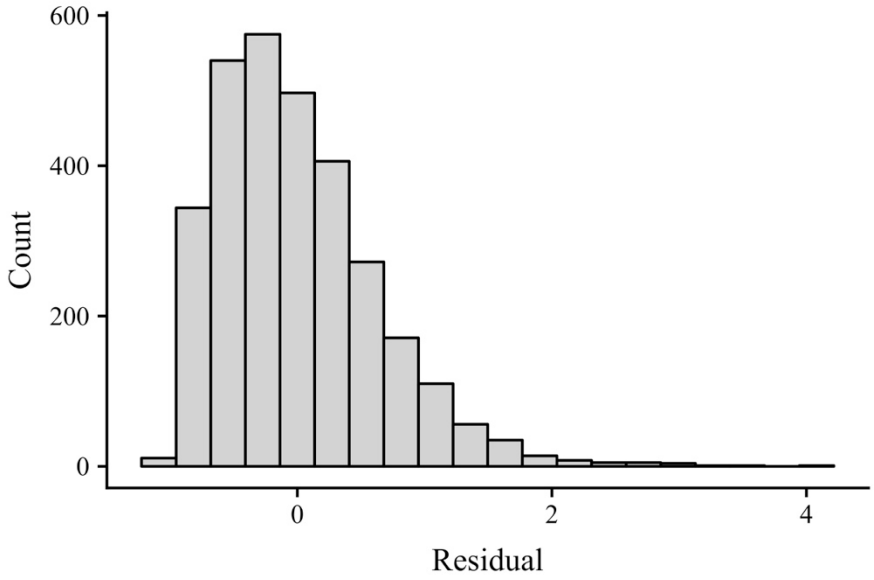


Figure 7-22, Distribution of residuals in negative binomial segmented regression model for all four attacks with time granularity of 10,000 seconds

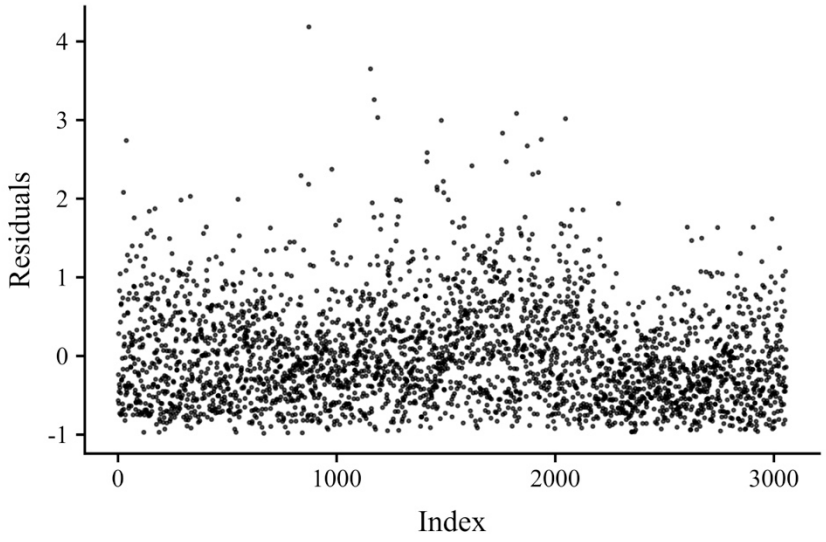


Figure 7-23, Residuals plot of negative binomial segmented regression model for all four attacks with time granularity of 10,000 seconds

7.2.5 | Significance testing for user behaviour following Islamist terrorist attacks

7.2.5.1 | Each attack modelled separately

For each day during the period studied, excluding the days for the terrorist attacks, I take the count of one-off Islamophobic tweeters. The distribution is well approximated by a discrete power law, with a minimum value of 6 and a scaling factor of 4.14. These values are estimated using the bootstrapping procedure outlined by Clauset et al. (Clauset, Shalizi, & Newman, 2009), implemented in R through the 'PoweRLaw' package (Gillespie, 2015). Testing also indicates that a power law distribution is a better fit than an exponential, Poisson or log-normal distribution. The distribution is plotted with logarithmic axes in Figure 7-24.

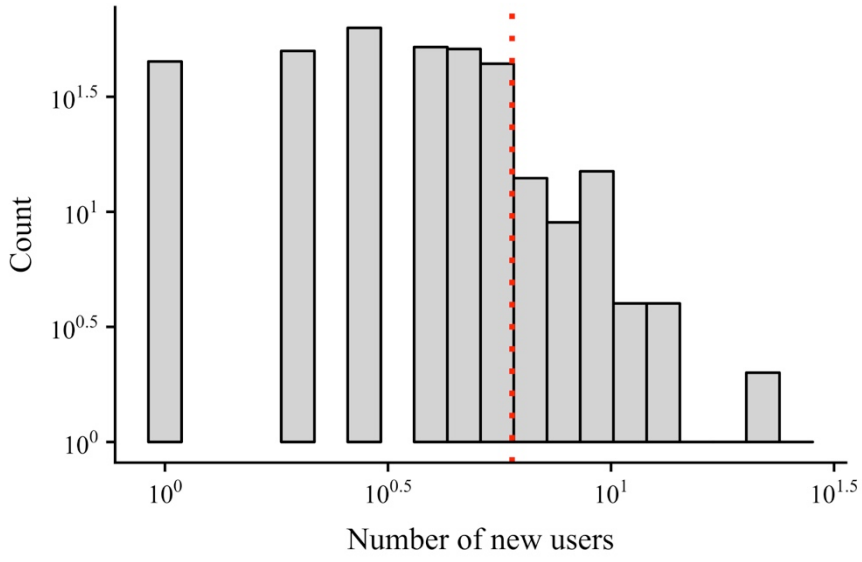


Figure 7-24, Number of one-off Islamophobic users each day

I compare the count of one-off Islamophobic tweeters for the four terrorist attacks with the other days and calculate the probability that they would each be generated by the distribution of counts. Significance values are estimated using a bootstrapping procedure, the results of which are shown in Table 7-9. The values are significant for the first three

of the attacks (which all involved the murder of citizens) but not the final one, Parsons Green.

Attack	Number of users	Significance
Westminster	23	$p < 0.0000001$
Manchester Arena	53	$p < 0.0000001$
London Bridge	76	$p < 0.0000001$
Parsons Green	5	N.S.

Table 7-9, Count of users who only tweet Islamophobically during the terrorist attack with p-values

7.2.5.2 | Attacks modelled together (over four days)

I take every combination of four days in the dataset (i.e. every combination without considering the order of days), less the four days for the terrorist attacks (which creates a dataset of 361 days). This creates 27,441,715 unique 4-day combinations. This is infeasible to model and as such I only take combinations from 1st December 2017 to 28th February 2018. As noted elsewhere, this can be understood as a period during which no Islamist terrorist attacks occur and the rate of Islamophobia is broadly stable. During these 90 days there are 2,555,190 unique 4-day combinations. The average count of one-off Islamophobic tweeters during other 4-day periods is 14. Appropriate statistical testing indicates the distribution of users who only tweet during each of these combinations is well-approximated by a power law. Note that whilst I only study 90 days, I compare the count of users who tweet on these days with the whole period of data to ensure a fair comparison. That is, for a user who tweets Islamophobically during a given 4-day combination to have *only* tweeted Islamophobically during that combination they must not send any Islamophobic tweets during any other point during the whole year, and not just the 90 day period I focus on. Statistical testing indicates that the observed count of

users who only tweet on the four days for the Islamist terrorist attacks (n = 157) is extremely significant, with a p-value equivalent to 0.

Appendix 7.3 | Chapter 7 Data overview

7.3.1 | User sampling

Most UK political parties, particularly those with elected representatives in national or supra-national bodies, have a very large number of followers on Twitter. I collect the number of followers for each party at the start of the period (1st March 2017), as shown in Table 7-10. Monitoring all of the tweets produced by this many users is infeasible, and accordingly followers must be sampled for each party. A sample size of 7,500 users is chosen based on relevant power tests. A balanced one-way ANOVA power test for just four samples of 7,500 users, with alpha of 0.001 and an effect size of 0.1, has power indistinguishable from 1 (Cohen, 1992). This means that it is highly likely that any true differences will be captured. It also suggests that for the full set of four sub-samples, even if a non-parametric significance test is used (which tend be less powerful), power will be sufficiently high (Mumby, 2002).

Party	Number of followers on 1st March 2017	Size of sub-sample
UKIP	153,623	7,500
Conservatives	236,306	7,500
Labour	502,465	7,500
BNP	12,895	7,500

Table 7-10, Number of followers for each political party

For the UKIP, Conservative and Labour sub-samples users are sampled based on the followership distribution – that is, the distribution of how many followers each follower of the party has. Previous research indicates that the number of followers is linked to other behavioural features of interest, such as the volume of tweets users produce and the

number of times their tweets are retweeted, which makes it an appropriate sampling variable (Bakshy et al., 2011; Cha, Haddai, Benevenuto, & Gummadi, 2010; Halberstam & Knight, 2016). The followership distribution for Twitter users is typically long-tailed (Muchnik et al., 2013; Riquelme & González-Cantergiani, 2016). In many cases this can be well-approximated with a power law, although in some cases it is necessary to use other non-normal distributions, such as the log-normal or exponential (Clauset et al., 2009; Virkar & Clauset, 2014).

Sampling from long-tailed distributions poses additional challenges compared with normally distributed data, particularly when the sample size is relatively small compared to the population. Sampling methods which are appropriate for normally distributed data, such as Random Node, Random Edge and Random Node-Edge, are unsuitable as they are likely to introduce biases for two reasons (Gjoka, Butts, Kurant, & Markopoulou, 2011). First, the sample is likely to be severely truncated whereby values at the edge of the distribution are not included. Second, the distribution of the sample is likely to be a poor approximation of the underlying distribution as certain parts are under- or over-sampled (Leskovec & Faloutsos, 2006; Pickering, Bull, & Sanderson, 1995). Accordingly, traditional random sampling methods will not necessarily produce a representative sample and any inferences may be invalid.

A further challenge is that the large number of followers, and followers of followers, for UKIP, the Conservatives and Labour means that it is infeasible to collect each party's full followership network before sampling. Therefore, at the point of sampling the true followership distribution is unknown. In such situations many widely-used sampling methods cannot be implemented and alternatives, such as those based on the random walk, must be used instead (Sarma, Nanongkai, Pandurangan, & Tetali, 2009). Drawing on the work of Leskovec and Faloutsos I implement random jump sampling as this is

likely to produce a representative sampling and is computationally efficient (Leskovec & Faloutsos, 2006). In this method, a starting node is selected at random. A random walk is then simulated from this node across the followership network (with the starting party excluded from the sub-sample) with a 0.15 probability at each step of jumping to a new randomly selected node. The jump probability is an extension on the random walk which reduces the chance of ending up in a 'sink'. This is where the starting node has few connections, which is a risk in sparsely connected graphs such as social media networks – the jump strategy also ensures that nodes which are only connected to the starting party are better represented. Nodes are sampled without replacement until the full sample size has been collected.

To ensure that I can disambiguate how followers of different parties vary I only include users in each sub-sample if they do *not* follow the other accounts in the chapter. For example, users can only be included in the Conservatives sub-sample if they do not follow UKIP, Labour or the BNP.

7.3.2 | Active users

During the period, many of the followers do not send any tweets. In line with the previous chapter, I term followers who send tweets 'active users', and these form the basis of this analysis. The number of users who are active followers for each party is shown in Table 7-11. There are noticeable differences, from Labour with 5,078 active users to the Conservatives, with just 4,142 (a difference of 936 users).

Party	Size of sub-sample	Number of active users	Percentage of users in sub-sample which are active
UKIP	7,500	4,287	57.2%
Conservatives	7,500	4,142	55.2%
Labour	7,500	5,078	67.7%

BNP	7,500	4,319	57.6%

Table 7-11, Number of active followers for each party

7.3.3 | Persistent followers

For each party, some users cease following during the data collection period. Similarly, for each party some users start following one of the other parties during the period. In line with the previous chapter, I call users who follow the same party over the whole period, 'persistent followers'. I only retain these users in the dataset. The relatively large number of users who are *not* persistent followers (approximately 88% of the dataset) is driven by users starting to follow one of the other parties, in particular Labour, which had a noticeable upsurge in its number of social media followers during 2017. The number of active persistent followers for each party is shown in Table 7-12.

Party	Number of active followers	Number of active persistent followers	Percentage of active followers which are persistent
UKIP	4,287	3,606	84.1%
Conservatives	4,142	3,430	82.8%
Labour	5,089	4,759	93.5%
BNP	4,319	3,891	90%

Table 7-12, Number of active persistent followers for each party

7.3.4 | Language

For each party, many users send tweets in languages other than English. I only keep tweets in 'English' and 'Undetermined' as the classifier developed in Chapter 5 is tuned only for English language. Removing just non-English *tweets* ensures that users who produce a mix of both English and non-English language tweets are kept in the dataset. The impact of this on the number of followers for each party is minimal, as shown in

Table 7-13. It should be noted that this has a larger impact on the overall number of tweets in each dataset, as some followers send tweet in multiple languages.

Party	Number of active persistent followers	Number, considering only tweets in English and Undetermined	Percentage retained
UKIP	3,606	3,541	98.2%
Conservatives	3,430	3,362	98.0%
Labour	4,759	4,707	98.9%
BNP	3,891	3,795	97.5%

Table 7-13, Number of active persistent followers for each party, considering only tweets in English and Undetermined

7.3.5 | Bots

In line with the arguments made in Chapter 6, and with the work of Kollanyi et al. (Kollanyi et al., 2016), I implement a rule-based system, removing accounts which tweet more than 40 times per day (14,600 in total during the period of data collection). As previously discussed, this is a crude but effective approach in which highly active accounts are removed – irrespective of whether they are fully automated bots, semi-automated accounts or hyper-active genuine users. The number of users removed under this criterion is shown in Table 7-14. Interestingly, there are some noticeable discrepancies, which could point to the potential use of botnets by different political parties. 67 followers of the BNP are removed compared with just 16 followers of the Conservatives.

Party	Number of active persistent followers, considering only tweets in English and Undetermined	Number after highly active accounts are removed	Number of highly active accounts removed	Average number of tweets per highly active account
UKIP	3,541	3,497	44	26,095
Conservatives	3,362	3,346	16	24,794
Labour	4,707	4,683	24	20,312
BNP	3,795	3,727	68	23,432

Table 7-14, Number of active persistent followers for each party, considering only tweets in English and Undetermined, after the removal of highly active users

7.3.6 | Data summary

Party	Number of followers	Total Number of tweets	Average number of tweets per user
UKIP	3,497	2,691,105	770
Conservatives	3,346	2,135,850	638
Labour	4,683	3,167,564	676
BNP	3,727	3,149,468	845
TOTAL	15,253	11,143,987	731

Table 7-15, Summary of final dataset for followers of each party

References

Abbas, T. (2004). After 9/11: British South Asian Muslims, Islamophobia, Multiculturalism, and the State. *The American Journal of Islamic Social Sciences*, *21*(3), 26–38.

Adorno, T. (1950). *The Authoritarian Personality*. New York: Harper & Brothers.

Agarwal, B. (2014). Personality Detection from Text: A Review. *International Journal of Computer Systems*, *1*(1), 1–4.

Aguilera-Carnerero, C., & Azeez, A. H. (2016). 'Islamonausea, not Islamophobia': The many faces of cyber hate speech. *Journal of Arab & Muslim Media Research*, *9*(1), 21–40. Retrieved from doi: https://doi.org/10.1386/jammr.9.1.21_1

Ahmed, W., Bath, P., & Demartini, G. (2017). Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In K. Woodfield (Ed.), *The Ethics of Online Research. Advances in Research Ethics and Integrity* (pp. 79–107). London: Emerald Publishing. https://doi.org/10.1016/j.jocn.2005.03.017

Aichholzer, J., & Zandonella, M. (2016). Psychological bases of support for radical right parties. *Personality and individual differences, 96(1),* 185-190.

Akkerman, T., & Rooduijn, M. (2015). Pariahs or Partners? Inclusion and Exclusion of Radical Right Parties and the Effects on Their Policy Positions. *Political Studies*, *63*(5), 1140–1157. https://doi.org/10.1111/1467-9248.12146

Akoka, J., Comyn-wattiau, I., & Laou, N. (2017). Research on Big Data – A systematic mapping study. *Computer Standards & Interfaces*, *54*(1), 105–115. https://doi.org/10.1016/j.csi.2017.01.004

All Party Parliamentary Group on British Muslims. (2018). *Islamophobia defined: the*

inquiry into a working definition of Islamophobia. London.

Allen, C. (1996). What's wrong with the "golden rule"? Conundrums of conducting ethical research in cyberspace. *Information Society*, *12*(2), 175–188. https://doi.org/10.1080/713856146

Allen, C. (2010a). Fear and loathing: the political discourse in relation to Muslims and Islam in the British contemporary setting. *Contemporary British Religion and Politics*, *329*(420), 221–236.

Allen, C. (2010b). *Islamophobia*. Surrey: Ashgate.

Allen, C. (2011). Opposing Islamification or promoting Islamophobia? Understanding the English Defence League. *Patterns of Prejudice*, *45*(4), 279–294. https://doi.org/10.1080/0031322X.2011.585014

Allen, C. (2013). Passing the dinner table test: Retrospective and prospective approaches to tackling Islamophobia in Britain. *SAGE Open*, *3*(2), 1–10. https://doi.org/10.1177/2158244013484734

Allen, C. (2017). Islamophobia and the Problematization of Mosques: A Critical Exploration of Hate Crimes and the Symbolic Function of "Old" and "New" Mosques in the United Kingdom. *Journal of Muslim Minority Affairs*, *37*(3), 294–308. https://doi.org/10.1080/13602004.2017.1388477

Allen, T. J. (2017). All in the party family? Comparing far right voters in Western and Post-Communist Europe. *Party Politics*, *23*(3), 274–285. https://doi.org/10.1177/1354068815593457

Allport, G. (1954). *The nature of prejudice*. Oxford: Addison-Wesley.

Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). The Enemy Among Us:

Detecting Hate Speech with Threats Based "Othering" Language Embeddings. https://doi.org/arXiv:1801.07495v3

Altemeyer, B. (1998). The Other "Authoritarian Personality." *Advances in Experimental Social Psychology*, *36*(1), 47–92.

Alvares, C., & Dahlgren, P. (2016). Populism, extremism and media: Mapping an uncertain terrain. *European Journal of Communication*, *31*(1), 46–57. https://doi.org/10.1177/0267323115614485

Amiri, M., Hashemi, M. R., & Rezaei, J. (2015). The representation of Islamophobia: A critical discourse analysis of Yahoo news. *International Journal of Control Theory and Applications*, *8*(2), 599–618. https://doi.org/10.17485/ijst/2015/v8i28/87385

Amnesty. (2017). Tackling hate crime in the UK: a background briefing paper. London: Amnesty International.

Anand, D. (2010). Generating Islamophobia in India. In S. Sayyid & A. Vakil (Eds.), *Thinking through Islamophobia: global perspectives*. London: C. Hurst.

Anastasiades, G., & Mcsharry, P. E. (2014). Extreme value analysis for estimating 50 year return wind speeds from reanalysis data. *Wind Energy*, *17*(1), 1231–1245. https://doi.org/10.1002/we

Anderson, C. (2008, June 23). The end of theory: the data deluge makes the scientific method obselete. *Wired*. Retrieved from https://www.wired.com/2008/06/pb-theory/

Anstead, N., & O'Loughlin, B. (2011). The emerging viewertariat and BBC question time: Television debate and real-time commenting online. *International Journal of Press/Politics*, *16*(4), 440–462. https://doi.org/10.1177/1940161211415519

Archer, L. (2009). Race, Face and Masculinity: the identities and local geographies of Muslim boys. In P. Hopkins & R. Gale (Eds.), *Muslims in Britain: Race, Place and Identities: Race, Place and Identities*. Edinburgh: Edinburgh University Press.

Asad, T., Butler, J., & Mahmood, S. (2009). *Is Critique Secular? Blasphemy, Injury, and Free Speech*. Berkeley: The University of California Press. https://doi.org/10.1525/california/9780982329412.001.0001

Atton, C. (2006). Far-right media on the internet: Culture, discourse and power. *New Media and Society*, *8*(4), 573–587. https://doi.org/10.1177/1461444806065653

Austin, J. (1962). *How to do things with words*. Oxford: Clarendon Press.

Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against muslims on social media. *Policy and Internet*, *6*(2), 133–150. https://doi.org/10.1002/1944-2866.POI364

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1–20. https://doi.org/10.5281/zenodo.58517

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behaviour*, *27*(1), 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Awan, I., & Zempi, I. (2017). "I will blow your face off" - Virtual and physical world anti-muslim hate crime. *British Journal of Criminology*, *57*(2), 362–380. https://doi.org/10.1093/bjc/azv122

Backman, K., & Kyngäs, H. A. (1999). Challenges of the grounded theory approach to a novice researcher. *Nursing and Health Sciences*, *1*(1), 147–153.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *ArXiv:1706.00188v1*, 1–2. https://doi.org/10.1145/3041021.3054223

Bahdi, R., & Kanji, A. (2018). What is Islamophobia? *University of New Brunswick Law Journal*, *69*(11), 324–363.

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1–22. https://doi.org/10.1002/jae.659

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on Twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM '11*, 1–10. https://doi.org/10.1145/1935826.1935845

Baldini, G., Bressanelli, E., & Massetti, E. (2018). Who is in control? Brexit and the Westminster Model. *The Political Quarterly*, *89*(4), 537–544. https://doi.org/10.1111/1467-923X.12596

Bale, T. (2018). Who leads and who follows? The symbiotic relationship between UKIP and the Conservatives – and populism and Euroscepticism. *Politics*, *38*(3), 263–277. https://doi.org/10.1177/0263395718754718

Balibar, E. (1991). Is there a neo-Racism? In E. Balibar & E. Wallerstein (Eds.), *Race, Nation and Class: ambiguous identities* (pp. 17–28). London: Verso.

Banton, M. (2015). Problem finding in ethnic and racial studies. In *Theories of race and ethnicity*. Cambridge: Cambridge University Press.

Barker, M. (1981). *The New Racism*. London: Junction Books.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). *Don't count, predict! A systematic comparison of context-counting vs . context-predicting semantic vectors*.

Bartlett, J., & Littler, M. (2011). *Inside the EDL. Populist politics in a digital age.* London: DEMOS. Retrieved from http://www.demos.co.uk/publications/insidetheedl

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2010). An overview of latent Markov models for longitudinal categorical data. *ArXiv:1003.2804v1*, 1–36. https://doi.org/10.1007/s40300-014-0035-2

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2015). LMest: an R package for latent Markov models for categorical longitudinal data. *Journal of Statistical Software*, *81*(4), 1–38. https://doi.org/10.18637/jss.v081.i04

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397–1410. https://doi.org/10.1002/asi

Bayrakli, E., & Hafez, F. (2016). *European Islamophobia Report 2015*. Istanbul: SETA. https://doi.org/978-605-4023-68-4

Bayrakli, E., & Hafez, F. (2017). *European Islamophobia Report 2016*. Istanbul: SETA.

Bayrakli, E., & Hafez, F. (2018). *European Islamophobia Report 2017*. Istanbul: SETA.

BBC. (2014, May 7). Nigel Farage: UKIP is not a racist party. *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-politics-27315328

BBC. (2017, June 19). What is Islamophobia? *BBC News*.

BBC. (2018a, January 30). Labour suspends councillor in Sarwar "Islamophobia" row.

BBC News. Retrieved from https://www.bbc.co.uk/news/uk-scotland-scotland-politics-42880755

BBC. (2018b, February 2). Finsbury Park attacker Darren Osborne jailed for minimum of 43 years. *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-42920929

BBC. (2018c, March 21). Cambridge Analytica: The story so far. *BBC News*. https://doi.org/10.1007/s13277-013-1336-4

BBC. (2018d, November 23). UKIP leader defends hiring Tommy Robinson. *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-politics-46308160

BBC. (2018e, December 27). Tech became "darker and more muddy" in 2018. *BBC News*. Retrieved from https://www.bbc.co.uk/news/technology-46675680%0A

Belinski, D. (2001). *The Advent of the Algorithm: The 300-Year Journey from an Idea to the Computer*. New York: Harcourt.

Bell, A., Fairbrother, M., & Jones, K. (2018). Fixed and random effects models: making an informed choice. *Quality and Quantity*, *60*(1), 1–24. https://doi.org/10.1007/s11135-018-0802-x

Belli, L., & Zingales, Ni. (2017). *Platform regulations: how platforms are regulated and how they regulate us* (Vol. December). Geneva.

Benford, R., & Snow, D. (2000). Framing processes and social movements: an overvie wand assessment. *Annual review of sociology, 26(1)*, 611—639.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology*, *609*(1), 223–239. https://doi.org/10.1007/978-1-60327-241-4_13

Benesch, S. (2012). *Dangerous Speech: A Proposal to Prevent Group Violence*.

Washington.

Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or Hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13. https://doi.org/10.1145/380995.380999

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced Text Analysis: Reproducible and Agile Production. *American Political Science Review*, *110*(2), 278–295. https://doi.org/10.1017/S0003055416000058

Benoit, K., & Matsuo, A. (2018). *R Package: 'spacyr.'* London.

Berger, J. M., Strathearn, B., & Meleagrou-hitchens, A. (2013). *Who Matters Online: Measuring influence, evaluating content and countering violent extremism in online social networks*. London.

Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, *46*(1), 348–355. https://doi.org/10.1093/ije/dyw098

Berry, D. M. (2011). The computatonal turn: Thinking about the digital humanities. *Culture Machine*, *12*(1), 1–22. https://doi.org/10.1007/s12599-014-0342-4

Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science*, *340*(6139), 1438–1441. https://doi.org/10.1126/science.1235823

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, *104*(17), 7301–7306. https://doi.org/10.1073/pnas.0610172104

Biggs, M., & Knauss, S. (2012). Explaining membership in the British National Party: A

multilevel analysis of contact and threat. *European Sociological Review*, *28*(5), 633–646. https://doi.org/10.1093/esr/jcr031

Bilge, S. (2010). Beyond subordination vs. resistance: An intersectional approach to the agency of veiled Muslim women. *Journal of Intercultural Studies*, *31*(1), 9–28. https://doi.org/10.1080/07256860903477662

Biran, O., & McKeown, K. (2017). Human-centric justification of machine learning predictions. *IJCAI International Joint Conference on Artificial Intelligence*, 1461–1467. https://doi.org/10.24963/ijcai.2017/202

Birt, J. (2009). Islamophobia in the construction of British Muslim identity politics. In P. Hopkins & R. Gale (Eds.), *Muslims in Britain: race, place and identities* (pp. 210–227). Edinburgh: Edinburgh University Press.

Bischof, D. (2017). Towards a renewal of the niche party concept: Parties, market shares and condensed offers. *Party Politics*, *23*(3), 220–235. https://doi.org/10.1177/1354068815588259

Blank, G. (2017). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, *35*(6), 679–697. https://doi.org/10.1177/0894439316671698

Bleich, E. (2011). What is Islamophobia and how much is there? theorizing and measuring an emerging comparative concept. *American Behavioral Scientist*, *55*(12), 1581–1600. https://doi.org/10.1177/0002764211409387

Blok, A., & Pedersen, M. A. (2014). Complementary social science? Quali-quantitative experiments in a Big Data world. *Big Data & Society*, *1*(2), 1–6. https://doi.org/10.1177/2053951714543908

Boellstorff, T. (2012). Rethinking digital anthropology. In H. A. Horst & D. Miller (Eds.),

Digital anthropology (pp. 39–60). New York: Berg Publishers. https://doi.org/10.1093/obo/9780199766567-0087

Bohannon, J. (2011). Human subject research. Social science for pennies. *Science*, *334*(6054), 307. https://doi.org/10.1126/science.334.6054.307

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606v2*, 1–12. https://doi.org/1511.09249v1

Bokde, N., Asencio-Cortés, G., Martínez-Álvarez, F., & Kulat, K. (2016). PSF: Introduction to R Package for Pattern Sequence Based Forecasting Algorithm. *The R Journal*, *1*(1), 1–10. https://doi.org/10.1007/s11605-003-0035-7

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 1–9. https://doi.org/10.1038/nature11421.A

Bonnefoy, L. (2004). *Modalities and limits of stigmatisation: French public institutions and Muslims after 11th September 2001*. New York.

Borell, K. (2015). When Is the Time to Hate ? A Research Review on the Impact of Dramatic Events on Islamophobia and Islamophobic Hate Crimes in Europe. *Islam and Christian-Muslim Relations*, *26*(4), 409–421. https://doi.org/10.1080/09596410.2015.1067063

Boulesteix, A., & Schmid, M. (2014). Discussion Machine learning versus statistical modeling. *Biometrical Journal*, *56*(4), 588–593. https://doi.org/10.1002/bimj.201300226

Bowen, J. (2005). Commentary on Bunzl. *American Ethnologist*, *32*(4), 524–525.

Bowman-Grieve, L. (2009). Exploring Stormfront: A virtual community of the radical right. *Studies in Conflict and Terrorism*, *32*(11), 989–1007. https://doi.org/10.1080/10576100903259951

Bowyer, B. (2008). Local context and extreme right support in England: The British National Party in the 2002 and 2003 local elections. *Electoral Studies*, *27*(4), 611–620. https://doi.org/10.1016/j.electstud.2008.05.001

boyd, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

boyd, & Crawford, K. (2015). *Six provocations for big data*. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society on 21 September 2011*. Oxford. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=101596204&site=ehost-live

Bravo López, F. (2017). Völkisch versus Catholic Islamophobia in Spain: The conflict between racial and religious understandings of Muslim identity. *Revista de Estudios Internacionales Mediterraneos*, *22*(22), 141–164. https://doi.org/10.15366/reim2017.22.007

Breen-smyth, M. (2014). Critical Studies on Terrorism Theorising the "suspect community": counterterrorism , security practices and the public imagination. *Critical Studies on Terrorism*, *7*(2), 223–240. https://doi.org/10.1080/17539153.2013.867714

Bridgman, P. W. (1938). Operational analysis. *Philosophy of Science*, *5*(1), 114–131.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third*

conference on Applied natural language processing (pp. 152–155). https://doi.org/10.3115/974499.974526

British Sociological Association. (2017). *Ethics Guidelines and Collated Resources for Digital Research*. London.

Buttler, J. (1988). Performative acts and gender constitution: an essay in phenomenology and theory. *Theatre Journal*, *40(4)*, 519-531.

Butler, J., Laclau, E., & Žižek, S. (2001). *Contingency, Hegemony and Universality*. London: Verso.

Harris, B. (2002). Xenophobia: a new pathology for a new South Africa ? In D. Hook & G. Eagle (Eds.), *Psychopathology and social prejudice* (pp. 169–184). Cape Town: University of Cape Town Press.

Brown, G. (1995). Why the BNP is Still Fascist. *Spotlight*, 29–32.

Brown, R. (2010). *Prejudice: Its social psychology*. Sussex: Wiley-Blackwell.

Brubaker, R. (2010). In the name of the nation: reflections on nationalism and patriotism. *Citizenship Studies*, *8*(2), 115–127. https://doi.org/10.1080/1362102042000214705

Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology*, *4*(1), 217–231.

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91–108. https://doi.org/10.1080/13645579.2012.756095

Buchanan, E. (2017). Considering the ethics of big data research: A case of Twitter and ISIS/ISIL. *PLoS ONE*, *12*(12), 1–6. https://doi.org/10.1371/journal.pone.0187155

Bunzl, M. (2005). Between anti-Semitism and Islamophobia: *American Ethnologist*, *32*(4), 499–508.

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., … Sloan, L. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*, *95*(1), 96–108. https://doi.org/10.1016/j.techfore.2013.04.013

Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(11), 2–15. https://doi.org/10.1140/epjds/s13688-016-0072-6

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., & Knight, V. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Annals*, *4*(206), 1–14. https://doi.org/10.1007/s13278-014-0206-4

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 205395171454028. https://doi.org/10.1177/2053951714540280

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *Annals of the American Academy of Political and Social Science*, *659*(1), 122–131. https://doi.org/10.1177/0002716215569441

Busher, J., & Macklin, G. (2015). Interpreting “Cumulative Extremism”: Six proposals

for enhancing conceptual clarity. *Terrorism and Political Violence*, *27*(5), 884–905. https://doi.org/10.1080/09546553.2013.870556

Byers, B. D., & Jones, J. A. (2007). The Impact of the Terrorist Attacks of 9/11 on Anti-Islamic Hate Crime. *Journal of Ethnicity in Criminal Justice*, *5*(1), 43–56. https://doi.org/10.1300/J222v05n01

Bzdok, D., & Meyer-lindenberg, A. (2018). Review Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Caiani, M., della Porta, D., & Wagemann, C. (2012). *Mobilizing on the Extreme Right*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199641260.001.0001

Caiani, M., & Kröll, P. (2015). The transnationalization of the extreme right and the use of the Internet. *International Journal of Comparative and Applied Criminal Justice*, *39*(4), 331–351. https://doi.org/10.1080/01924036.2014.973050

Caiani, M., & Wagemann, C. (2009). Online networks of the Italian and German Extreme right. *Information Communication and Society*, *12*(1), 66–109. https://doi.org/10.1080/13691180802158482

Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, *22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

Candia, C., Rodriguez-sickert, C., Barabási, A., & Hidalgo, C. A. (2019). The universal decay of collective memory and attention. *Nature Human Behaviour*, *3*(January), 82–91. https://doi.org/10.1038/s41562-018-0474-5

Canovan, M. (2002). Taking Politics to the People: Populism as the Ideology of Democracy. In S. Mény Y. (Ed.), *Democracies and the populist Challenge* (pp. 25–44). London: Palgrave Macmillan UK.

Cardiff University. (2018). Rise in Brexit related Hate Crime to be Focus of New Research Lab. Retrieved December 29, 2018, from https://www.cardiff.ac.uk/news/view/1393983-rise-in-brexit-related-hate-crime-to-be-focus-of-new-research-lab

Caren, N., Jowers, K., & Gaby, S. (2012). A Social Movement Online Community: Stormfront and the White Nationalist Movement. In J. Earl & D. A. Rohlinger (Eds.), *Media, Movements, and Political Change (Research in Social Movements)*. London: Emerald Publishing.

Carrillo, H., Brodersen, K. H., & Castellanos, J. (2014). Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy. In *ROBOT2013: First Iberian Robotics Conference, Advances in Intelligent Systems and Computing 252* (pp. 347–368). https://doi.org/10.1007/978-3-319-03413-3

Carter, A. J. (2017). Cumulative extremism: escalation of movement–countermovement dynamics in Northern Ireland between 1967 and 1972. *Behavioral Sciences of Terrorism and Political Aggression*, *9*(1), 37–51. https://doi.org/10.1080/19434472.2016.1236830

Carter, E. (2018). Right-wing extremism/radicalism: reconstructing the concept. *Journal of Political Ideologies*, *23*(2), 1–26. https://doi.org/10.1080/13569317.2018.1451227

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/S1745855207005029

Castelli Gattinara, P., & Pirro, A. L. P. (2018). The far right as social movement. *European Societies*, *0*(0), 1–16. https://doi.org/10.1080/14616696.2018.1494301

Castells, M. (2015). *Networks of outrage and hope*. London: Wiley-Blackwell.

Castles, S., & Davidson, A. (2000). *Citizenship and Migration: Globalization and the Politics of Belonging*. New York: Routledge.

Catalano, R., Novaco, R. W., & McConnell, W. (2002). Layoffs and Violence Revisited. *Aggressive Behavior*, *28*(3), 233–247. https://doi.org/10.1002/ab.80003

Centola, D., & Macy, M. (2007). Complex contagion and the weakness of long ties. *American Journal of Sociology*, *113*(3), 702–734.

Cha, M., Haddai, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring User Influence in Twitter : The Million Follower Fallacy. *International AAAI Conference on Weblogs and Social Media*, 10–17. https://doi.org/10.1.1.167.192

Chadwick, A., & Stromer-Galley, J. (2016). Digital Media, Power, and Democracy in Parties and Election Campaigns: Party Decline or Party Renewal? *International Journal of Press/Politics*, *21*(3), 283–293. https://doi.org/10.1177/1940161216646731

Chakraborti, N., & Garland, J. (2015). *Reconceptualising hate crime victimisation through the lens of vulnerability and 'difference'*. Birmingham.

Chandrasekharan, E., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E., & Pavalanathan, U. (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM Human-Computer Interactions*, *1*(2), 1–22. https://doi.org/10.1145/3134666

Cheng, J., Bernstein, M., Danescu-niculescu-mizil, C., & Leskovec, J. (2017). Anyone

Can Become a Troll : Causes of Trolling Behavior in Online Discussions. *ArXiv:1702.01119v1*, 1–14.

Cheung, C. M. K., Chiu, P., & Lee, M. K. O. (2011). Online social networks: Why do students use Facebook ? *Computers in Human Behavior*, *27*(4), 1337–1343. https://doi.org/10.1016/j.chb.2010.07.028

Christ, T. W. (2014). Scientific-Based Research and Randomized Controlled Trials, the "Gold" Standard? Alternative Paradigms and Mixed Methodologies. *Qualitative Inquiry*, *20*(1), 72–80. https://doi.org/10.1177/1077800413508523

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine*, *357*(1), 370–379.

Christakis, N. A., & Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. *The New England Journal of Medicine*, *358*(1), 2249–2258.

Ciampaglia, G. L., Flammini, A., & Menczer, F. (2015). The production of information in the attention economy. *Scientific Reports*, *5*, 1–6. https://doi.org/10.1038/srep09452

Cihon, P., & Yasseri, T. (2016). A Biased Review of Biases in Twitter Studies on Political Collective Action. *Frontiers in Physics*, *4*(1), 1–8. https://doi.org/10.3389/fphy.2016.00034

Clauset, A., Shalizi, C. R., & Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1055/s-2008-1043820

Cleen, B. De, & Stavrakakis, Y. (2017). Distinctions and Articulations: A Discourse Theoretical Framework for the Study of Populism and Nationalism. *Javnost: The Public*, *0*(0), 1–19. https://doi.org/10.1080/13183222.2017.1330083

CNet. (2018, September 6). Hey, Twitter and Facebook: Your wild west era's coming to an end. *CNet*.

Cockbain, E. (2013). Grooming and the "Asian sex gang predator": The construction of a racial crime threat. *Race and Class*, *54*(4), 22–32. https://doi.org/10.1177/0306396813475983

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1).

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cole, A. (2005). Old right or new right ? The ideological positioning of parties of the far right. *European Journal of Political Research*, *44*(1), 203–230.

Coles, A. B., & West, M. (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, *60*, 233–244. https://doi.org/10.1016/j.chb.2016.02.070

Collier, D., & Mahon, J. E. (1993). Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis. *The American Political Science Review*, *87*(4), 845–855.

Collingwood, L., & Wilkerson, J. (2011). Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, (4), 1–28.

Commerer, B. (2010). *Populist Radical Right Homophobia*. Portland.

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on Twitter. In *Association for the advancement of artificla intelligence: ICWSM* (Vol. 133, pp. 89–96). https://doi.org/10.1021/ja202932e

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., … Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal: Special Topics*, *214*(1), 325–346. https://doi.org/10.1140/epjst/e2012-01697-8

Copsey, N. (1994). Fascism: The Ideology of the British National Party. *Politics*, *14*(3), 101–108.

Copsey, N. (2007). Changing course or changing clothes? Reflections on the ideological evolution of the British National Party 1999-2006. *Patterns of Prejudice*, *41*(1), 61–82. https://doi.org/10.1080/00313220601118777

Copsey, N., Dack, J., Littler, M., & Feldman, M. (2013). Anti-Muslim Hate Crime and the Far Right. *Centre for Fascist, Anti-Fascist and Post-Fascist Studies*, 1–27.

Correa, T., Hinsley, A. W., & Zúñiga, H. G. De. (2010). Who interacts on the Web? The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247–253. https://doi.org/10.1016/j.chb.2009.09.003

Cowls, J., & Brown, I. (2015). *Check the Web*. Dublin: VOX-Pol.

Cowls, J., & Schroeder, R. (2015). Causation, Correlation, and Big Data in Social Science Research. *Policy and Internet*, *7*(4), 447–472. https://doi.org/10.1002/poi3.100

CPS. (2017). Hate Crime: public statement on prosecuting racist and religious hate crime. London: Crown Prosecution Service.

CPS. (2018). Social Media: guidelines on prosecuting cases involving communications sent via social media. London: Crown Prosecution Service.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*,

93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Crane, H. (2015). Clustering from Categorical Data Sequences. *Journal of the American Statistical Association*, *110*(510), 810–823. https://doi.org/10.1080/01621459.2014.983521

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum: Journal of Media and Cultural Studies*, *23*(4), 525–535. https://doi.org/10.1080/10304310903003270

Credibility Coalition. (2018). Credibility Coalition: About.

Crilley, R., & Gillespie, M. (2019). What to do about social media? Politics, populism and journalism. *Journalism*, *20*(1), 173–176. https://doi.org/10.1177/1464884918807344

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, *27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Croucher, S. (2011). Social networking and cultural adaption: a theoretical model. *Journal of International and intercultural communication, 4(4),* 259-264.

CSEW. (2018). CSEW: User guide to crime statistis for England and Wales 2018. London: Office for National Statistics.

Curran, P. J., Obeidat, K., & Losardo, D. (2011). Twelve frequently asked questions about Growth Curve Modelling. *Journal of Cognitive Development*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969.Twelve

Curthoys, A. (2013). *Identifying the Effect of Unemployment on Crime*. *Syracuse University Honors Program Capstone Projects* (Vol. May). https://doi.org/10.1086/320275

D'Orazio, V., Kenwick, M., Lane, M., Palmer, G., & Reitter, D. (2016). Crowdsourcing the measurement of interstate conflict. *PLoS ONE*, *11*(6), 1–21. https://doi.org/10.1371/journal.pone.0156527

Dadvar, M., Trieschnigg, D., & Jong, F. de. (2014). Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Advances in Artificial Intelligence. AI 2014. Lecture Notes in Computer Science (volume 8436)* (pp. 275–281). New York: Springer.

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. *ECIR 2013: Advances in Information Retrieval*, *LNCS*, 693–696. https://doi.org/10.1007/978-3-642-36973-5_62

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document Embedding with Paragraph Vectors. *ArXiv:1507.07998v1*, 1–8. Retrieved from http://arxiv.org/abs/1507.07998

Dancygier, R. (2017). The Left and Minority Representation: The Labour Party, Muslim Candidates, and Inclusion Tradeoffs. *Comparative Politics*, *46*(1), 1–21.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., … Chuang, I. (2014). Privacy, Anonymity, and Big Data in the Social Sciences. *Communications of the ACM*, *57*(9), 56–63.

DataRobot. (2018). DataRobot - about us. Retrieved October 5, 2018, from https://www.datarobot.com

Davenport, S. W., Bergman, S. M., Bergman, J. Z., & Fearrington, M. E. (2014). Twitter versus Facebook: Exploring the role of narcissism in the motives and usage of different social media platforms. *Computers in Human Behavior*, *32*, 212–220.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *ArXiv:1703.04009v1*, 1–4.

https://doi.org/10.1561/1500000001

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *ArXiv:1602.00975v1*, 4–5. https://doi.org/10.1145/2872518.2889302

de Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, (August), 1–8. https://doi.org/10.3115/1608858.1608859

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

DEMOS. (2017). *Anti-Islamic Hate on Twitter*. London.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. *Proceedings of the Recent Advances in Natural Language Processing*, (September), 198–206. Retrieved from http://www.aclweb.org/website/old_anthology/R/R13/R13-1026.pdf

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR) Surveys*, *27*(3), 326–327.

Djupe, P. A., & Sokhey, A. E. (2011). Interpersonal Networks and Democratic Politics. *PS: Political Science & Politics*, (January), 55–60. https://doi.org/10.1017/S1049096510001861

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipat, N. (2015). Hate Speech Detection with Comment Embeddings. *WWW 2015 Companion, May 18-25, Florence, Italy*, 29–30.

https://doi.org/10.1128/JB.186.10.3117-3123.2004

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), 1–57. https://doi.org/10.1371/journal.pone.0026752

Dodd, V, & Marsh, S. (2017). Anti-Muslim hate crimes increase fivefold since London Bridge attacks. *The Guardian*. Retreived from https://www.theguardian.com/uk-news/2017/jun/07/anti-muslim-hate-crimes-increase-fivefold-since-london-bridge-attacks

Doerr, N. (2017a). Bridging language barriers, bonding against immigrants: A visual case study of transnational network publics created by far-right activists in Europe. *Discourse and Society*, *28*(1), 3–23. https://doi.org/10.1177/0957926516676689

Doerr, N. (2017b). How right-wing versus cosmopolitan political actors mobilize and translate images of immigrants in transnational contexts. *Visual Communication*, *16*(3), 315–336. https://doi.org/10.1177/1470357217702850

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. https://doi.org/10.1145/2347736.2347755

Donnelly, J. P., & Woodruff, J. D. (2007). Intense hurricane activity over the past 5,000 years controlled by El Nino and the West African monsoon. *Nature*, *447*(May), 465–468. https://doi.org/10.1038/nature05834

Dranove, D. (2012). Practical Regression: Fixed Effects Models. *Kellog School of Management: Technical Notes*.

Druce, K. L., McBeth, J., van der Veer, S. N., Selby, D. A., Vidgen, B., Georgatzis, K., … Dixon, W. G. (2017). Recruitment and Ongoing Engagement in a UK

Smartphone Study Examining the Association Between Weather and Pain: Cohort Study. *JMIR MHealth and UHealth*, *5*(11), 1–14. https://doi.org/10.2196/mhealth.8162

Druce, K. L., Veer, S. N. Van Der, Beukenhorst, A. L., Laksshminarayana, R., Schultz, D. M., & Mcbeth, J. (2017). *Poster 299: Engagement in a United Kingdom smartphone study*.

Dunbar, R. I. M. (2016). Do Online Social Media Cut Through The Constraints That Limit the Size of Offline Social Networks? *Royal Society Open Science*, *3*(1), 1–9. https://doi.org/http://dx.doi.org/10.1098/rsos.150292

Dunn, K., & Hopkins, P. (2016). The Geographies of Everyday Muslim Life in the West. *Australian Geographer*, *47*(3), 255–260. https://doi.org/10.1080/00049182.2016.1191138

Dunn, K. K., Klocker, N., & Salabay, T. (2007). Contemporary racism and Islamaphobia in Australia: Racializing religion. *Ethnicities*, *7*(4), 564–589. https://doi.org/10.1177/1468796807084017

Eatwell, R. (1996). The Esoteric Ideology of National Front in the 1980s. In M. Cronin (Ed.), *The Failure of British Fascism* (pp. 99–117). London: Macmillan Publishers Limited.

Eatwell, R. (2006). Community Cohesion and Cumulative Extremism in Britain. *Political Quarterly*, *77*(2), 204–216.

Eatwell, R., & Goodwin, M. (2010). *The new extremism in the 21st century Britain*. London: Routledge.

ECRI. (2016). *ECRI Report on the United Kingdom: fifth monitoring cycle*. European Union: European Commission against Racism and Intolerance.

Edwards, G. O. (2012). A comparative discourse analysis of the construction of "in-groups" in the 2005 and 2010 manifestos of the British National Party. *Discourse and Society*, *23*(3), 245–258. https://doi.org/10.1177/0957926511433477

Ekman, M. (2015). Online Islamophobia and the politics of fear: manufacturing the green scare. *Ethnic and Racial Studies*, *38*(11), 1986–2002. https://doi.org/10.1080/01419870.2015.1021264

Ellinas, A. A. (2013). The Rise of Golden Dawn: The New Face of the Far Right in Greece. *South European Society and Politics*, *18*(4), 543–565. https://doi.org/10.1080/13608746.2013.782838

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: how politicians spread a fragmented ideology. *Information, Communication and Society*, *20*(8), 1109–1126. https://doi.org/10.1080/1369118X.2016.1207697

Ennser, L. (2012). The homogeneity of West European party families: The radical right in comparative perspective. *Party Politics*, *18*(2), 151–171. https://doi.org/10.1177/1354068810382936

Erdenir, B. (2010). Islamophobia qua racial discrimination. In A. Triandafyllidou (Ed.), *Muslims in 21st Century Europe: structural and cultural perspectives* (pp. 27–44). Abingdon: Routledge.

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization*,

10(Spring), 1–49.

European Monitoring Centre on Racism and Xenophobia. (2006). *Muslims in the European Union: Discrimination and Islamaphobia*. *EUMC Report*. Brussels. https://doi.org/ISBN: 92-9192-018-5

Europol. (2017). *TESAT: European Union Situation and Trend Report (2017)*. EU: European Union Ageny for Law Enforcement Cooperation. https://doi.org/10.2813/237471

Europol. (2018). *TESAT: European Union Situation and Trend Report (2018)*. EU: European Union Ageny for Law Enforcement Cooperation. https://doi.org/10.2813/00041

Evans, G., & Mellon, J. (2016). Working class votes and Conservative losses: Solving the UKIP puzzle. *Parliamentary Affairs*, *69*(2), 464–479. https://doi.org/10.1093/pa/gsv005

Evans, H., Ginnis, S., & Bartlett, J. (2015). #SocialEthics a guide to embedding ethics in social media research, (November).

Facebook (2019). Facebook: Hate Speech. Retrieved May 30, 2019, from https://www.facebook.com/communitystandards/hate_speech

FactMata. (2018). FactMata: About us.

Faliq, A. (2010). Islamophobia and anti-Muslim hatred: causes and remedies. *The Cordoba Foundation: Cultures in Dialogue*, *4*(7), 1–12.

Falkheimer, J., & Olsson, E. K. (2015). Depoliticizing terror: The news framing of the terrorist attacks in Norway, 22 July 2011. *Media, War and Conflict*, *8*(1), 70–85. https://doi.org/10.1177/1750635214531109

Feldman, M. (2015). *Tell MAMA Reporting 2015/2015: Annual Monitoring, Cumulative Extremism, and Policy Implications*. Teesside University.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, *5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Amorim Fernández-Delgado, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. https://doi.org/10.1016/j.csda.2008.10.033

Ferrara, E., Wang, W. Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10047 LNCS*, 22–39. https://doi.org/10.1007/978-3-319-47874-6_3

Feuerriegel, S., & Proellochs, N. (2018). *R Package: 'SentimentAnalysis.'* London.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society*, *4*(1), 1–14. https://doi.org/10.1177/2056305118763366

Filimonov, K., Russmann, U., & Svensson, J. (2016). Picturing the Party: Instagram and Party Campaigning in the 2014 Swedish Elections. *Social Media and Society*, *2*(3), 1–11. https://doi.org/10.1177/2056305116662179

Firth, J. R. (1957). *A synopsis of linguistic theory (1930-1955)*. Oxford: Oxford University Press.

Fishbein, M., & Ajzen, I. (2010). *Predicting and Changing Behavior*. Hove: Psychology

Press.

Fisher, J., Fieldhouse, E., & Cutts, D. (2014). Members Are Not the Only Fruit: Volunteer Activity in British Political Parties at the 2010 General Election. *The British Journal of Politics and International Relations*, *16*(1), 75–95. https://doi.org/10.1111/1467-856X.12011

Fitzpatrick, P. J. (2005). *Hurricanes A Reference Handbook*. Oxford: Wiley-Blackwell.

Five Thirty Eight. (2016, March 21). The World's Most Prolific Twitter User Tweets Mostly About Nothing. *Five Thirty Eight*. Retrieved from https://fivethirtyeight.com/features/the-worlds-most-prolific-twitter-user-tweets-mostly-about-nothing/

Fleck, C., & Müller, A. (1998). Front-stage and back-stage: the problem of measuring post-Nazi antisemitism in Austria. In S. U. Larsen (Ed.), *Modern Europa after Fascism, 1943-1980s*. New York: Social Science Monographs.

Flick, U. (2006). *An introduction to qualitative research*. London: SAGE.

Ford, R. (2010). Who might vote for the BNP? Survey evidence on the electoral potential of the extreme right in Britain. In R. Eatwell & M. Goodwin (Eds.), *The New Extremism in 21st Century Britain*. London: Routledge.

Ford, R., & Goodwin, M. (2010). Angry white men: Individual and contextual predictors of support for the British National Party. *Political Studies*, *58*(1), 1–25. https://doi.org/10.1111/j.1467-9248.2009.00829.x

Ford, R., & Goodwin, M. (2014a). *Revoilt on the right*. London: Routledge.

Ford, R., & Goodwin, M. (2014b). Understanding UKIP: Identity, Social Change and the Left Behind. *The Political Quarterly*, *85*(3), 277–284.

https://doi.org/10.1111/j.1467-923X.2014.00000.x

Ford, R., Goodwin, M., & Cutts, D. (2012). Strategic Eurosceptics and polite xenophobes: Support for the United Kingdom Independence Party (UKIP) in the 2009 European Parliament elections. *European Journal of Political Research*, *51*(2), 204–234. https://doi.org/10.1111/j.1475-6765.2011.01994.x

Ford, R., Jennings, W., & Somerville, W. (2017). Public Opinion, Responsiveness and Constraint: Britain's Three Immigration Policy Regimes. *Journal of Ethnic and Migration Studies*, *41*(9), 1391–1411. https://doi.org/10.1080/1369183X.2015.1021585

Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., … Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *ArXiv:1802.00393v2*. Retrieved from http://arxiv.org/abs/1802.00393

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ*, *337*(1), 1–9. https://doi.org/10.1136/bmj.a2338

Franceschini, F., & Maisano, D. (2016). Do Scopus and WoS correct "old" omitted citations? *Scientometrics*, *107*(2), 321–335. https://doi.org/10.1007/s11192-016-1867-8

Freelon, D. (2014). On the Interpretation of Digital Trace Data in Communication and Social Computing Research. *Journal of Broadcasting and Electronic Media*, *58*(1), 59–75. https://doi.org/10.1080/08838151.2013.875018

Freeman, J. B., & Ambady, N. (2011). A Dynamic Interactive Theory of Person Construal. *Psychological Review*, *118*(2), 247–279.

https://doi.org/10.1037/a0022327

Friemel, T. N. (2016). The digital divide has grown old: Determinants of a digital divide among seniors. *New Media and Society*, *18*(2), 313–331. https://doi.org/10.1177/1461444814538648

Froio, C. (2018). Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the French far right. *Perspectives on Politics*, *16*(3), 696–709. https://doi.org/10.1017/S1537592718001573

Froio, C., & Ganesh, B. (2018). The transnationalisation of far right discourse on Twitter: Issues and actors that cross borders in Western European democracies. *European Societies*, *0*(0), 1–27. https://doi.org/10.1080/14616696.2018.1494295

Fuchs, C. (2017). *Social media: a critical introduction*. London: SAGE Publications.

Futrell, R., & Simi, P. (2017). The [Un]Surprising Alt-Right. *Contexts*, *16*(2), 76–85. https://doi.org/10.1177/1536504217714269

Gabriel, F. (2013). Sexting, selfies and self-harm: young people, social media and the performance of self-development. *Media International Australia*, *151*(1), 104–112. https://doi.org/10.1177/1329878X1415100114

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Paris: United Nations Educational, Scientific and Cultural Organization.

Gallie, W. B. (1956). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*, 167–198.

Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. *Association for Computational Linguistics*, (7491), 85–90. Retrieved from http://www.aclweb.org/anthology/W17-

3013%0Ahttps://aclanthology.info/pdf/W/W17/W17-3013.pdf

Garcia-Gavilanes, R., Mollgaard, A., Tsvetkova, M., & Yasseri, T. (2017). Memory Remains : Understanding Collective Memory in the Digital Age. *Science Advances*, *3*(1), 1–7.

García-Gavilanes, R., Tsvetkova, M., & Yasseri, T. (2016). Dynamics and biases of online attention: The case of aircraft crashes. *Royal Society Open Science*, *3*(10), 1–13. https://doi.org/10.1098/rsos.160460

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays by Clifford Geertz*. New York: Basic Books, Inc. https://doi.org/10.1007/BF00695328

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. New York. https://doi.org/10.1037/a0037714

Gerbaudo, P. (2012). *Tweets and the streets: social media and contemporary activist*. London: Pluto Press.

Gerodimos, R. (2015). The Ideology of Far Left Populism in Greece: Blame, Victimhood and Revenge in the Discourse of Greek Anarchists. *Political Studies*, *63*(3), 608–625. https://doi.org/10.1111/1467-9248.12079

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, *69*, 214–224. https://doi.org/10.1016/j.eswa.2016.10.043

Gibson, R., & Ward, S. (2009). Parties in the Digital Age—a Review Article. *Representation*, *45*(1), 87–100. https://doi.org/10.1080/00344890802710888

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes. *Criminology and Public Policy*, *16*(1), 99–117. https://doi.org/10.1111/1745-9133.12249

Gill, P., Corner, E., Thornton, A., & Conway, M. (2015). *What Are the Roles of the Internet in Terrorism? Measuring online behaviours of convicted UK terrorists*. Retrieved from http://voxpol.eu/wp-content/uploads/2015/11/DCUJ3518_VOX_Lone_Actors_report_02.11.15_WEB.pdf

Gillespie, C. S. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*, *64*(2), 1–16. https://doi.org/10.18637/jss.v000.i00

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Gjoka, M., Butts, C. T., Kurant, M., & Markopoulou, A. (2011). Multigraph Sampling of Online Social Networks. *ArXiv:1008.2565v2*.

Glynos, J., & Howarth, D. (2007). *Logics of Critical Explanation in Social and Political Theory*. London: Routledge.

Glynos, J., Howarth, D., Norval, A., & Speed, E. (2009). ESRC National Centre for Research Methods Review paper Discourse Analysis : Varieties and Methods, (August), 1–41.

Goh, K. I., & Barabási, A. L. (2008). Burstiness and memory in complex systems. *Epl: A Letters Journal Exploring the Frontiers of Physics*, *81*(4), 1–5. https://doi.org/10.1209/0295-5075/81/48002

Golbeck, J. (2018). *Online harassment*. New York: Springer.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. New York: Morgan & Claypool Publishers.

Golder, M. (2016). Far Right Parties in Europe. *Annual Review of Political Science*, *19*(1), 477–497. https://doi.org/10.1146/annurev-polisci-042814-012441

Goodhart, D. (2014). Racism: less is more. *The Political Quarterly*, *85*(3), 251–258. https://doi.org/10.1111/j.1467-923X.2014.00000.x

Goodwin, M. (2006). The rise and faults of the Internalist Perspective in Extreme Right Studies. *Representation*, *42*(4), 347–364. https://doi.org/10.1080/00344890600951924

Goodwin, M. (2010). Activism in contemporary extreme right parties: The case of the British National Party (BNP). *Journal of Elections, Public Opinion and Parties*, *20*(1), 31–54. https://doi.org/10.1080/17457280903450690

Goodwin, M. (2011). *New British Fascism: Rise of the British National Party*. London: Routledge.

Goodwin, M. (2013a). Forever a False Dawn? The Collapse of the British National Party. *Parliamentary Affairs*, *1*(1), 1–20.

Goodwin, M. (2013b). The Roots of Extremism: The English Defence League and the Counter-Jihad Challenge. *Chatam House*. Retrieved from http://www.chathamhouse.org/publications/papers/view/189767

Goodwin, M., Ford, R., & Cutts, D. (2013). Extreme right foot soldiers, legacy effects and deprivation: A contextual analysis of the leaked British National Party (BNP) membership list. *Party Politics*, *19*(6), 887–906.

https://doi.org/10.1177/1354068811436034

Goodwin, M. J. (2008). Research, revisionists and the radical right. *Politics*, *28*(1), 33–40.

Google. (2018). word2vec Introduction. Retrieved October 17, 2018, from https://code.google.com/archive/p/word2vec/

Gottlieb, J. V. (2004). Women and British Fascism Revisited: Gender, the Far-Right, and Resistance. *Journal of Women*, *16*(3), 108–123. https://doi.org/10.1353/jowh.2004.0065

Gottschalk, P., & Greenberg, G. (2008). *Islamophobia: making Muslims the Enemy*. London: Rowman & Littlefield.

Graham, T., Jackson, D., & Broersma, M. (2014). New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns. *New Media and Society*, *18*(5), 765–783. https://doi.org/10.1177/1461444814546728

Green, E. G. T., Sarrasin, O., Fasel, N., & Staerkle, C. (2011). Nationalism and patriotism as predictors of immigration attitudes in Switzerland: A municipality-level analysis. *Swiss Political Science Review*, *17*(4), 369–393. https://doi.org/10.1111/j.1662-6370.2011.02030.x

Green, M. J. (2014). Latent class analysis was accurate but sensitive in data simulations. *Journal of Clinical Epidemiology*, *67*(10), 1157–1162. https://doi.org/10.1016/j.jclinepi.2014.05.005

Green, P., Glaser, J., & Rich, A. (1998). From lynching to gay bashing: the elusive connection between economic conditions and hate crime. *Journal of Personality and Social Psychology*, *75*(1), 82–92.

Greevy, E. P., & Smeaton, A. F. (2004). Text Categorisation of Racist Texts Using a Support Vector Machine. In *JADT: 7th International Journal for the Statistical Analysis of Textual Data* (pp. 533–544).

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet*, *359*(1), 248–252. https://doi.org/http://dx.doi.org/10.1016/S0140-6736(02)07451-2

Grimmer, J. (2014). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS - Political Science and Politics*, *48*(1), 80–83. https://doi.org/10.1017/S1049096514001784

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gruenewald, J., Chermak, S., & Freilich, J. D. (2013). Distinguishing "loner" attacks from other domestic extremist violence: A comparison of far-right homicide incident and offender characteristics. *Criminology and Public Policy*, *12*(1), 65–91. https://doi.org/10.1111/1745-9133.12008

Gruzd, A. (2014). Investigating Political Polarization on Twitter: A Canadian Perspective. *Policy and Internet*, *6*(1), 28–45. https://doi.org/10.1002/1944-2866.POI354

Guo, C., & Saxton, G. D. (2018). Speaking and Being Heard: How Nonprofit Advocacy Organizations Gain Attention on Social Media. *Nonprofit and Voluntary Sector Quarterly*, *47*(1), 5–26. https://doi.org/10.1177/0899764017713724

Hainmueller, J., & Hopkins, D. J. (2014). Public Attitudes Toward Immigration. *Annual Review of Political Science*, *17*(1), 225–249. https://doi.org/10.1146/annurev-

polisci-102512-194818

Halberstam, Y., & Knight, B. (2016). Homophily , group size , and the diffusion of political information in social networks : Evidence from Twitter ☙. *Journal of Public Economics*, *143*, 73–88. https://doi.org/10.1016/j.jpubeco.2016.08.011

Hale, S. A. (2014). Global connectivity and multilinguals in the Twitter network. In *Proceedings of the 2014 ACM Annual Conference on Human Factors in Computing Systems* (pp. 185–196). Montreal, Canada. https://doi.org/10.1007/s12026-013-8436-5

Hale, S. A., John, P., Margetts, H., & Yasseri, T. (2018). How digital design shapes political participation: A natural experiment with social information. *PLoS ONE*, *13*(4), 1–20. https://doi.org/10.1371/journal.pone.0196068

Halikiopoulou, D., & Vasilopoulou, S. (2014). Support for the Far Right in the 2014 European Parliament Elections : A Comparative Perspective. *The Political Quarterly*, *85*(3), 285–288. https://doi.org/10.1111/j.1467-923X.2014.00000.x

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quantative Methosd Psychology*, *8*(1), 23–34.

Halliday, F. (1999). "Islamophobia" reconsidered. *Ethnic and Racial Studies*, *22*(5), 892–902. https://doi.org/10.1080/014198799329305

Hamann, J., & Suckert, L. (2018). Temporality in Discourse: Methodological Challenges and a Suggestion for a Quantified Qualitative Approach. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *19*(2), 1–20. https://doi.org/10.17169/fqs-19.2.2954

Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in Practice*. London:

Routledge.

Hanes, E., & Machin, S. (2014). Hate Crime in the Wake of Terror Attacks: Evidence From 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, *30*(3), 247–267. https://doi.org/10.1177/1043986214536665

Haque, A., Tubbs, C. Y., Kahumoku-Fessler, E. P., & Brown, M. D. (2018). Microaggressions and Islamophobia: Experiences of Muslims Across the United States and Clinical Implications. *Journal of Marital and Family Therapy*, *44*(2).

Harris, R. M. B., Beaumont, L. J., Vance, T. R., Tozer, C. R., Nicotra, A. B., Mcgregor, S., … Bowman, D. M. J. S. (2018). Biological responses to the press and pulse of climate trends and extreme events. *Nature Climate Change*, *8*(1), 579–587. https://doi.org/10.1038/s41558-018-0187-9

Harzing, A., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787–804. https://doi.org/10.1007/s11192-015-1798-9

Harzing, A. W., & Adams, D. (2009). *Publish or Perish: realising Google Scholar's potential to democratize citation analysis*. Retrieved from http://www.harzing.com/pop.htm

Hayton, R. (2010). Towards the Mainstream? UKIP and the 2009 Elections to the European Parliament. *Politics*, *30*(1), 26–35.

Hee, C. Van, Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G. De, & Daelemans, W. (2015). Detection and Fine-Grained Classification of Cyberbullying Events. *International Conference Recent Advances in Natural Language Processing (RANLP)*, 672–680.

Heiss, R., Schmuck, D., & Matthes, J. (2018). What drives interaction in political actors'

Facebook posts? Profile and content predictors of user engagement and political actors' reactions. *Information Communication and Society*, *0*(0), 1–17. https://doi.org/10.1080/1369118X.2018.1445273

Hemsley, J., Jacobson, J., Gruzd, A., & Mai, P. (2018). Social Media for Social Good or Evil: An Introduction. *Social Media + Society*, *1*(1), 1–5. https://doi.org/10.1177/2056305118786719

Hermundstad, A. M., Brown, K. S., Bassett, D. S., & Carlson, J. M. (2011). Learning, memory, and the role of neural network architecture. *PLoS Computational Biology*, *7*(6), 1–12. https://doi.org/10.1371/journal.pcbi.1002063

Herring, S., Hoerling, M., Kossin, J., Peterson, T., & Stott, P. (2015). Explaining extreme events of 2014 from a climate perspective. *Bulletin of the American Meteorological Society*, *96*(12), 1–180.

Hervik, P. (2015). Xenophobia and nativism. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioural sciences* (pp. 796–801). Oxford: Elsevier. https://doi.org/10.1016/j.ssresearch.2009.02.002

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*(1), 575–604.

Hill, M. (2018). *The Terrorism Acts in 2017*. London: Independent Review of Terrorism Legislation.

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., … Blackburn, J. (2016). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. Retrieved from http://arxiv.org/abs/1610.03452

HM Government. (2012). *Deputy Prime Minister extends funding to tackle hate crime*

against Muslims. London.

HM Government. (2017a). *Hate crime: abuse, hate and extremism online*. London: House of Commons.

HM Government. (2017b). *Hate Crime, England and Wales, 2016/2017*. London: Home Office.

HM Government (2019) *Online harms: white paper*. London: Department of Digital, Culture, Media & Sport.

Hodgkin, A. (2017). *Following Searle on Twitter: How Words Create Digital Institutions*. Chicago: University of Chicago Press.

Hogg, M. A., Abrams, D., Otten, S., & Hinkle, S. (2004). Intergroup Relations, Self-Conception, and Small Groups. *Small Group Research*, *35*(3), 246–276. https://doi.org/10.1177/1046496404263424

Hogg, M. A., & Tindale, R. S. (2001). *Blackwell Handbook of Social Psychology: Group Processes*. London: Blackwell Publishers.

Hollenbaugh, E. E., & Ferris, A. L. (2014). Facebook self-disclosure: Examining the role of traits, social cohesion, and motives. *Computers in Human Behavior*, *30*(1), 50–58. https://doi.org/10.1016/j.chb.2013.07.055

Home Affairs Select Committee. (2017). *Home Affairs Committee Hate crime: abuse, hate and extremism online*. London.

Hope Not Hate. (2015). *Hope Not Hate: State of Hate 2015*. London: Hope Not Hate.

Hope Not Hate. (2017). *Hope Not Hate: State of Hate 2017*. London: Hope Not Hate.

Hopkins, L. (2008). Young Turks and new media: The construction of identity in an age of Islamophobia. *Media International Australia*, (126), 54–66.

https://doi.org/10.1177/1329878X0812600107

Hopkins, P., & Gale, R. (2009). *Muslims in Britain: Race, Place and Identities: Race, Place and Identities*. (P. Hopkins & R. Gale, Eds.). Edinburgh: Edinburgh University Press.

Housley, W., Procter, R., Edwards, A., Burnap, P., Williams, M., Sloan, L., … Greenhill, A. (2014). Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society*, *1*(2), 1–15. https://doi.org/10.1177/2053951714545135

Howard, P. N., & Parks, M. R. (2012). Social Media and Political Change: Capacity, Constraint, and Consequence. *Journal of Communication*, *62*(2), 359–362. https://doi.org/10.1111/j.1460-2466.2012.01626.x

Howard, P. N., Woolley, S., Calo, R., & Howard, P. N. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, *15*(2), 81–93. https://doi.org/10.1080/19331681.2018.1448735

Howarth, D. (2016). *Post-structuralism and after*. London: Routledge.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems (JAIS)*, *12*(12), 767–797. https://doi.org/10.1007/BF00733441

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector

machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. https://doi.org/10.1109/72.991427

Huang, Z. (1998). Extensions to the K-Means algorithm for clustering large datasets with categorical values. *Data Mining and Knowledge Discovery*, *2*(1), 283–304. https://doi.org/10.1023/A:1009769707641

Humble, Á. M. (2009). Technique Triangulation for Validation in Directed Content Analysis. *International Journal of Qualitative Methods*, *8*(3), 34–51.

Hussain, A. (2012). Confronting Misoislamia: Teaching Religion and Violence in Courses on Islam. In B. K. Pennington (Ed.), *Teaching religion and violence* (pp. 118–148). Oxford: Oxford University Press.

Hyde, K. F. (2000). Recognising deductive and inductive processes in qualitative research. *Qualitative Market Research: An International Journal*, *3*(2), 82–90.

Ignazi, P. (1997). New Challenges: Post-materialism and the Extreme Right. *Developments in West European Politics*, (December), 300–319.

Ignazi, P. (2003). *Extreme right parties in Western Europe*. Oxford: University of Oxford.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789. https://doi.org/10.1017/S0003055411000414

Imhoff, R., & Recker, J. (2012). Differentiating Islamophobia: Introducing a New Scale to Measure Islamoprejudice and Secular Islam Critique, *33*(6). https://doi.org/10.1111/j.1467-9221.2012.00911.x

Ingham-Barrow, I. (2018). *More than words: approaching a definition of Islamophobia*. (I. Ingham-Barrow, Ed.). London: MEND.

Ipsos MORI. (2016). *Attitudes to potentially offensive language on TV and radio*. London. Retrieved from http://www.ipsos-mori.com/terms.

Iqbal, Z. (2010). Islamophobia or Islamophobias: Towards Developing A Process Model. *Islamic Studies*, *49*(1), 81–101. Retrieved from http://www.jstor.org/stable/41429246

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014). Political Ideology Detection Using Recursive Neural Networks. *Acl-2014*, 1113–1122. https://doi.org/10.1017/CBO9781107415324.004

Jacks, W., & Adler, J. R. (2015). A proposed typology of online hate crime. *Open Access Journal of Forensic Psychology*, *7*(1), 64–89.

Jackson, L. R. (2018). *Islamophobia in Britain: the making of a Muslim enemy*. London: Palgrave Macmillan UK.

Jackson, P., & Feldman, M. (2011). *The EDL: Britain's 'New Far Right' social movement*. Northampton.

Jakubowicz, A. (2017). Alt-right white lite: trolling, hate speech and cyber racism on social media. *Cosmopolitan Civil Societies: An Interdisciplinary Journal*, *9*(3), 41–60.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis) (2015)*, 1–21. https://doi.org/10.2312/eurovisstar.20151113

Jaromir Antoch, Jan Hanousek, Lajos Horvath, Marie Huskova, S. W. (2017). Structural breaks in panel data: Large number of panels and short length time series. *CEPR Discussion Paper Series*, *11891*(0), 1–28. https://doi.org/10.1080/07474938.2018.1454378

Jetten, J., Spears, R., & Postmes, T. (2004). 'Intergroup Distinctiveness and Differentiation: A Meta-Analytic Integration', *Journal of personality and social psychology, 86(6),* 862-879.

Jha, A., & Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 7–16. https://doi.org/10.18653/v1/W17-2902

John, P., & Margetts, H. (2009). The Latent Support for the Extreme Right in British Politics. *West European Politics*, *32*(3), 496–513.

John, P., Margetts, H., Rowland, D., & Weir, S. (2004). The BNP: the roots of its appeal. Colchester: University of Essex.

Johns, A. H., & Saeed, A. (2002). Muslims in Australia: the building of a community. In Y. Y. Haddad & J. I. Smith (Eds.), *Muslim minorities in the West: visible and invisible*. Oxford: Altamira Press.

Jones, B., Norton, P., & Daddow, O. (2018). *Politics UK [9th edition]*. London: Routledge.

Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., & Uskali, T. (2016). Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication*, *4*(4), 75–85. https://doi.org/10.17645/mac.v4i4.579

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient

Text Classification. *ArXiv:1607.01759v3*, (2), 759–760. https://doi.org/1511.09249v1

Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, *13*(1), 72–91. https://doi.org/10.1080/19331681.2015.1132401

JUST. (2018). Rethinking PREVENT: A case for an alternative approach. Yorkshire: JUST Yorkshire

Kamkarhaghighi, M., & Makrehchi, M. (2017). Content Tree Word Embedding for document representation. *Expert Systems With Applications*, *90*, 241–249. https://doi.org/10.1016/j.eswa.2017.08.021

Kantha, L. (2006). Time to Replace the Saffir- Simpson Hurricane Scale ? *Eos*, *87*(3–6), 2005–2006.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Karsai, M., Kaski, K., Barabási, A. L., & Kertész, J. (2012). Universal features of correlated bursty behaviour. *Scientific Reports*, *2*, 1–7. https://doi.org/10.1038/srep00397

Kassimeris, G., & Jackson, L. (2015). The Ideology and Discourse of the English Defence League: "Not Racist, Not Violent, Just No Longer Silent." *British Journal of Politics and International Relations*, *17*(1), 171–188. https://doi.org/10.1111/1467-856X.12036

Katz, Richard S.; Mair, P., Haid, C. J., Press, C., Borz, G., & Janda, K. (2008). The Three Faces of Party Organization. *Political Science*, *46*(February), 593–617.

https://doi.org/10.2307/40732121

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup Perception and Cognition: An Integrative Framework for Understanding the Causes and Consequences of Social Categorization. *Advances in Experimental Social Psychology*, *55*(1), 1–80. https://doi.org/10.1016/bs.aesp.2016.10.001

Khoury, M. J., & Ioannidis, J. P. A. (2014). Big data meets public health. *Science*, *346*(6213), 1054–1055. https://doi.org/10.1126/science.aaa2709

Kietzmann, J. H., Hermkens, K., Mccarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(1), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Kim, N., & Wojcieszak, M. (2018). Intergroup contact through online comments: Effects of direct and extended contact on outgroup attitudes. *Computers in Human Behavior*, *81*(1), 63–72. https://doi.org/10.1016/j.chb.2017.11.013

King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology*, *51*(4), 871–894. https://doi.org/10.1111/1745-9125.12022

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 205395171663113. https://doi.org/10.1177/2053951716631130

Klug, B. (2012). Islamophobia: A concept comes of age. *Ethnicities*, *12*(5), 665–681. https://doi.org/10.1177/1468796812450363

Klug, B. (2014). The limits of analogy: comparing Islamophobia and antisemitism. *Patterns of Prejudice*, *48*(5), 442–459. https://doi.org/10.1080/0031322X.2014.964498

Kollanyi, B., Howard, P. N., & Woolley, S. C. (2016). *Bots and automation over Twitter during the first U.S. presidential debate*. *COMPROP Data Memo* (Vol. 14 October). Oxford.

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis. *BMJ*, *350*(2750), 1–4. https://doi.org/10.1136/bmj.h2750

Korba, K. A., & Arbaoui, F. (2018). SVM Multi-Classification of Induction Machine's bearings defects using Vibratory Analysis based on Empirical Mode Decomposition. *International Journal of Applied Engineering Research*, *13*(9), 6579–6586.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. https://doi.org/10.1115/1.1559160

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 8788–8790. https://doi.org/10.1073/pnas.1412469111

Kuckartz, U. (2014). *Qualitative text analysis: a guide to methods, practice and using software*. London: SAGE.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. *Proceedings of the First Workshop on Trolling,*

Aggression and Cyberbullying, 1–11. Retrieved from http://aclweb.org/anthology/W18-4401

Kundnani, A. (2007). Integrationism: the politics of anti-Muslim racism. *Race & Class*, *48*(4), 24–44. https://doi.org/10.1177/0306396807077069

Kunst, J. R., Sam, D. L., & Ulleberg, P. (2013). Perceived Islamophobia: Scale development and validation. *International Journal of Intercultural Relations*, *37*(2), 225–237. https://doi.org/10.1016/j.ijintrel.2012.11.001

Kutneski, P. J. (2009). *Structural Attributes Associated with the Prevalence of Hate Groups: A State-Level Analysis*. Nevada.

Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 1621–1622.

Labrinidis, A., & Jagadish, H. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032–2033. https://doi.org/10.14778/2367502.2367572

Laclau, E. (1996). *Emancipation(s)*. London: Verso.

Laclau, E. (2005a). *On Populist Reason*. London: Verso.

Laclau, E. (2005b). Power and social communication. *Ethical Perspectives*, *7*(2–3), 139–145. https://doi.org/10.2143/EP.7.2.503799

Laclau, E., & Mouffe, C. (1985). *Hegemony and socialist strategy*. London: Routledge.

Lai, C., Guo, S., Cheng, L., & Wang, W. (2017). A comparative study of feature selection methods for the discriminative analysis of temporal lobe epilepsy. *Frontiers in Neurology*, *8*(1), 1–13. https://doi.org/10.3389/fneur.2017.00633

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding.

IEEE Intelligent Systems, *31*(6), 5–14. https://doi.org/10.1109/MIS.2016.45

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2267–2273.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *International Biometric Society1*, *33*(1), 159–174.

Larsson, A. O., & Hallvard, M. (2015). Bots or journalists? News sharing on Twitter. *Communications*, *40*(3), 361–370. https://doi.org/10.1515/commun-2015-0014

Larsson, G. (2007). Cyber-Islamophobia? The case of WikiIslam. *Cont. Islam*, *1*(1), 53–67. https://doi.org/10.1007/s11562-007-0002-2

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). "The whole is always smaller than its parts" - a digital test of Gabriel Tardes' monads. *British Journal of Sociology*, *63*(4), 590–615. https://doi.org/10.1111/j.1468-4446.2012.01428.x

Law Commission. (2014). *Hate Crime: Should the current offences be extended?* London: The Law Commission.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Fly: Traps in Big Data Analysis. *Science*, *343*(March), 1203–1206. https://doi.org/10.1126/science.1248506

Lazer, D. M. J., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Alstyne, M. Van. (2009). Computational social science. *Science*, *323*(6), 721–723. https://doi.org/10.1126/science.1167742

Lazer, D., & Radford, J. (2017). Data Ex Machina: Introduction to Big Data. *Annual*

Review of Sociology, *43*(1), 1–21. https://doi.org/10.1146/annurev-soc-060116-053457

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, pp. 1–9). Beijing, China. https://doi.org/10.1145/2740908.2742760

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, B. (2016). Why we fight: Understanding the counter-jihad movement. *Religion Compass*, *10*(10), 257–265. https://doi.org/10.1111/rec3.12208

Lee, B. (2017). A Day in the " Swamp ": Understanding Discourse in the Online Counter-Jihad Nebula. *Democracy and Security*, *11*(3), 248–274. https://doi.org/10.1080/17419166.2015.1067612

Lee, S. A., Gibbons, J. A., Thompson, J. M., & Timani, H. S. (2009). The Islamophobia Scale: Instrument Development and Initial Validation. *The International Journal for the Psychology of Religion*, *19*(2), 92–105. https://doi.org/10.1080/10508610802711137

Leruth, B., & Taylor-Gooby, P. (2018). Does political discourse matter? Comparing party positions and public attitudes on immigration in England. *Politics*, *0*(0), 1–16. https://doi.org/10.1177/0263395718755566

Leskovec, J., & Faloutsos, C. (2006). Sampling from Large Graphs. In *KDD '06*. Philadelphia, Pennsylvania.

Levi-Strauss, C. (1952). *Race and History*. New York: UNESCO.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties,

and time: a new social network dataset using Facebook.com. *Social Networks*, *30*(1), 330–342. https://doi.org/10.1016/j.socnet.2008.07.002

Liberty. (2018). Liberty: Speech offences. Retreived May 11, 2019, from https://www.libertyhumanrights.org.uk/human-rights/free-speech-and-protest/speech-offences

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Liu, P., Choo, K. R., Wang, L., & Huang, F. (2017). SVM or deep learning? A comparative study on remote sensing image classification. *Soft Computing*, *21*(23), 7053–7065. https://doi.org/10.1007/s00500-016-2247-2

Lowe, B. (1985). *Islam and the media*. Sydney.

Lowry, P. B., Roberts, T. L., Romano, N. C., Cheney, P. D., & Hightower, R. T. (2006). The Impact of Group Size and Social Presence on Small-Group Communication: Does Computer-Mediated Communication Make a Difference? *Small Group Research*, *37*(6), 631–661. https://doi.org/10.1177/1046496406294322

Lucassen, G., & Lubbers, M. (2012). Who Fears What? Explaining Far-Right-Wing Preference in Europe by Distinguishing Perceived Cultural and Economic Ethnic Threats. *Comparative Political Studies*, *45*(5), 547–574. https://doi.org/10.1177/0010414011427851

Ludemann, D. (2018). /pol/emics: Ambiguity, scales, and digital discourse on 4chan. *Discourse, Context and Media*, *24*, 92–98. https://doi.org/10.1016/j.dcm.2018.01.010

Luqiu, L. R., & Yang, F. (2018). Islamophobia in China: news coverage, stereotypes, and

Chinese Muslims' perceptions of themselves and Islam. *Asian Journal of Communication*, *1*(1), 1–22. https://doi.org/10.1080/01292986.2018.1457063

Lynch, P., Whitaker, R., & Loomes, G. (2011). *The UK Independence Party: analysing its candidates and supporters*. Leicester.

Macklin, G. (2013). Transnational Networking on the Far Right: The Case of Britain and Germany. *West European Politics*, *36*(1), 176–198. https://doi.org/10.1080/01402382.2013.742756

Magu, R., Joshi, K., & Luo, J. (2017). Detecting the Hate Code on Social Media. *ArXiv:1703.05443v1*, 1–5. https://doi.org/S0764-4469(99)80037-7 [pii]

Maio, G. R., Haddock, G., & Verplanken, B. (2019). *The psychology of attitudes and attitude change (3rd edition)*. London: SAGE.

Malik, M. (2009). Anti-Muslim prejudice in the West, past and present: an introduction. *Patterns of Prejudice*, *43*(3–4), 207–212. https://doi.org/10.1080/00313220903109144

Malmasi, S., & Zampieri, M. (2017). Detecting Hate Speech in Social Media. *ArXiv Preprint:1712.06427v2*. https://doi.org/10.26615/978-954-452-049-6_062

Mammone, A., Godin, E., & Jenkins, B. (2012). *Mapping the Extreme Right in Contemporary Europe: From Local to Transnational*. Abingdon: Routledge.

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities*, *1*(1), 1–17. https://doi.org/http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *Computer Science*

Review, *27*(1), 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

Margetts, H. (2006). Cyber parties. In R. S. Katz & W. Crotty (Eds.), *Handbook of Party Politics* (pp. 528–535). London: SAGE.

Margetts, H. (2017a). Political behaviour and the acoustics of social media. *Nature Human Behaviour*, *1*(4), 1–3. https://doi.org/10.1038/s41562-017-0086

Margetts, H. (2017b). The Data Science of Politics. *Political Studies Review*, *15*(2), 201–209. https://doi.org/10.1177/1478929917693643

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: how social media shape collective action*. Oxford: Princeton University Press.

Margetts, H., John, P., & Weir, S. (2004). Latent Support for the Far-Right in British Politics: The BNP and UKIP in the 2004 European and London Elections. In *PSA EPOP Conference, September 10-12* (pp. 1–24). Oxford.

Markham, A., & Buchanan, E. (2012). Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee. *Recommendations from the AoIR Ethics Working Committee (Version 2.0)*, 19. https://doi.org/Retrieved from www.aoir.org

Marranci, G. (2004). Multiculturalism, Islam and the clash of civilisations theory: rethinking Islamophobia. *Culture and Religion*, *5*(1), 105–117. https://doi.org/10.1080/0143830042000200373

Martin, J., & Hall, D. B. (2016). R2 measures for zero-inflated regression models for count data with excess zeros. *Journal of Statistical Computation and Simulation*, *86*(18), 3777–3790. https://doi.org/10.1080/00949655.2016.1186166

Martin, K., & Shilton, K. (2016). Why Experience Matters to Privacy: How Context-

Based Experience Moderates Consumer Privacy Expectations. *Journal of the Association for Information Science and Technology*, *67*(8), 1871–1882. https://doi.org/10.1002/asi

Maruf, H. Al, Meshkat, N., Ali, M. E., & Mahmud, J. (2015). Human behaviour in different social medias: A case study of Twitter and Disqus (pp. 1–5). https://doi.org/10.1145/2808797.2809395

Marwick, A. E., & boyd. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Marwick, A. E., & boyd. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media and Society*, *16*(7), 1051–1067. https://doi.org/10.1177/1461444814543995

Mathison, S. (1988). Why Triangulate ? *Educational Researcher*, *17*(2), 13–17.

Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). *Words that wound*. London: Routledge.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work and think*. London: John Murray.

Maynard, J., & Benesch, S. (2016). Dangerous speech and dangerous ideology: an integrated model for monitoring and prevention. *Genocide studies and prevention: an international journal*, *9(3),* 70-95.

Maynard, D., & Greenwood, M. a. (2014). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* (pp. 4238–4243). Reykjavik,

Iceland. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/67_Paper.pdf

Mccall, L., Crenshaw, K., & Cho, S. (2013). Toward a field of intersectionality studies: theory, applications and praxis. *Signs: Journal of Women in Culture and Society*, *38*(4), 785–810.

McCambridge, J., Witton, J., & Elbourne, D. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

McCauley, C., & Moskalenko, S. (2014). Toward a Profile of Lone Wolf Terrorists: What Moves an Individual From Radical Opinion to Radical Action. *Terrorism and Political Violence*, *26*(1), 69–85. https://doi.org/10.1080/09546553.2014.849916

Mcclurg, S. D. (2006). Networks on Political Participation, *50*(3), 737–754.

McCosker, A., & Wilken, R. (2014). Rethinking “big data” as visual knowledge: The sublime and the diagrammatic in data visualisation. *Visual Studies*, *29*(2), 155–164. https://doi.org/10.1080/1472586X.2014.887268

Mcdowall, D., Mccleary, R., Meidinger, E. E., & Hay, R. A. (1980). Introduction to Interrupted Time Series Analysis. In *Interrupted Time Series Analysis* (pp. 9–15). London: SAGE.

Mchugh, M. L. (2013). The Chi-square test of independence Lessons in biostatistics. *Biochemia Medica*, *23*(2), 143–149. https://doi.org/10.11613/BM.2013.018

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, *22*(3), 276–282.

Mckelvey, K., Digrazia, J., & Rojas, F. (2014). Twitter publics: how online political communities signaled electoral outcomes in the 2010 US house election. *Information, Communication & Society*, *17*(4), 436–450. https://doi.org/10.1080/1369118X.2014.892149

McPhillips, L. E., Chang, H., Chester, M. V., Depietri, Y., Friedman, E., Grimm, N. B., … Shafiei Shiva, J. (2018). Defining Extreme Events: A Cross-Disciplinary Review. *Earth's Future*, *6*(3), 441–455. https://doi.org/10.1002/2017EF000686

Meddaugh, P. M., & Kay, J. (2009). Hate Speech or "Reasonable Racism?" The Other in Stormfront. *Journal of Mass Media Ethics*, *24*(4), 251–268. https://doi.org/10.1080/08900520903320936

Meehl, G. A., Karl, T., Easterling, D. R., Changnon, S., Jr, R. P., Changnon, D., … Zwiers, F. (2000). An Introduction to Trends in Extreme Weather and Climate Events: Observations , Socioeconomic Impacts, Terrestrial Ecological Impacts, and Model Projections. *Bulletin of the American Meteorological Society*, *81*(3), 413–416.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Meer, N. (2013). Racialization and religion: race, culture and difference in the study of antisemitism and Islamophobia. *Ethnic and Racial Studies*, *36*(3), 385–398. https://doi.org/10.1080/01419870.2013.734392

Meer, N., & Modood, T. (2009). Refutations of racism in the 'Muslim question.' *Patterns of Prejudice*, *43*(3), 335–354. https://doi.org/10.1080/00313220903109250

Mehdad, Y., & Tetreault, J. (2016). Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (September), 299–303. https://doi.org/10.18653/v1/W16-3638

Mellon, J., & Evans, G. (2016). Class, electoral geography and the future of UKIP: Labour's secret weapon? *Parliamentary Affairs*, *69*(2), 492–498. https://doi.org/10.1093/pa/gsv013

Mesirov, J. P. (2010). Accessible Reproducible Research. *Science*, *327*(January), 415–417.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, 1–12. https://doi.org/10.1162/153244303322533223

Milgram, J., Cheriet, M., & Sabourin, R. (2006). "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? *Tenth International Workshop on Frontiers in Handwriting Recognition*, 1–6. Retrieved from http://hal.inria.fr/inria-00103955

Miller, D., & Horst, H. A. (2012). Introduction to Digital Anthropology. In H. A. Horst & D. Miller (Eds.), *Digital Anthropology*. London: Berg.

Mimno, D., Magnusson, M., Barrling, K., & Ohrvall, R. (2017). *Voices from the far right: a text analysis of Swedish parliamentary debates*.

Miro-Llinares, F., & Rodriguez-Sala, J. J. (2016). Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design and Nature and Ecodynamics*, *11*(3), 406–415. https://doi.org/10.2495/DNE-V11-N3-406-415

Mirza, H. S. (2013). "A second skin": Embodied intersectionality, transnationalism and

narratives of identity and belonging among Muslim women in Britain. *Women's Studies International Forum*, *36*, 5–15. https://doi.org/10.1016/j.wsif.2012.10.012

Modood, T. (2003). Muslims and the Politics of Difference. *Political Quarterly*, *74*(1), 100–115.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Mondon, A., & Winter, A. (2017). Articulations of Islamophobia: from the extreme to the mainstream? *Ethnic and Racial Studies*, *40*(13), 2151–2179. https://doi.org/10.1080/01419870.2017.1312008

Mongeon, P. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Monroe, B. L., Roberts, M. E., Diego, S., & Sinclair, B. (2015). No! Formal theory, causal inference, and big data are not contradictory trends in political science. *P.S. Political Science & Politics*, (January), 71–74. https://doi.org/10.1017/S1049096514001760

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS ONE*, *12*(9), 1–12. https://doi.org/10.1371/journal.pone.0184148

Moosavi, L. (2015). The Racialization of Muslim Converts in Britain and Their Experiences of Islamophobia. *Critical Sociology*, *41*(1), 41–56. https://doi.org/10.1177/0896920513504601

Moretti, F. (2013). *Distant Reading*. London: Verso.

Morse, J. M. (2003). Principles of mixed method and multimethod research design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 189–208). London: SAGE. https://doi.org/10.4135/9781446268308.n19

Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the technology acceptance model. *International Journal of Information Management*, *36*(6), 1248–1259. https://doi.org/10.1016/j.ijinfomgt.2016.07.007

Moten, A. (2012). Understanding and ameliorating Islamophobia. *Culture, International Journal of Philosophy of Culture and Axiology*, *9*(1), 155–178.

Mouffe, C. (2005). *On the Political*, London: Routledge.

Mouffe, C. (2009). *The Democratic Paradox*, London: Verso.

Muchnik, L., Pei, S., Parra, L. C., Reis, S. D. S., Andrade, J. S., Havlin, S., & Makse, H. A. (2013). Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, *3*, 1–8. https://doi.org/10.1038/srep01783

Mudde, C. (2002). *The Ideology of the Extreme Right*. Manchester: Manchester University Press. https://doi.org/10.7228/manchester/9780719057939.001.0001

Mudde, C. (2007a). *Populist radical right parties in Europe*. London: Routledge.

Mudde, C. (2007b). The war of words defining the extreme right party family. *West European Politics*, *19*(2), 225–248. https://doi.org/10.1080/01402389608425132

Mudde, C. (2008). *The populist radical right: a pathological normalcy*. *Willy Brandt series of working papers in international migration and ethnic relations* (Vol. 3). Malmo.

Mudde, C. (2009). Populist Radical Right Parties in Europe Redux. *Political Studies Review*, *7*(3), 330–337. https://doi.org/10.1111/j.1478-9302.2009.00194.x

Mudde, C. (2014). Fighting the system? Populist radical right parties and party system change. *Party Politics*, *20*(2), 217–226. https://doi.org/10.1177/1354068813519968

Mudde, C. (2017). The single-issue party thesis: Extreme right parties and the immigration issue. *West European Politics*, *22*(3), 182–197. https://doi.org/10.1080/01402389908425321

Mudde, C., & Kaltwasser, R. (2007). *Voices of the peoples: populism in Europe and Latin America compared* (July 2011 No. 378). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.004

Mulinari, D., & Neergaard, A. (2011). The Sweden democrats, racisms and the construction of the Muslim Threat. In G. Morgan & S. Poynting (Eds.), *Global Islamophobia: Muslims and moral panic in the West* (pp. 4–27). Routledge.

Müller, K., & Schwarz, C. (2017). *Fanning the Flames of Hate: Social Media and Hate Crime*. *SSRN*. https://doi.org/10.2139/ssrn.3082972

Mumby, P. J. (2002). Statistical power of non-parametric tests: A quick guide for designing sampling strategies. *Marine Pollution Bulletin*, *44*(1), 85–87. https://doi.org/10.1016/S0025-326X(01)00097-2

Mummolo, J., & Peterson, E. (2018). Improving the Interpretation of Fixed Effects Regression Results. *Political Science Research and Methods*, 1–7. https://doi.org/10.1017/psrm.2017.44

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Nacos, B. L., & Torres-Reyna, O. (2007). *Fueling our Fears: Stereotyping, Media Coverage and Public Opinion of Muslim Americans*. New York: Rowman & Littlefield.

Nadal, K. L., Griffin, K. E., Hamit, S., Leon, J., Tobio, M., & Rivera, D. P. (2012). Subtle and Overt Forms of Islamophobia: Microaggressions toward Muslim Americans. *Journal of Muslim Mental Health*, *6*(2), 15–37.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

NASA. (2018). NASA: What are hurricanes? Retrieved December 21, 2018, from https://www.nhc.noaa.gov/aboutsshws.php

National Commission. (1979). *The Belmont Report: ethical principles and guidelines for research involving human subjects*. Washington. https://doi.org/10.1021/bi00780a005

Neil Adger, W., Hughes, T. P., Folke, C., Carpenter, S., & Rockström, J. (2005). Social-Ecological Resilience to Coastal Disasters. *Science*, *309*(August), 1036–1040.

Nemer, D. (2016). Celebrities Acting up: A Speech Act Analysis in Tweets of Famous People. *Social Networking*, *5*(1), 1–10. https://doi.org/10.4236/sn.2016.51001

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *WWW '16 Proceedings of the 25th International Conference on World Wide Web*, *Montreal,*(April 11-15, 2016), 145–153.

Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*.

New York: NYU Press.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Oliver, J. E., & Rahn, W. M. (2016). Rise of the Trumpenvolk: populism in the 2016 election. *The Annals of the American Academy*, *667*(September), 189–206. https://doi.org/10.1177/0002716216662639

Onwuegbuzie, A. J., & Collins, K. M. T. (2007). A Typology of Mixed Methods Sampling Designs in Social Science Research. *The Qualitative Report*, *12*(2), 281–316.

Open Science. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–952. https://doi.org/10.1126/science.aac4716

Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: challenging the separation of technology, work and organization. *The Academy of Management Annals*, *2*(1), 433–474. https://doi.org/10.1080/19416520802211644

Panayi, P. (2014). *An immigration history of Britain: multicultural racism since 1800*. London: Routledge.

Parfitt, D. (1987). *Reasons and Persons*. Oxford: Oxford University Press.

Parikh, R., Parikh, S., Mathai, A., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50.

Parisier, E. (2012). *The filter bubble: what the Internet is hiding from you*. London:

Penguin.

Park, J. H., & Fung, P. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. *ArXiv Preprint*. https://doi.org/10.18653/v1/W17-3006

Parsell, M. (2008). Pernicious virtual communities: Identity, polarisation and the Web 2.0. *Ethics and Information Technology*, *10*(1), 41–56. https://doi.org/10.1007/s10676-008-9153-y

Peddell, D., Eyre, M., McManus, M., & Bonworth, J. (2016). Influences and vulnerabilities in radicalised lone-actor terrorists: UK practitioner perspectives. *International Journal of Police Science & Management*, *18*(2), 63–76. https://doi.org/10.1177/1461355716638686

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(December), 1226–1228.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perkmann, M., Fini, R., Ross, J., Salter, A., Silvestri, C., & Tartari, V. (2015). Accounting for universities' impact: using augmented data to measure academic engagement and commercialization by academic scientists. *Research Evaluation*, *24*(September), 380–391. https://doi.org/10.1093/reseval/rvv020

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57–75.

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice?

Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*(1), 922–934. https://doi.org/10.1002/ejsp

Pettigrew, T. F., Tropp, L. R., Wagner, U., & Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, *35*(3), 271–280. https://doi.org/10.1016/j.ijintrel.2011.03.001

Pew Research. (2016). *Social Media Update 2016*. New York. Retrieved from www.pewresearch.org

Phua, J., Venus, S., & Jay, J. (2017). Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, Twitter, Instagram, and Snapchat. *Computers in Human Behavior*, *72*, 115–122. https://doi.org/10.1016/j.chb.2017.02.041

Pickering, G., Bull, J. M., & Sanderson, D. J. (1995). Sampling power-law distributions, *248*, 1–20.

Pogorelis, R., Maddens, B., Swenden, W., & Fabre, E. (2006). Issue salience in regional and national party manifestos in the UK. *West European Politics*, *28*(5), 992–1014. https://doi.org/10.1080/01402380500310667

Poole, E. (2002). *Reporting Islam: media representations of British Muslims*. London: I. B. Tauris.

Porta, D. Della, & Diani, M. (2006). *Social movements: an introduction*. London: Wiley-Blackwell.

Postmes, T., Spears, R., & Lea, M. (2002). Intergroup Differentiation in Computer-Mediated Communication: Effects of Depersonalization. *Group Dynamics: Theory, Research and Practice*, *6*(1), 3–16. https://doi.org/10.1037//1089-2699.6.1.3

Poynting, S., & Briskman, L. (2018). Islamophobia in Australia: From far-right deplorables to respectable liberals. *Social Sciences*, *7*(11), 1–17. https://doi.org/10.3390/socsci7110213

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle., B. F. (1994). Social Dominance Orientation: A Personality Variable Predicting Social and Political Attitudes. *Journal of Personality and Spcial Psychology*, *67*(4), 741–763.

Proferes, N. (2017). Information Flow Solipsism in an Exploratory Study of Beliefs About Twitter. *Social Media + Society*, *3*(1), 1–17. https://doi.org/10.1177/2056305117698493

Puschmann, C., Bastos, M. T., & Schmidt, J. (2017). Birds of a feather petition together? Characterizing e-petitioning through the lens of platform data. *Information, Communication & Society*, *20*(2), 203–220. https://doi.org/10.1080/1369118X.2016.1162828

Quinn, E. (2018). Deer crossing. Moose crossing. Old people crossing. Children crossing: Reading Islamophobia through a vegan lens in Kamila Shamsie's Burnt Shadows. *The Journal of Commonwealth Literature*, *53*(1), 109–123. https://doi.org/10.1177/0021989416634767

Ramasco, J., Moro, E., Pujol, J. M., Eguiluz, V. M., & Grabowicz, P. A. (2012). Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, *7*(1), 1–9. https://doi.org/10.1371/journal.pone.0029358

Rana, J. (2007). The Story of Islamophobia. *Souls*, *9*(2), 148–161. https://doi.org/10.1080/10999940701382607

Ranstorp, M. (2010). *Understanding violent radicalisation: Terrorist and jihadist movements in Europe*. (M. Ranstorp, Ed.), *Introduction*. Abingdon: Routledge.

https://doi.org/10.4324/9780203865743

Renton, D. (2004). Examining the sucess of the British National Party, 1999-2003. *Race and Class*, *45*(2), 76–85.

Reuters. (2018, January 19). Social media companies accelerate removals of online hate speech: EU. *Reuters*. Retrieved from https://www.reuters.com/article/us-eu-hatespeech/social-media-companies-accelerate-removals-of-online-hate-speech-eu-idUSKBN1F806X

Rezaeinia, S. M., Ghodsi, A., & Rahmani, R. (2017). Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis. Retrieved from http://arxiv.org/abs/1711.08609

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W. (2018). Characterizing and Detecting Hateful Users on Twitter. https://doi.org/10.475/123_4

Richardson, J. E., & Wodak, R. (2008). The Impact of Visual Racism: Visual Arguments in Political Leaflets of Austrian and British Far- right Parties. *Controversia*, *6*(2), 45–77. Retrieved from http://engl894dwrs13.courses.digitalodu.com/wp-content/uploads/2013/07/Wodak.pdf

Richardson, J. E., & Wodak, R. (2017). Recontextualising Fascist ideologies of the past: right-wing discourses on employment and nativism in Austria and the United Kingdom. *Critical Discourse Studies*, *6*(4), 251–267. https://doi.org/10.1080/17405900903180996

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing and Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *WWW'11 Proceedings of the 20th International Conference on World Wide Web*, 695–704. https://doi.org/10.1145/1963405.1963503

Römmele, A. (2003). Political parties, party communication and new information and communication technologies. *Party Politics*, *9*(1), 7–20. https://doi.org/10.1177/135406880391002

Rosenfeld, M. (2012). Hate speech in constitutional jurisprudence: a comparative analysis. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: rethinking regulation and responses*. Cambridge: Cambridge University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *ArXiv:1701.08118v1*, 1–4. https://doi.org/10.17185/duepublico/42132

Rudas, C., Surányi, O., Yasseri, T., & Török, J. (2017). Understanding and coping with extremism in an online collaborative environment: A data-driven modeling. *PLoS ONE*, *12*(3), 1–16.

Runnymede Trust. (1997). *Islamophobia: a challenge for us all*. London.

Runnymede Trust. (2017). *Islamophobia: still a challenge for us all*. London.

Ruths, B. D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Ryan, M. E., & Leeson, P. T. (2011). Hate groups and hate crime. *International Review of Law and Economics*, *31*(4), 256–262. https://doi.org/10.1016/j.irle.2011.08.004

Rydgren, J. (2010). Radical right-wing populism in Demark and Sweden: explaining party system change and stability. *SAIS Review*, *1*(Winter-Spring), 57–71.

Sadek, N. (2017). Islamophobia, shame, and the collapse of Muslim identities. *International Journal of Applied Psychoanalysis Studies*, *14*(1), 200–221. https://doi.org/10.1002/aps.1534

Sadock, B. J., Sadock, V. A., & Ruiz, P. (2014). *Kaplan & Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry, 11th Edition*. New York: Lippincott Williams & Wilkins.

Saeed, A. (2007). Media, Racism and Islamophobia: The Representation of Islam and Muslims in the Media. *Sociology Compass*, *2*(1), 443–462.

Said, E. (1978). *Orientalism*. New York: Pantheon Books.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *ArXiv:1709.10159v1*, 1–9. Retrieved from http://arxiv.org/abs/1709.10159

Salminen, J., Almerekhi, H., Milenkovi, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *12th International AAAI Conference on Web and Social Media (ICWSM)* (pp. 330–339).

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1158/0008-5472.CAN-11-3472

Sarma, A. Das, Nanongkai, D., Pandurangan, G., & Tetali, P. (2009). Efficient Distributed Random Walks with Applications. *ArXiv:0911.3195v2*.

Sartori, G. (1970). Concept Misinformation in Comparative Politics. *The American Political Science Review*, *64*(4), 1033–1053.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885–899. https://doi.org/10.1177/0038038509105420

Sayyid, S. (2010). Out of the Devil's dictionary. In S. Sayyid & A. Vakil (Eds.), *Thinking through Islamophobia: global perspectives*. London: Hurst & Co.

Sayyid, S. (2014). A Measure of Islamophobia. *Islamophobia Studies Journal*, *2*(1), 10–25.

Schiffer, S., & Wagner, C. (2011). Anti-Semitism and Islamophobia - new enemies, old patterns. *Race and Class*, *52*(3), 77–84. https://doi.org/10.1177/0306396810389927

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, (2012), 1–10. https://doi.org/10.18653/v1/W17-1101

Schumann, S., Klein, O., Douglas, K., & Hewstone, M. (2017). When is computer-mediated intergroup contact most promising? Examining the effect of out-group members' anonymity on prejudice. *Computers in Human Behavior*, *1*(1), 1–14.

Searle, J. R. (1969). *Speech acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press.

Seaver, N. (2018). What Should an Anthropology of Algorithms Do? *Cultural Anthropology*, *33*(3), 375–385. https://doi.org/10.14506/ca33.3.04

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Semati, M. (2010). Islamophobia, culture and race in the age of empire. *Cultural Studies*, *24*(2), 256–275. https://doi.org/10.1080/09502380903541696

Severs, G. J. (2017). The 'obnoxious mobilised minority': homophobia and homohysteria in the British National Party, 1982 – 1999. *Gender and Education*, *29*(2), 165–181. https://doi.org/10.1080/09540253.2016.1274384

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods and Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820

Shapiro, I. (2002). Problems, methods, and theories in the study of politics, or: what 's wrong with political science and what to do about it. *Political Theory*, *30*(4), 596–619. https://doi.org/10.1177/0090591702030004008

Sheldon, P., & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior*, *58*(1), 89–97. https://doi.org/10.1016/j.chb.2015.12.059

Shmueli, G. (2009). To Explain or To Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.2139/ssrn.1351252

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, *12*(3), 248–279. https://doi.org/10.1177/1088868308319226

Siddiqui, K. (1992). *the Muslim Manifesto: a strategy for survival*. London.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the Targets of Hate in Online Social Media. *ArXiv:1603.07709v1*, 1–4. Retrieved from http://arxiv.org/abs/1603.07709

Sinno, A. I. I., & Tatari, E. (2009). Muslims in UK Institutions: Effective Representation or Tokenism? In *Muslims in Western Politics* (pp. 113–134). Indianapolis: University of Indiana Press.

Skopek, J. M. (2014). Anonymity, the production of goods, and institutional design. *Fordham Law Review*, *82*(4), 1751–1809. Retrieved from http://fordhamlawreview.org/wp-content/uploads/assets/pdfs/Vol_82/Skopek_March.pdf

Snijders, A. T. A. B. (2005). Fixed and random effects. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioural Science* (Vol. 2, pp. 664–665). New York: Wiley.

Snijders, C., Matzat, U., & Reips, U.-D. (2012). "Big data": Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science*, *7*(1), 1–5. https://doi.org/10.3923/ijds.2012.1.10

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Soltani, S., Amin, S., Seno, H., Nezhadkamali, M., Budirato, R., & Author, S. (2014). A Survey On Real World Botnets And Detection Mechanisms. *International Journal of Information & Network Security (IJINS)*, *3*(2), 116–127. Retrieved from http://iaesjournal.com/online/index.php/

Song, X., Xia, Y., & Zhu, H. (2017). Hidden Markov latent variable models with multivariate longitudinal data. *Biometrics*, *73*(1), 313–323. https://doi.org/10.1016/j.trsl.2014.08.005.The

Sowles, S. J., Mcleary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-craft, E.

E., … Cavazos-rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image*, *24*(1), 137–144. https://doi.org/10.1016/j.bodyim.2018.01.001

Spanje, J. Van, & Vreese, C. de. (2015). The good, the bad and the voter: The impact of hate speech prosecution of a politician on electoral support for his party. *Party Politics*, *21*(1), 115–130. https://doi.org/10.1177/1354068812472553

Spedicato, G. A., Kang, T. S., Yalamanchi, S. B., & Bhargav, S. (2017). *The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R*. *CRAN*. Retrieved from ftp://sunsite2.icm.edu.pl/site/cran/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf

Sponholz, L. (2016). Islamophobic Hate Speech: What is the Point of Counter-Speech? The Case of Oriana Fallaci and The Rage and the Pride. *Journal of Muslim Minority Affairs*, *36*(4), 502–522. https://doi.org/10.1080/13602004.2016.1259054

Springs, J. A. (2015). A tale of two Islamophobias: the paradoxes of civic nationalism in contemporary Europe and the United States. *Soundings: An Interdisciplinary Journal*, *93*(3), 289–321.

Stanford. (2018). Stanford GloVe. Retrieved October 12, 2018, from https://nlp.stanford.edu/projects/glove/

Statham, P., & Geddes, A. (2017). Elites and the 'organised public': Who drives British immigration politics and in which direction? *West European Politics*, *29*(2), 248–269. https://doi.org/10.1080/01402380500512601

Steckler, A., McLeroy, K. R., Goodman, R. M., Bird, S. T., & McCormick, L. (1992). Toward integrating qualitative and quantitative methods: an introduction. *Health*

Education Quarterly, *19*(1), 1–8.

Stephan, W., Ybarra, O., & Bachman, G. (1999) Prejudice towards immigrants. *Journal of applied social psychology, 29(11),* 2221-2237.

Stewart, B. M., & Zhukov, Y. M. (2009). Use of force and civil–military relations in Russia: an automated content analysis. *Small Wars & Insurgencies*, *20*(2), 319–343. https://doi.org/10.1080/09592310902975455

Stockemer, D., & Barisione, M. (2017). The 'new' discourse of the Front National under Marine Le Pen: A slight change with a big impact. *European Journal of Communication*, *32*(2), 100–115. https://doi.org/10.1177/0267323116680132

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sunstein, C. (2001). *Echo Chambers*. New Jersey: Princeton University Press.

Szmania, S., & Fincher, P. (2017). Countering Violent Extremism Online and Offline. *Criminology and Public Policy*, *16*(1), 119–125. https://doi.org/10.1111/1745-9133.12267

Taras, R. (2012). *Xenophobia and Islamophobia in Europe*. Edinburgh: Edinburgh University Press.

Taras, R. (2013). 'Islamophobia never stands still': race, religion, and culture. *Ethnic and Racial Studies*, *36*(3), 417–433. https://doi.org/10.1080/01419870.2013.734388

Tausczik, Y., & Pennebaker, J. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Taylor, J., Peignon, M., & Chen, Y.-S. (2017). Surfacing contextual hate speech words

within social media. *ArXiv:1711.10093v1*. Retrieved from http://arxiv.org/abs/1711.10093

Tell Mama. (2015). *We Fear for our Lives: Offline and Online Experiences*. London.

Tell Mama. (2016). *Tell Mama annual report 2015: the geography of anti-Muslim hatred*. London.

Tell Mama. (2017). *Tell MAMA 2016 annual report - A Constructed Threat: Identity, Intolerance and the Impact of Anti-Muslim Hatred*. London.

Tell Mama. (2018a). *Beyond the incident: outcomes for victims of anti-Muslim prejudice*. London. https://doi.org/10.1053/j.jvca.2010.06.032

Tell Mama. (2018b). *Gendered: anti-Muslim hatred and Islamophobia*. London.

Tench, R., & Jones, B. (2015). Social media: the Wild West of CSR communications. *Social Responsibility Journal*, *11*(2), 290–305. https://doi.org/https://doi.org/10.1108/SRJ-12-2012-0157

The Economist. (2010, February 25). The data deluge. *The Economist*. Retrieved from https://www.economist.com/leaders/2010/02/25/the-data-deluge

The Economist. (2017). Advances in AI are used to sport signs of sexuality: machines that read faces are coming.

The EDL. (2018). English Defence League About. Retrieved December 12, 2018, from http://www.englishdefenceleague.org.uk

The Guardian. (2014, May 16). Nigel Farage's LBC interview - the key moments. *The Guardian*. Retrieved from https://www.theguardian.com/politics/2014/may/16/nigel-farage-lbc-interview-key-moments

The Guardian. (2008, Jul 18). Young Muslims to get citizenship lessons in bid to combat extremism. *The Guardian*. Retreived from https://www.theguardian.com/politics/2008/jul/18/education.terrorism

The Guardian. (2017, Oct 14). Government's new online hate crime hub given just £200,000. *The Guardian*. Retreived from https://www.theguardian.com/society/2017/oct/14/government-criticised-for-low-funding-level-to-tackle-online-hate

The Guardian. (2018, October 16). Hate crime surge linked to Brexit and 2017 terrorist attacks. *The Guardian*.

The Independent. (2018, June 11). Islamophobia "very widespread" in Conservative Party, says Baroness Warsi. *The Independent*. Retrieved from https://www.independent.co.uk/news/uk/politics/conservatives-islamophobia-tory-party-racism-baroness-warsi-a8394271.html

The UN. (2007). *Racism, racial discrimination, xenophobia and related forms of intolerance: follow-up to and implementation of the D'urban declaration and programme of action*. New York.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173.

Thieltges, A., Schmidt, F., & Hegelich, S. (2016). The devil's triangle: Ethical considerations on developing bot detection methods. *AAAI Spring Symposium - Technical Report*, *SS-16-01*, 253–257.

Tierney, K., Bevc, C., & Kuligowski, E. (2006). Metaphors matter: disater myths, media frames, and their consequences in Hurricane Katrina. *The Annals of the American*

Academy, (March), 57–81. https://doi.org/10.1177/0002716205285589

Topolski, A. (2018). Good Jew, bad Jew… good Muslim, bad Muslim: "managing" Europe's others. *Ethnic and Racial Studies*, *41*(4), 1–18. https://doi.org/10.1080/01419870.2018.1391402

Townsend, L., & Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. Aberdeen. https://doi.org/http://dx.doi.org.ez.sun.ac.za/10.4135/9781526413642

Treadwell, J., & Garland, J. (2011). Masculinity, marginalization and violence: A case study of the English Defence League. *British Journal of Criminology*, *51*(4), 621–634. https://doi.org/10.1093/bjc/azr027

Tsvetkova, M., Yasseri, T., Meyer, E. T., Pickering, J. B., Engen, V., Walland, P., … Bravos, G. (2015). Understanding Human-Machine Networks: A Cross-Disciplinary Survey, (645043). https://doi.org/10.1145/3039868

Tufekci, Z. (2013). "Not This One": Social Movements, the Attention Economy, and Microcelebrity Networked Activism. *American Behavioral Scientist*, *57*(7), 848–870. https://doi.org/10.1177/0002764213479369

Twitter. (2018). Twitter: Docs. Retrieved October 17, 2018, from https://developer.twitter.com/en/docs.html

Twitter. (2019). Twitter: Hateful conduct policy. Retrieved May 30, 2019, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Uprichard, E. (2013). Big data, little questions? *Discover Society*, pp. 1–8. https://doi.org/10.1016/B978-0-12-027905-0.50010-7

Vaccari, C., & Valeriani, A. (2016). Party Campaigners or Citizen Campaigners? How Social Media Deepen and Broaden Party-Related Engagement. *International*

Journal of Press/Politics, *21*(3), 294–312. https://doi.org/10.1177/1940161216642152

Vaisey, S., & Miles, A. (2017). What You Can, and Can't, Do With Three-Wave Panel Data. *Sociological Methods and Research*, *46*(1), 44–67. https://doi.org/10.1177/0049124114547769

van Dijck, J. (2013). *The Culture of Connectivity: a Critical History Of Social Media*. Oxford: Oxford University Press.

van Dijck, J. (2017). Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief. *Surveillance & Society*, *12*(2), 197–208.

van Der Brug, W., Fennema, M., & Tillie, J. (2000). Anti-immigrant parties in Europe: Ideological or protest vote? *European Journal of Political Research*, *37*(1), 77–102. https://doi.org/10.1111/1475-6765.00505

van Deursen, A. J. A. M., & van Dijk, J. A. G. M. (2014). The digital divide shifts to differences in usage. *New Media and Society*, *16*(3), 507–526. https://doi.org/10.1177/1461444813487959

van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *ArXiv:1703.03107v2*, 1–11. Retrieved from http://arxiv.org/abs/1703.03107

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., & Ã, J. H. M. (2007). A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction. *Genetic Epidemiology*, *31*(1), 306–315. https://doi.org/10.1002/gepi

Veugelers, J., & Magnan, A. (2005). Conditions of far-right strength in contemporary Western Europe: an application of Kitschelt's theory. *European Journal of Political Research*, *44*(1), 837–860.

Vidgen, B. (2017, December 9). Mythbusting Britain First - 5 things you need to know. *Open Democracy*. Retrieved from https://www.opendemocracy.net/uk/bertie-vidgen/mythbusting-britain-first-5-things-you-need-to-know

Vidgen, B., & Yasseri, T. (2016). P-values: misunderstood and misused. *Frontiers in Physics*, *4*(6), 1–5. https://doi.org/10.3389/fphy.2016.00006

Vidgen, B., & Yasseri, T. (2018a). Detecting weak and strong Islamophobic hate speech on social media. *ArXiv:1812.10400*, 1–6.

Vidgen, B., & Yasseri, T. (2018b, January 2). Detecting weak and strong Islamophobia on social media. *The Conversation*. Retrieved from https://theconversation.com/how-we-built-a-tool-that-detects-the-strength-of-islamophobic-hate-speech-on-twitter-102536

Vieten, U. M., & Poynting, S. (2016). Contemporary Far-Right Racist Populism in Europe. *Journal of Intercultural Studies*, *37*(6), 533–540. https://doi.org/10.1080/07256868.2016.1235099

Virkar, Y., & Clauset, A. (2014). Power-law distributions in binned empirical data. *Annals of Applied Statistics*, *8*(1), 89–119. https://doi.org/10.1214/13-AOAS710

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 939–951. https://doi.org/10.1145/2818048.2820078

Vogel, T., & Wanke, M. (2016). *Attitudes and attitude change*. London: Psychology Press.

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2015). Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. *ArXiv:1509.01692v4*, 1–12. Retrieved from http://arxiv.org/abs/1509.01692

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the Gdpr. *Harvard Journal of Law & Technology*, *31*(2), 1–52. https://doi.org/10.2139/ssrn.3063289

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, *27*(4), 299–309. https://doi.org/10.1046/j.1365-2710.2002.00430.x

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets. *ArXiv:1606.00930v1*, (2014), 1–36. Retrieved from http://arxiv.org/abs/1606.00930

Wall Street Journal. (2014, April 11). Report: 44% of Twitter Accounts Have Never Sent a Tweet. *The Wall Street Journal*. Retrieved from https://blogs.wsj.com/digits/2014/04/11/new-data-quantifies-dearth-of-tweeters-on-twitter/

Walther, J. B., Hoter, E., Ganayem, A., & Shonfeld, M. (2015). Computers in Human Behavior Computer-mediated communication and the reduction of prejudice: A controlled longitudinal field experiment among Jews and Arabs in Israel. *Computers in Human Behavior*, *52*(1), 550–558. https://doi.org/10.1016/j.chb.2014.08.004

Wang, T. (2013). Big Data Needs Thick Data. *Ethnographpy Matters*.

Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceeding LSM '12 Proceedings of the Second Workshop on Language in Social Media*, (Lsm), 19–26. Retrieved from http://dl.acm.org/citation.cfm?id=2390374.2390377

Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, *7*(8), 882–890. https://doi.org/10.1111/2041-210X.12552

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. https://doi.org/10.18653/v1/W16-5618

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *ArXiv:1705.09899v2*, 1–17. Retrieved from http://arxiv.org/abs/1705.09899

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New Media and Society*, *20*(5), 1813–1831. https://doi.org/10.1177/1461444817707349

Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*(7127), 489–489. https://doi.org/10.1038/445489a

Webb, P., & Bale, T. (2014). Why Do Tories Defect to UKIP ? Conservative Party Members and the Temptations of the. *Political Studies*, *62*(1), 961–970. https://doi.org/10.1111/1467-9248.12130

Weber, M. (2017). *Methodology of Social Sciences [collected edition]*. (E. A. Shills & H. FInch, Eds.). New York: Routledge.

Webster, P. J., Holland, G. J., Curry, J. A., & Chang, H. (2005). Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment. *Science*, *309*(September), 1844–1846.

Wesley, J. (2010). Qualitative document analysis in political science. *T2PP Workshop, Vrije Universiteit Amstardam*, (April), 9–10. Retrieved from http://www2.let.vu.nl/oz/cltl/t2pp/docs/ws2010/papers/P2-Wesley.pdf

West, G. B., Brown, J. H., & Enquist, B. J. (1999). The Fourth Dimension of Life: Fractal Geometry and Allometric Scaling of Organisms. *Science*, *284*(5420), 1677–1679. https://doi.org/10.1126/science.284.5420.1677

White, F. A., Harvey, L. J., & Abu-rayya, H. M. (2015). Improving Intergroup Relations in the Internet Age: A Critical Review. *Review of General Psychology*, *19*(2), 129–139.

Williams, M. (2019). *Online Hate Speech Report*. Forthcoming.

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., & Burnap, P. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views , Online Context and Algorithmic Estimation. *Sociology*, *1*(1), 1–20. https://doi.org/10.1177/0038038517708140

Williamson, M., & Khiabany, G. (2010). UK: The veil and the politics of racism. *Race and Class*, *52*(2), 85–96. https://doi.org/10.1177/0306396810377003

Witten, I., Frank, E., & Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques*. London: Morgan Kaufmann.

Wodak, R. (2016). *The Politics of Fear: what right wing discourses mean*. London: SAGE.

Wojcieszak, M. (2008). False consensus goes online: impact of ideologically homogeneous groups on false consensus. *Political Opinion Quarterly*, *4*(4), 781–791. https://doi.org/10.1093/poq/nfn056

Wojcieszak, M. (2010). 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media and Society*, *12*(4), 637–655. https://doi.org/10.1177/1461444809342775

Wojcieszak, M. E., & Mutz, D. C. (2009). Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement? *Journal of Communication*, *59*(1), 40–56. https://doi.org/10.1111/j.1460-2466.2008.01403.x

Wood, C., & Finlay, W. M. L. (2008). British National Party representations of Muslims in the month after the London bombings: homogeneity, threat, and the conspiracy tradition. *British Journal of Social Psychology*, *47*(4), 707–726. https://doi.org/10.1348/014466607X264103

Woolley, S. C., & Guilbeault, D. R. (2017). Computational Propaganda in the United

States of America : Manufacturing Consensus Online. *Computational Propaganda Research Project*, *1*(5), 1–29. Retrieved from http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-USA.pdf

Wright, H. (2014). YBMs: religious identity and consumption amoung young British Muslims. *International Journal of Market Research*, *57*(1), 151–163.

Wright, S., Graham, T., & Jackson, D. (2017). Third Space and Everyday Online Political Talk: Deliberation, Polarisation, Avoidance. In *The 67th Annual Conference of the International Communication Association* (pp. 1–18). San Diego, California. Retrieved from http://eprints.whiterose.ac.uk/119308/%0Ahttp://eprints.whiterose.ac.uk/

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data Mining with Big Data. *Knowledge and Data Engineering, IEEE Transactions*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Wulczyn, E., Thain, N., & Dixon, L. (2016). Ex Machina: Personal Attacks Seen at Scale. In *WWW 2017 - IW3C2* (pp. 1391–1399). Perth, Australia. https://doi.org/10.1145/3038912.3052591

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology : Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yasseri, T., Sumi, R., & Kertesz, J. (2012). Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE*, *7*(1), 1–8. https://doi.org/10.1371/Citation

Yasseri, T., & Vidgen, B. (2016, March 7). P-values are widely used in the social sciences, but often misunderstood: and that's a problem. *Policy & Internet Blog*.

Retrieved from https://blogs.oii.ox.ac.uk/policy/many-of-us-scientists-dont-understand-p-values-and-thats-a-problem/

Yuan, L., B, Z. H., Chen, L., & Cai, Q. (2016). Clustering categorical sequences with variable-length tuples representation. In F. Lehner & N. Fteimi (Eds.), *Knowledge Science, Engineering and Management 9th International Conference* (Vol. 214, pp. 15–27). Passau, Germany. https://doi.org/10.1007/978-3-642-54930-4

Zafarani, R., & Liu, H. (2013). Connecting Users across Social Media Sites: A Behavioral-Modeling Approach. In *KDD '13, August 11-14*. Chicago, USA.

Zaghloul, W., Lee, S. M., & Trimi, S. (2009). Text classification: neural networks vs support vector machines. *Industrial Management & Data Systems*, *109*(5), 708–717.

Zannettou, S., Caulfield, T., Cristofaro, E. De, Kourtellis, N., Leontiadis, I., Sirivianos, M., … Blackburn, J. (2017). The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. *ArXiv:1705.06947v2*.

Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., Hansen, B., & Merkle, E. C. (2015). *Package "strucchange" in R. R CRAN Repository Report*. London.

Zempi, I., & Chakraborti, N. (2015). "They Make Us Feel Like We're a Virus": The Multiple Impacts of Islamophobic Hostility Towards Veiled Muslim Women. *International Journal for Crime, Justice and Social Democracy*, *4*(3), 44–56. https://doi.org/10.5204/ijcjsd.v3i2.236

Zhang, M., & Zhou, Z. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837.

Zhang, Y., Wells, C., Wang, S., & Rohe, K. (2017). Attention and amplification in the hybrid media system: The composition and activity of Donald Trump's Twitter

following during the 2016 presidential election. *New Media and Society*, *0*(0), 1–22. https://doi.org/10.1177/1461444817744390

Zhang, Y., & Wildemuth, B. M. (2009). Qualitative Analysis of Content. *Applications of Social Research Methods to Questions in Information and Library Science*, *1*(1), 421. https://doi.org/10.1002/hbm.20661

Zimmer, M. (2010). ‘‘But the data is already public": on the ethics of research in Facebook. *Ethics of Information Technology*, *12*(1), 313–325. https://doi.org/10.1007/s10676-010-9227-5

Zimmer, M., & Proferes, J. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. https://doi.org/http://dx.doi.org/10.1108/MRR-09-2015-0216

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015). Integrating and Evaluating Neural Word Embeddings in Information Retrieval. *Proceedings of the 20th Australasian Document Computing Symposium on ZZZ - ADCS ’15*, 1–8. https://doi.org/10.1145/2838931.2838936

Zúquete, J. P. (2008). The European extreme-right and Islam: New directions? The European extreme-right and Islam: New directions? *Journal of Political Ideologies*, *13*(3), 321–344. https://doi.org/10.1080/13569310802377019