EUROPOL

The Hague, 31/05/2017

Computer support to analyze IS propaganda

This paper was presented at the 1st European Counter Terrorism Centre (ECTC) conference on online terrorist propaganda, 10-11 April 2017, at Europol Headquarters, The Hague.
The views expressed are the authors' own and do not necessarily represent those of Europol.

Authors: Lisa Kaati, Magnus Sahlgren,
Tim Isbister, Babak Toghiani-Rizi and
Katie Cohen
Swedish Defence Research Agency (FOI)

Abstract

This paper discusses the use of computer support to analyze IS propaganda. We focus on thematic analysis of the textual content of various types of online propaganda, and investigate two different ways to identify themes in the data. One approach is theory-driven and builds on previous analysis of IS propaganda, and the other approach is data-driven and automatically identifies prevalent themes in the data. We exemplify the various approaches using different types of IS propaganda and we also provide examples of how the computer-assisted analysis can be used to analyze variation in the data over time. Both approaches indicate an increase in more violent themes in more recent IS propaganda.

1. Introduction

The Islamic State (IS) has realized the value of the Internet for spreading propaganda. "Online jihad" is by now a well-established phenomenon that has spread at a fast pace over the past fifteen years [1], with professionally edited propaganda material as well as user-generated content being distributed both on traditional websites and on social media.

The content and narratives of IS propaganda are useful tools to understand the motivating factors for IS supporters, as well as the development of the terrorist organization in itself. Analysis of the content of terrorist propaganda can also provide insights into if and how counter-narratives can and should be used. Previous work in this area has mostly focused on manually analyzing small segments of IS propaganda; Winters [2] studies 1,146 propaganda events (released during 17 July to 15 August 2015), Zelin [3] analyzes IS propaganda during one sample week (from April 18 to April 24, 2015, with a total of 123 different media releases), and Ingram [4] analyzes nine issues of the Dabiq magazine. Common for these previous studies is that they rely on manual analysis of comparatively small amounts of data from a limited time period.

In this work, we investigate the use of computer-assisted analysis of bigger amounts of data from a longer time period. In particular, we focus on the use of thematic analysis of the textual content of IS propaganda[1]. We study two different

[1] We acknowledge the fact that IS propaganda contains a significant amount of imagery (both still and moving), and that analysis of such material can provide useful information, but we focus here exclusively on the textual content.

approaches to thematic analysis; a theory-driven approach, in which the themes are provided a priori based on previous work, and a data-driven approach in which the themes are extracted directly from the data. We exemplify the different approaches using a variety of IS-related material, including the Dabiq and Rumiyah magazines, a selection of pro-IS Twitter accounts, and a selection of pro-IS blogs. We also include examples of how the thematic analysis can be used to monitor the development of themes in the data over time.

2. Previous work

Berger [5] states that "The Islamic State uses social media to activate a sense of "apocalyptic time" among its supporters online." Berger defines two different important themes that he considers when reasoning about IS social media output: apocalyptic and millenarian. Apocalyptic communication is defined as communication concerned with the imminent end or complete and radical transformation of the world. Millenarian is defined as communication concerned with the creation of a perfect society that will transform the world and establish a utopian reign on earth.

In the study by Winter [2], 1,146 propaganda events are manually coded into seven different narratives: brutality, victimhood, mercy, belonging, war and utopia. The brutality narrative is an important component of IS propaganda, including documentations in high resolution photos and videos of executions and beheadings. The victimhood narrative is closely related to brutality, but with the aim to spread the perception of an alleged global war on Sunni Islam. The mercy narrative presents a choice for populations that are under attack by IS that they can either resist and be killed, or repent, submit and be rewarded with mercy. The belonging narrative is used to attract new recruits to the organization. Belonging is characterized by depictions of people relaxing and enjoying themselves together in the caliphate. In the war narrative, the focus is on the organization's military gains by communicating about military parades and the frontline. Finally, the utopia narrative describes the Islamic State's establishment and implementation of a perceived Islamic utopia – the caliphate. In Winter's study, pictures make up 78 percent of the material while only 11 percent of the material is in written form.

In the study by Zelin [3], the propaganda is manually coded into eight variables: date, wilayat (province), country, city/village/region, media center, language(s), medium, and types. Zelin also provides a classification of the media releases into eleven identified topics. The topics are military, governance, da'wa (Islamic

missionary), hisba (control and enforcement of Islamic principles), promotion of the Caliphate, enemy attack, news, martyrdom, execution, denying enemy reports and other. Among these topics, Zelin identifies the six most important topics as: military, governance, da'wa, hisba, promotion of the caliphate, and enemy attack. Zelin concludes that the one week sample that he examined is not fully representative of everything IS releases over longer periods of time and that his study may serve as inspiration for other researchers to take a more holistic look at IS media over time. Ingram's [4] study of Dabiq aims to explore the strategic logic of IS's communications campaign targeting Western Muslim. In his analysis, Ingram identifies three different types of items in Dabiq: articles, statements and advertisements. An article is a longer written piece, a statement a short written piece (several sentences to three paragraphs in length) while the advertisements consists of short statements or excerpts from Islamic texts accompanied by colorful imagery. For each identified item, Ingram evaluates the "primary focus" of the item and then groups the items according to whether the item can be seen as a value-, dichotomy- or crisis-reinforcing message. Ingram's analysis shows that IS prioritizes dichotomy-reinforcing messages with solution and crisis narratives.

3. Thematic analysis

The previous studies mentioned above perform manual analysis to identify the main thematic content of the material. We take our starting point in themes identified in these previous studies and investigate the use of computer support to construct, refine, and maintain dictionaries representing the various themes, which are then analyzed with respect to their distribution over bigger samples of IS propaganda. We also investigate the use of a completely data-driven approach in which the themes are extracted from the data itself, without any reliance on a priori theories. For both approaches, we adhere to the standard way of operationalizing thematic analysis by developing dictionaries containing words that represent the themes in question, and then counting the relative frequencies of such words in the text material to be analyzed. The relative frequencies of the dictionary words are then averaged, producing an aggregated score for each theme that represent its relative frequency of occurrence in the data. This gives an indication of which themes are more prevalent in the data, and how they develop over time (assuming that the data has a temporal dimension).

There are several possible critiques of this approach. One is that the meaning of words can be context-dependent, which means that words may have several different meanings depending on the context, something that was noticed in for

example [6]. One example of a context-dependent word “execute”, which can be used in the meaning “to kill”, but also in the meaning “to carry out a task”. Counting only the frequency of occurrence of “execute” with no regards to the context thus risks leading to an inaccurate analysis that over-estimates the significance of the theme. Another criticism of dictionary-based (or keyword-based) analysis is that the dictionaries are often defined a priori, without any consideration of the properties of the actual data. This risks introducing bias, and it also risks making the analysis sensitive to vocabulary variation, which is introduced by slang words, different spellings, and domain specific terminology. Inability to handle such variation risks under-estimating the significance of a theme, which thus is the inverse problem compared to context-sensitivity. One example of domain specific terminology is the word “monkey”, which has a completely different meaning on an immigrant critic right-wing discussion forum than in a biology forum.

Context-sensitivity is a difficult problem for which there are no established solutions. In the field of natural language processing, techniques for word-sense disambiguation [7] based on various types of machine learning or various types of lexical resources have been suggested, but they either require significant amounts of manually-labeled training data or lexical resources of significant size and domain-relevance, which makes them extremely time-consuming and costly to apply in practice. The problem of vocabulary variation is arguably more approachable, in particular in light of the recent developments in distributional semantics [8] and representation learning, where semantic technologies and (unsupervised) machine learning algorithms are used to learn task- and domain-specific terminology that can be included (semi-) automatically in the dictionaries (see further Section 3.2). The most important take-away when working with dictionary-based analysis is to include experts with significant domain knowledge of the environments that are studied.

Although the samples of IS propaganda used in this study is bigger than what has been previously used, we are still only analyzing a small fraction of available IS propaganda. However, the methods and techniques described here are generic, and can be applied to other data sources as well. We exemplify the application of the proposed techniques on English data, but the same approach can be extended to operate on other languages.

3.1 Theory-driven approach

In the theory-driven approach, the idea is to build on theories from previous research in psychology and social science, and to use computer-assisted methods to

facilitate the analysis of bigger amounts of data. In the case of analyzing the thematic content of IS propaganda, such theories could for example be the themes presented by Winter [2] and Zelin [3]. We use the themes that Winter identified in his work [2,9]. As mentioned before, Winter identified six different themes in IS propaganda: brutality, mercy, belonging, victimhood, war and utopia. We have excluded mercy as a theme in our analysis, since only a very small part of the material (0.45%) that Winter analyzed included the mercy theme. For each of Winter's remaining themes, a dictionary with words representing the expression of the theme was created. The dictionaries were built from words occurring in Winter's descriptions of the themes. In a second step, the dictionaries were augmented with more words that were extracted from relevant data using two types of distributional semantic models, one which is developed in-house, and one that is developed by Facebook called FastText [10]. Each of the words suggested by the distributional semantic models was manually verified by experts before inclusion in the dictionary. An example of what the resulting dictionaries can look like is presented in Table 1.

Table 1. *Winter's themes and some sample words from the dictionaries representing each theme.*

Themes	Sample words
Brutality	beheaded, execute, punish, warning, admonish
Belonging	brotherhood, friendship, eid, collective, together, uniting
Victimhood	losses, survival, victim, victimhood, innocent, civilian
War	battlefield, operation, sniper, infiltration, survival, target
Utopia	culture, harvest, plenty, industry, agriculture, health care

3.2 Data-driven approach

In the data-driven approach, the idea is to extract the themes directly from the data, without reliance on any predefined theory, and to use these data-driven themes to perform the same type of count-based analysis as with the theory-driven themes. In computational approaches to text analysis, the arguably most common way to perform data-driven analysis of the thematic content of text data is to use a topic model [11, 12] which employs the document structure of the data to infer a probabilistic distribution of latent factors (i.e. topics) over the vocabulary. The result of using a topic model is a predetermined number of word lists that contain words that are representative of the various topics. The number of topics and representative words in each topic is normally on the order of ten to twenty. Topic

models are theoretically well understood, and they are thoroughly used both in computer science and in social science as a tool for thematic analysis. However, there are certain conditions in which topic models are less suitable. One is when data is not clearly formatted in topically coherent pieces of text (i.e. what we call "documents"); data may for example arrive in streaming format, or in chunks larger or smaller than what can be described as topical units. This is particularly common when dealing with web data, which can range from very short and terse formulations (typical in e.g. Twitter and Facebook status updates) to very long and verbose essays that cover a wide range of independent topics (more typical in blogs and traditional websites). Loss of document structure may also occur when converting data to text from other formats such as pdf or images.

An alternative to traditional topic models in such conditions is the use of distributional semantic models (also known as word embeddings), which are models that encode words as vectors whose relative directions indicate semantic similarity [13]. Such models do not rely on the data being structured into documents, but instead only considers the local context (which is often defined as simply the nearest surrounding words) around each word. The contextual information is aggregated over the data, resulting in one vector for each word that represents that word's contextual profile. Words that often occur with the same contexts will end up with similar vectors, leading to a representation that embodies Firth's famous dictum "you shall know a word by the company it keeps" [14]. By applying a clustering algorithm to the vector space defined by the word vectors, we can achieve a similar result as when using a topic model, even if the data lacks document structure.

4. Datasets

We consider three different digital sources of IS-related material in this study: tweets, magazines and blogs. All data used is publicly available and not protected using passwords. Each dataset is described below.

Tweets: The set of tweets we use in our analysis is from a dataset that was released by the data science community Kaggle [15] and contains 17,000 tweets from several pro-IS Twitter accounts. The tweets were published between 6 January 2015 and 13 May 2016.

Magazines: IS regularly publishes propaganda magazines in different languages. The magazines are editorially produced with glossy images and professional layout,

and they are printed in several languages including English. This study uses the English versions of Dabiq (fifteen issues) and Rumiyah (seven issues).

Blogs: The set of blogs consist of ten manually identified pro-IS blogs. The blogs are from different platforms and vary in size. The amount of multimedia content of the blogs varies, some of the blogs contain more images and pdf-files. We only consider the textual data and exclude images and pdf-files.

4.1 Preprocessing

Each dataset is subject to a pipeline of preprocessing steps that is required in order to prepare the data for computational analysis. The pipeline includes extracting the text content from the pdf-files (for the magazines), segmenting the text into sentences, tokenizing the text into words (which includes separating punctuation from the words), identifying common multiword terms and phrases, and identifying the names of persons and places. Each of these processing steps can be non-trivial, depending on the qualities of the data, and the final processed data should thus not be expected to be flawless. The frequent occurrence of transliteration of Arabic in this data further complicates the preprocessing, since it often uses non-standard Unicode characters, and there is a large degree of variation in the transliteration (there are many different ways to transliterate Arabic into Latin script).

5. Results

In this Section we present the results of our two different analysis approaches: the theory driven approach (using Winter's [2] themes) and the data-driven approach.

5.1 Winter's themes

In our first analysis we compare the presence of the different themes identified by Winter in the different datasets. Figure 1 shows the results of our analysis (the numbers are presented in Table 2). As can be seen in the figure, the themes war, brutality and victimhood are more present in all IS propaganda sources compared to the baseline. The themes belonging and utopia on the other hand is more common in the baseline than in the IS propaganda. A comparison of the presence of the themes in the different datasets can be found in Figure 2. The figures show the distribution of the different themes in relation to each other. Even though our

themes are inspired by Winter's work, our results cannot be directly compared with his. Winter classified propaganda outputs (mostly pictures) into different themes while we focus on measuring the presence of themes in text. In our analysis the theme war dominates the communication in all propaganda sources. It is mostly communicated in Rumiyah (70%), the blogs (71%) and the tweets (77%) and less in Dabiq (57%).

In Winter's study from 2015 [2] utopia was the most common narrative followed by war. However, in a recent study by Winter propaganda from one month in 2017 (Januari 31 - February 28) was studied, and Winter's analysis shows that the use of narratives has shifted and war is now the most used narrative followed by utopia. The results from Winter's manual analysis shows results similar to our analysis. This is perhaps not a surprise since most of the data we consider in our analysis are from late 2016 to early 2017.

Table 2. *The occurrences of words from each dictionary (in percent) using Winter's themes.*

	Belonging	Brutality	War	Utopia	Victimhood
Blogs	0.16	0.03	2.96	0.83	0.17
Tweets	0.09	0.06	3.44	0.64	0.24
Rumiyah	0.2	0.06	2.75	0.64	0.27
Dabiq	0.13	0.08	1.95	0.76	0.48

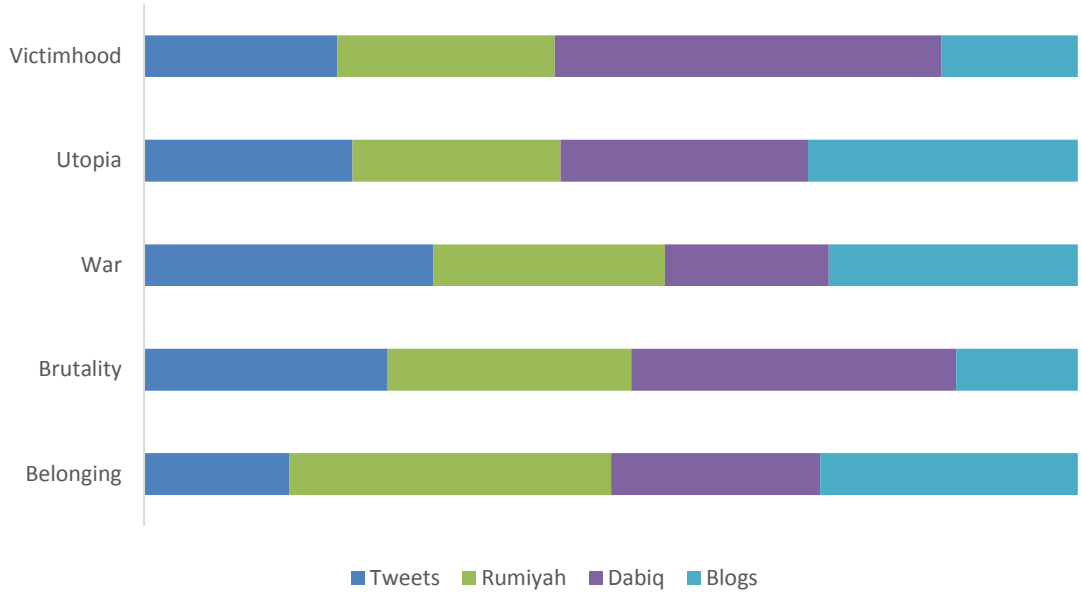

Figure 1. The presence of Winter's themes in the different datasets expressed in percent.

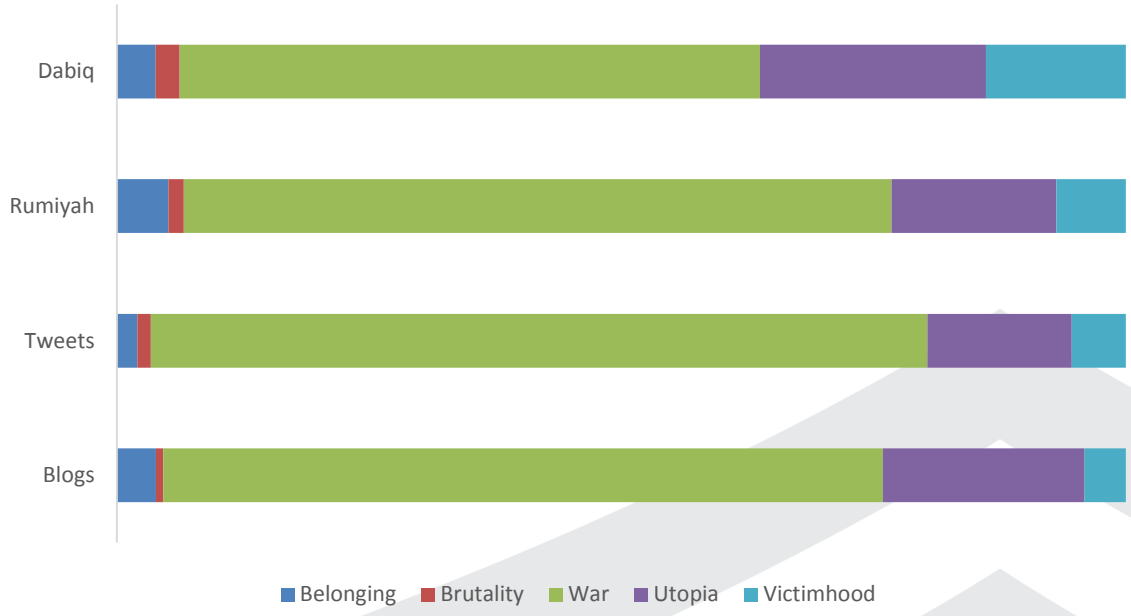

Figure 2. The distribution of Winter's themes in the different datasets.

A comparison between the different sources shows that the topic of belonging occurs to a greater extent in the blogs and journals (especially Rumiyah) than in the tweets. The theme war occurs most in the tweets. This could be explained by the fact that the tweets often consist of reports from battlefields, while the blogs and journals have a greater variety of subjects. Another possible explanation is the format. Longer texts may be better suited for descriptions of belonging and community. A third explanation may have to do with the target groups. At least blogs are largely aimed at women, who are not expected to participate in the armed battle.

The themes victimhood and brutality occur mostly in Dabiq (especially victimhood, brutality is almost as common on Twitter). Overall, the proportion of brutality was very small in all material. This despite the fact that extreme violence in the form of executions has become closely connected with IS in the West. The extremely violent propaganda is produced primarily in the form of films or pictures, which could explain that it represented such a small part of our analysis. However, even in Winter's analysis, mostly based on images, the proportion of brutality was small. Just because they are so extreme and shocking by nature, these images get more attention than, for example, the representations of daily life in the Caliphate, which actually represent a much larger part of propaganda. Extreme violence is reproduced to a greater extent by western media, and also makes a greater impression in the individual's consciousness [2].

5.2 Data-driven themes

In order to automatically identify prevalent themes in the data, we apply an in-house distributional semantic model to the pro-IS magazines (Dabiq and Rumiyah), and feed the resulting semantic representations into a clustering algorithm that automatically determines the number of clusters in the representation space. We average the clustering solution over multiple runs of the clustering algorithm by manually selecting and merging similar clusters. This is done in order to reduce the variance of the clustering solution, and to introduce at least a minimal amount of manual quality control. We select a subset of the themes automatically extracted in this analysis, which we label artillery, base (referring to various types of military installations), conquest, killing, knife, prohibition, slave, and vehicles. Each of these themes are treated as a dictionary, in the same way as the theory-driven themes that were manually identified. Examples of dictionary entries from each of the data-driven themes are presented in Table 3. As can be noted, most of the data-driven themes are concerned with military operations and war and should perhaps be seen as sub-categories to Winter's war narrative.

Table 3. *The data-driven themes that we have used in our analysis and some sample words.*

Theme	Example words
Death	killed, killing, wounded, wounding, injured, injuring, deaths, death
Explosive	explosive, explosives, detonated, belt, detonating, device, bomb
Family	daughter, girl, birth, born, wives, woman, marry, child, baby
Knife	cut, knife, knives, cutting, handle, blade, fixed, throat, stabbing, sharp
Prohibition	prohibition, fornication, drugs, ribā, fāhishah, cigarettes, sins
Slave	slave, master, concubine, servants, mastery, worshipper, slavegirl
Vehicles	bulldozer, vehicle, tanks, drive, armored, truck, hummers, 4wheel, car

Table 4. *The occurrences of words from the data-driven themes (in percent).*

	Death	Explosive	Family	Knife	Prohibition	Slave	Vehicles
Blogs	0.0522	0.0318	0.0133	0.0037	0.0022	0.0095	0.0382
Dabiq	0.0207	0.0082	0.0153	0.0068	0.0047	0.0140	0.0140
Rumiyah	0.0407	0.0153	0.0141	0.0082	0.0031	0.0135	0.0267
Tweets	0.0864	0.0283	0.0210	0.0093	0.0024	0.0031	0.0344

Figure 3 shows the relative occurrence of the different themes in the different datasets. The occurrence of each theme (in percent) is provided in Table 4. As can be seen in Table 4, the occurrence of the conquest theme dominates the analysis and the theme occurs in all different sources.

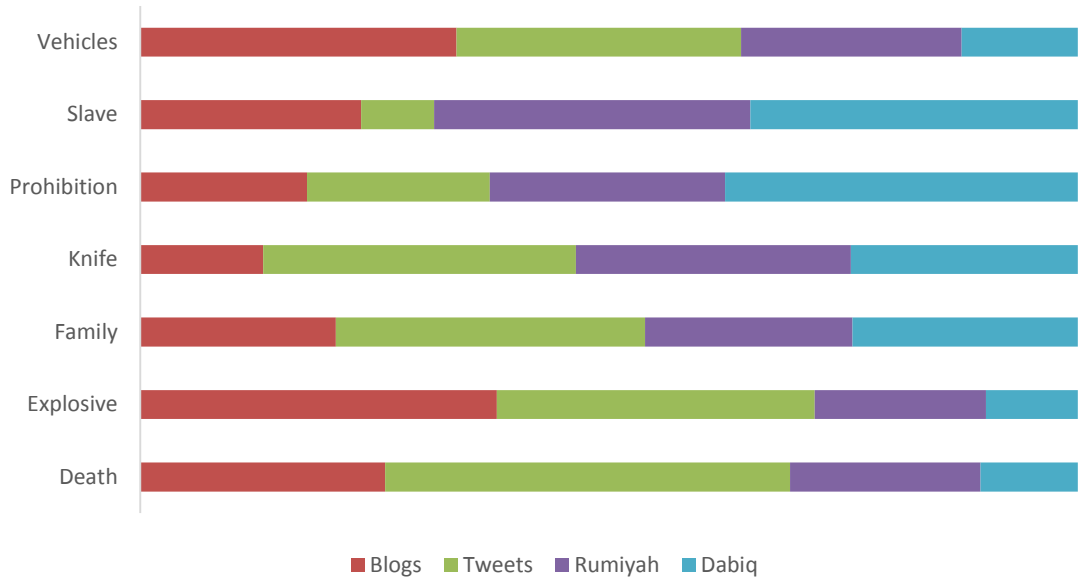

Figure 3. The presence of the data-driven themes in the different datasets expressed in percent.

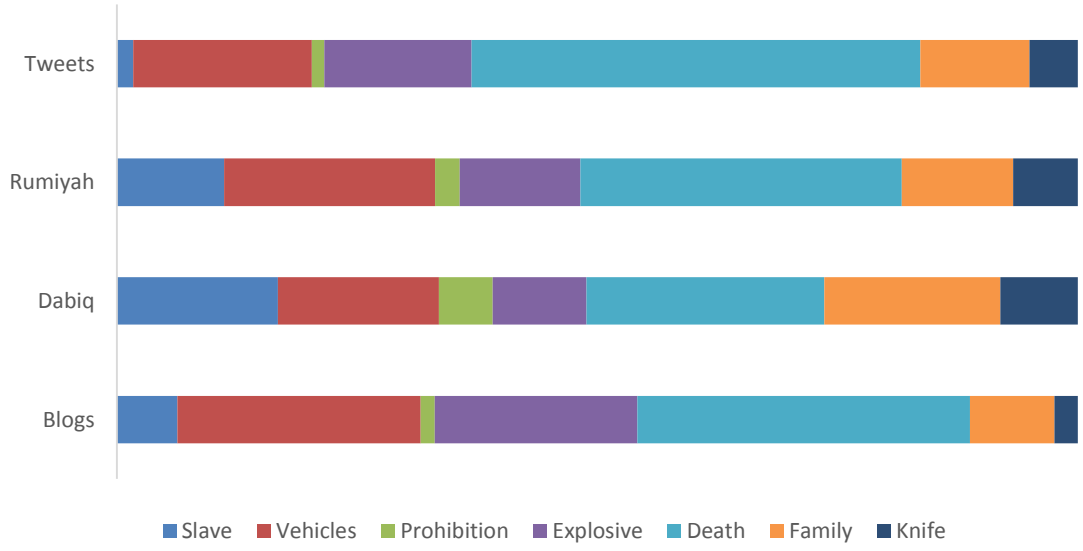

Figure 4. The distribution of the data-driven themes in the different IS propaganda datasets.

In Figure 4 we display the distribution of the themes in relation to each other. As can be seen in the figure, the topic death is dominant on Twitter, something that could be linked to the fact that IS propaganda on Twitter often contains reports

from battlefields, but perhaps also death has an important role for IS through martyrdom. Another notable difference was that explosives occur more in blogs than in any of the other sources. Knives, explosives and vehicles are common approaches used in recent terror attacks in Europe, and therefore future research should focus on analysing more closely the occurrence of these themes in different sources and what conclusions can be drawn from the results. It is not impossible that, for example, trends in terrorist attacks may be detected at an early stage of propaganda.

6. Variation over time

In this section, we provide a temporal analysis of the development of the various themes over time in the IS magazines Dabiq and Rumiyah. The reason for only performing this analysis on the magazines is that we rely on the assumption that the magazines are published in consecutive order over time. It would be of great interest to perform a similar analysis on the other data sources and compare how the occurrences of themes shifts.

We display the results in the form of stacked area charts, in which the proportion of each theme is represented by the respective colored area. To compare the frequency of different themes, we should look at the relative sizes of each theme's area in the chart. The *y*-axis represents the cumulated total; if one theme increases significantly for a particular data point, so will the cumulated total.

Figure 5 shows the development of Winter's themes over the different issues of the IS magazines Dabiq and Rumiyah. As can be seen in the figure, the theme war increases significantly in Rumiyah. The increase of the use of the narrative war in propaganda was also noticed by Winter in a recent study [16].

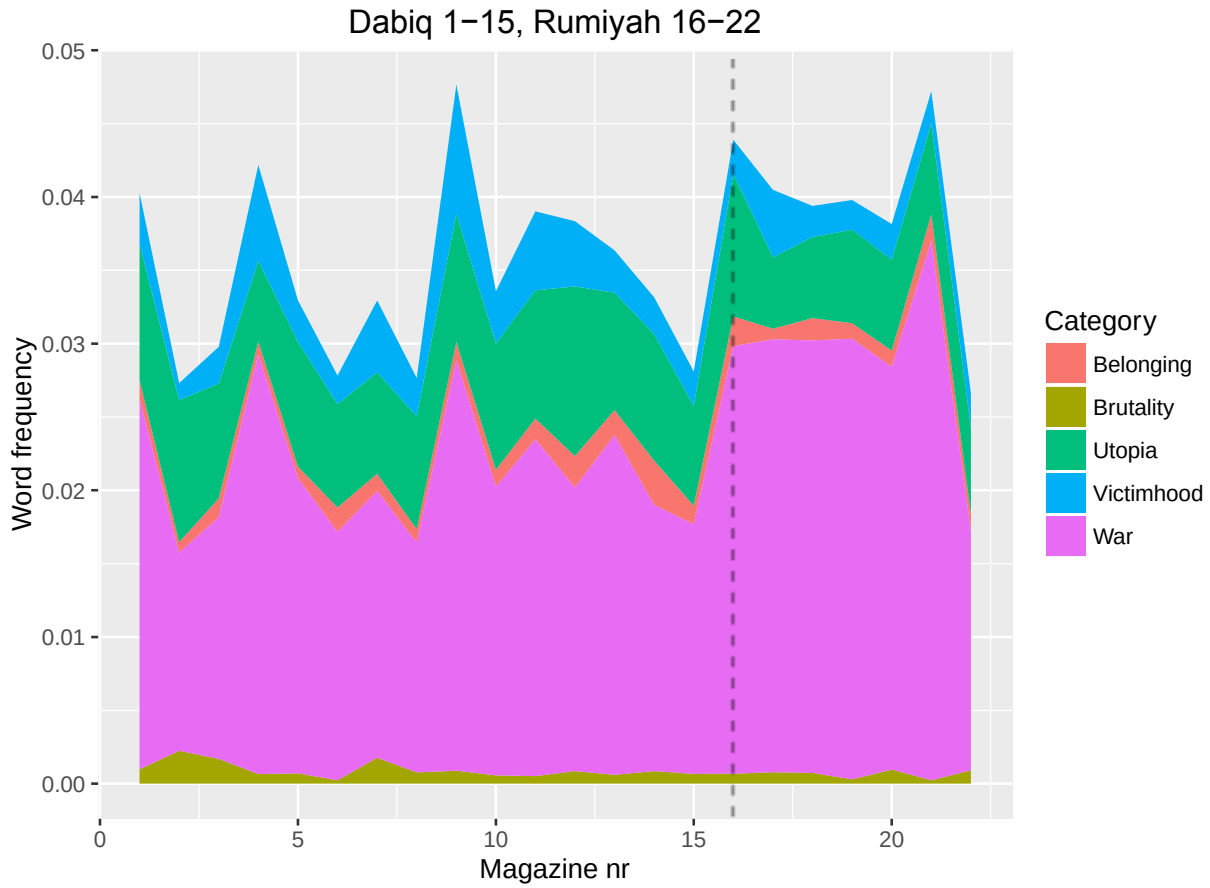

Figure 5. Distribution of Winter's themes over time in the IS magazines. The vertical dotted line indicates the change from Dabiq to Rumiyah.

Figure 6 shows the development of the data-driven themes over the different issues of Dabiq and Rumiyah. The most striking aspect of this analysis is the significant increase in most of the data-driven themes in the Rumiyah publications; the only themes that do not increase drastically is the slave and prohibition themes. This development seems consistent with Winter's analysis.

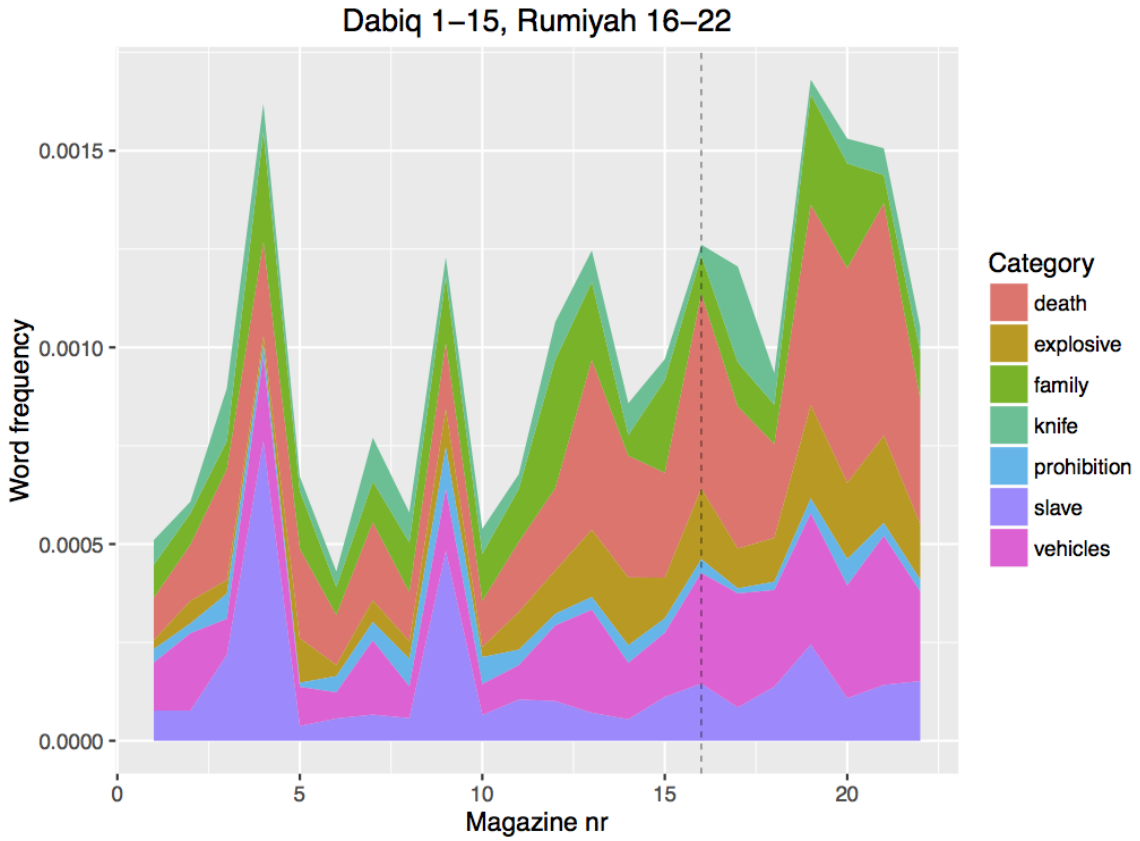

Figure 6. Distribution of the data-driven themes over time in the IS magazines. The vertical dotted line indicates the change from Dabiq to Rumiyah.

7. Discussion

In this study the aim has been to investigate how computer support can be used to analyze narratives in IS propaganda. Developing techniques and methods that can be used to analyze large amounts of propaganda is an important task, since the amount of propaganda that IS is releasing is enormous to the extent that it is impossible for humans to analyze. A deeper understanding of the different narratives in propaganda, how the narrative differs in different media sources, and how they shift over time are important knowledge when developing counter-narratives and counter-messaging strategies. Winter [9] formulates the need for analyzing the narratives in propaganda with the question "How, for example, can we be expected to develop a counter-narrative without knowing what narratives we are countering?".

The results of our analysis using Winter's themes shows similar results as Winter's [16] resent analysis where war is the most dominant narrative followed by utopia and victimhood. The use of narratives did not vary much between the different sources that we have studied - only small variations can be identified. One difference that was noticeable was that Dabiq had a higher use of utopia than the other sources. This could be a consequence of the fact that Dabiq is older than the rest of the sources. Winter [16] also notices a shift in the use of the narrative utopia in his comparison of propaganda from 2015 with propaganda from 2017. The data-driven themes mainly consist of themes that could be seen as sub-categories to Winter's war theme. It is therefore not surprising that almost all data-driven themes increase significantly in Rumiyah compared to Dabiq. Among the data-driven themes is an increase in the theme of death, but also in the themes knives, vehicles and explosives. The increase may be due to the fact that the first numbers of Rumiyah contains texts that are legitimizing terrorist attacks in other countries, including descriptions of how such attacks should be done. The increase of these themes also coincides with an actual increase in terrorist attacks where the approach has included vehicles, knives and explosives. However, more in-depth investigations are needed to determine the relationship between these changes and actual terrorist attacks – something that can be seen as a direction for future research.

The data-driven themes have a high variation in the different issues of the magazines. There is, for example, a significant increase in the theme slave in Dabiq number 4 and Dabiq number 9. This is due to the fact that these issues contain articles about slavery. In Dabiq number 4 there is an article that justifies having women as sex slaves and in Dabiq number 9 there is an article in which IS both acknowledges the existence of sex slaves and justifies slavery (in the same article, IS threatens to sell Michelle Obama, then first lady of the USA, as a sex slave).

Studies of how the themes of propaganda change over time are valuable for understanding how terrorist organizations change and develop over time. Computer-assisted analyses are useful for such studies because they quickly create an image of what is written in large amounts of text. As we have shown here, computer-aided analysis can be used both to find general trends, such as the rise of the theme of death, or temporary changes, such as slavery, for example. That the increase in the slavery theme can be identified in this way means that anyone who wants to form an idea of IS's view of slavery only needs to read Dabiq numbers 4 and 9 instead of going through all 15 issues of the journal. Thus, in this relatively small amount of data, there will be 90 percent less material to go through. Using the same approach to large amounts of data saves even more time and manpower.

There are always pros and cons with all kinds of analysis methods. One of the benefits with computerized methods is that large amounts of data can be analyzed and that it is easy to study changes over time. In this study we use a dictionary-based approach where we measure the presence of words from a dictionary. Each dictionary represents a theme or a narrative, we use both narratives identified by experts and data-driven themes that are automatically identified by the computer. However, computer support should only be seen as a help for conducting analysis - experts need to be included in all phases of the analysis.

This study is performed on a subset of IS propaganda. There are many interesting directions for future work. One obvious direction would be to include more data and to study how narratives in different sources vary over time.

References

[1] Prucha, N. Is and the jihadist information highway – projecting influence and religious identity via telegram. Perspectives on Terror. 10 (2016).

[2] Winter, C. Documenting the virtual 'caliphate'. Quilliam (2015).

[3] Zelin, A. Picture or it didn't happen: A snapshot of the islamic state's official media output. Perspectives on Terror. 9 (2015).

[4] Ingram, H. J. An analysis of islamic state's dabiq magazine. Aust. J. Polit. Sci. (2016).

[5] Berger, J. The metronome of apocalyptic time: Social media as carrier wave for millenarian contagion. Perspectives on Terror. 9 (2015).

[6] Mehl, M., Robbins, M. & Holleran, S. How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. J. Methods Meas. Soc. Sci. 3 (2013).

[7] Agirre, E. & Edmonds, P. Word Sense Disambiguation: Algorithms and Applications (Springer Publishing Company, Incorporated, 2007).

[8] Sahlgren, M. The distributional hypothesis. Italian J. Linguist. 20, 31–51 (2008).

[9] Winter, C. The virtual 'caliphate': Understanding islamic state's propaganda strategy. Quilliam (2015).

[10] Facebook. Fasttext. URL https://research.fb.com/projects/fasttext/.

[11] Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, 50–57 (ACM, New York, NY, USA, 1999).

[12] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).

[13] Turney, P. D. & Pantel, P. From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res. 37, 141–188 (2010). URL http://dl.acm.org/citation.cfm?id=1861751.1861756.

[14] Firth, J. R. A synopsis of linguistic theory 1930-55. In Studies in Linguistic Analysis (special volume of the Philological Society), vol. 1952-59, 1–32 (The Philological Society, Oxford, 1957).

[15] Kaggle. URL http://www.kaggle.com.

[16] Winter, C. The isis propaganda decline. ICSR Insight (2017).