"Like Sheep Among Wolves": Characterizing Hateful Users on Twitter

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, Wagner Meira Jr.
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
{manoelribeiro,pcalais,yurisantos,virgilio,meira}@dcc.ufmg.br

ABSTRACT

Hateful speech in Online Social Networks (OSNs) is a key challenge for companies and governments, as it impacts users and advertisers, and as several countries have strict legislation against the practice. This has motivated work on detecting and characterizing the phenomenon in tweets, social media posts and comments. However, these approaches face several shortcomings due to the noisiness of OSN data, the sparsity of the phenomenon, and the subjectivity of the definition of hate speech. This works presents a user-centric view of hate speech, paving the way for better detection methods and understanding. We collect a Twitter dataset of 100,386 users along with up to 200 tweets from their timelines with a random-walk-based crawler on the retweet graph, and select a subsample of 4,972 to be manually annotated as hateful or not through crowdsourcing. We examine the difference between user activity patterns, the content disseminated between hateful and normal users, and network centrality measurements in the sampled graph. Our results show that hateful users have more recent account creation dates, and more statuses, and followees per day. Additionally, they favorite more tweets, tweet in shorter intervals and are more central in the retweet network, contradicting the "lone wolf" stereotype often associated with such behavior. Hateful users are more negative, more profane, and use less words associated with topics such as hate, terrorism, violence and anger. We also identify similarities between hateful/normal users and their 1-neighborhood, suggesting strong homophily.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; **Empirical studies in collaborative and social computing**;

KEYWORDS

hate speech, online social networks, hateful users

ACM Reference Format:
Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, Wagner Meira Jr.. 2018. "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.475/123_4

MIS2, 2018, Marina Del Rey, CA, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

Hate speech can be defined as *"language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"* [8]. The importance of understanding the phenomenon in Online Social Networks (OSNs) is manifold. For example, countries such as Germany have strict legislation against the practice [34], the presence of such content may pose problems for advertisers [16] and users [30], and manually inspecting all possibly hateful content in OSNs is unfeasible [31]. Furthermore, the blurry line between banning such behavior from platforms and censoring dissenting opinions is a major societal issue [25].

This scenario has motivated a body of work that attempts to characterize and automatically detect such content [4, 10, 19, 21, 31, 37]. These create representations for tweets, posts or comments in an OSN, *e.g.* word2vec [24], and then classify content as hateful or not, often drawing insights on the nature of hateful speech on the granularity level of tweets or comments. However, in OSNs, the meaning of such content is often not self-contained, referring, for instance, to some event which just happened, and the texts are packed with informal language, spelling errors, special characters and sarcasm [9, 28]. Furthermore, hate speech itself is highly subjective, reliant on temporal, social and historical context, and occurs sparsely [31]. These problems, although observed, remain largely unaddressed [8, 21].

Fortunately, the data in posts, tweets or messages, are not the only signals we may use to study hate speech in OSNs. Most often, these signals are linked to a profile representing a person or institution. Characterizing and detecting hateful *users* shares much of the benefits of detecting hateful content and presents plenty of opportunities to explore a richer feature space. Twitter's guideline for hateful conduct captures this intuition, stating that *some Tweets may seem to be abusive when viewed in isolation, but may not be when viewed in the context of a larger conversation* [35].

Analyzing hateful *users* rather than *content* is also attractive because other dimensions may be explored, such as the user's activity and connections in the network. For example, in *Twitter*, it is possible to see the number of tweets, followers, and favorites a user has. It is also possible to extract influence links among users who retweet each other, analyzing them in a larger network of influences. This allows us to use network-based metrics, such as betweenness centrality [12] and also to analyze the neighborhood of such users. Noticeably, although several studies characterize hateful speech in text [8, 40], no study that the authors are aware of focuses on the dimension of hateful *users* in OSNs.

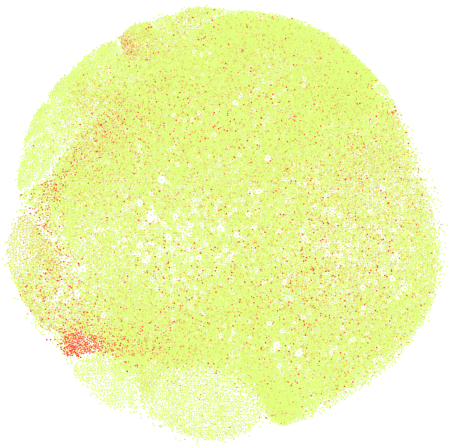

Figure 1: Network of 100, 386 users sampled from Twitter after our diffusion process. Red nodes indicate the proximity of users to those who employed words in our lexicon.

In this paper we focus on identifying and characterizing hateful users on Twitter, which we define in accordance with Twitter's hateful conduct guidelines [35]. We collect a dataset of 100, 386 users along with up to 200 tweets from their timelines with a random-walk-based crawler on Twitter's retweet-induced graph. We identify users that employed a set of hate speech related words, and generate a subsample selecting users that are in different "distances" to these to be manually annotated as hateful or not through crowdsourcing. This is explained in Section 3. We create a dataset containing 4, 972 manually annotated users, of which 544 were labeled as hateful. We ask the following research questions:

***Q1**: Are the attributes of and the content associated with hateful users different from normal ones?*

***Q2**: How are hateful users characterized in terms of their global position in the network and their local neighborhood of interactions?*

To address these questions, we perform experiments in our collected dataset. We *(i)* examine attributes provided by Twitter's API, such as number of followers and creation date as well as attributes related to users activity; *(ii)* perform a sentiment and lexical analysis on the content present in each user's timeline; and *(iii)* compare centrality measures such as betweenness and eigenvector centrality between hateful and normal users. We also examine these statistics for users in the 1-neighborhood on the retweet graph.

Our results show that hateful users tweet more and within smaller intervals, and favorite other tweets significantly more than the normal ones. They also are more negative according to lexicon-based sentiment analysis and use more swear words. Hateful users have follow more people per day than normal ones, and use vocabulary related to categories such as hate, anger, shame, violence and terrorism *less* frequently. Also, the median hateful user have higher network centrality according to several metrics, contradicting the "lone wolf" behavior often associated with the practice [3]. This analysis held similar results when we looked at the 1-neighborhood of hateful and normal users. Our code is available online [1].

[1] https://github.com/manoelhortaribeiro/AbusiveUsersOSNs

2 DEFINITIONS

Retweet-Induced Graph. We define the retweet-induced graph G as a directed graph $G = (V, E)$ where each node $u \in V$ represents a user in Twitter, and each edge $(u_1, u_2) \in E$ represents a retweet in the network, where the user u_1 has retweeted user u_2. Retweet-graphs have been largely used in the social network analysis, with previous work suggesting that retweets are better than followers to judge the influence of users [6]. Notice that influence flows in the opposite direction of retweets, and thus we actually work on the graph with inverted edges. Intuitively, given that a lot of people retweet u_i and u_i retweets nobody, u_i may still be a central and influential node.

Hateful Users. Defining which users are hateful is non-trivial as it derives from the definition of hateful speech, which is not widely agreed upon [32]. We choose to define hateful users in accordance to Twitter's hateful conduct guidelines, which state users *may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories* [35].

Offensive Language. Other concept we employ is that of offensive language, which has been shown to be correlated with hateful content [8]. While there doesn't exist a universal definition of offensive language, we employ Zerak et. al definition of explicit abusive language, which defines it as *language that is unambiguous in its potential to be abusive, for example language that contains racial or homophobic slurs* [39]. Importantly, the use of this kind of language does not necessarily imply hate speech.

3 DATA COLLECTION

Most existing work that detecthate speech on Twitter employ a lexicon-based data collection, which involves sampling only tweets that contain certain words [4, 8, 21, 40], such as `wetb*cks` of `fagg*t`. As we are trying to characterize hateful users, it would not be appropriate to rely solely on this technique, as we would get a sample heavily biased towards users who used these words. Furthermore this methodology presents problems even for dealing with the problem strictly on a tweet-based level. Some examples are:

- Statements may subtly disseminate hate with no offensive words, as in the sentence `"Who convinced Muslim girls they were pretty?"` [8, 31, 40]
- Hate groups may employ code words that are apparently benign, such as `"skypes"`, to reference minorities demeaningly, creating a truly adversarial setting [21, 23].

Thus, we employ a more elaborate data collection process, which involves collecting a sample of Twitter's English speaking users, selecting a subsample of these users to be annotated as hateful or not hateful, and, finally, annotating them using a crowdsourcing service. These are described in the upcoming paragraphs.

Sampling Twitter. As we do not have access to the full Twitter graph, we are faced with the challenge of obtaining a representative sample of it. Although there are several ways which users relate to each other in Twitter, we choose the retweet graph, in accordance

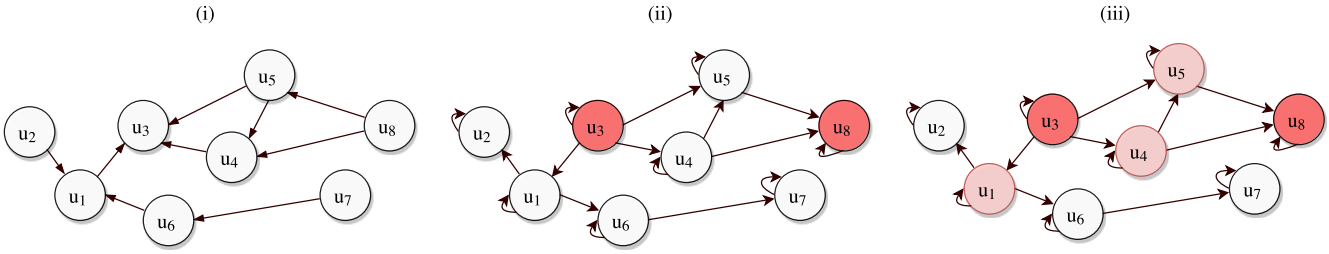

Figure 2: Depiction of our diffusion process. *(i)* We begin with graph G from the retweet-induced graph we sample from twitter, *(ii)* We revert the direction of the edges (as it is the way influence flows), add self loops to every node, and mark the users who employed one of the words in our lexicon, *(iii)* We iteractively update the belief of other nodes.

with existing literature [6]. Sampling the retweet-induced is hard as we can only observe out-coming edges, or in other words, given a user's timeline, we can obtain all users he or she retweeted, but not all users who retweeted them (due to API limitations). Furthermore, it is known that any unbiased in-degree estimation is impossible without sampling most of these "hidden" edges in the graph [26]. Acknowledging these limitations, we employ Ribeiro et al. Direct Unbiased Random Walk (*DURW*), algorithm, which constructs an undirected graph in real time and estimates out-degrees distribution efficiently by occasionally jumping to a random node in the undirected graph [27]. Fortunately, however, in the retweet graph the outcoming edges of a user represent the other users they (usually [17]) endorse. With this strategy, we collect 100,386 users and 2,286,592 retweet edges along with the 200 most recent tweets for each one of the users (including quotes, retweets and replies).

Selecting a Subsample to Annotate. After sampling Twitter, we are faced with the problem of selecting the subset of the data which will be annotated as hateful or not. If we choose the users uniformly at random, we risk having a very insignificant percentage of hate speech in the subsample. On the other hand, if we choose only users that use obvious hate speech related features, such as offensive racial slurs, we will bias our sample with only tweets with this language. In this case, for example, we would not capture codewords as the ones mentioned in Magu et. al [21]. We:

(1) Create a lexicon of words that are mostly used in the context of hate speech. This is unlike other work [8], as we don't consider words that are employed in a hateful context but often used in the everyday life in a harmless way (*e.g.* `n*gger`);
(2) Run a diffusion process on the graph based on DeGroot's Learning Model [15], assigning a initial belief $p_i^{(0)} = 1$ to each user u_i who employed the words in the lexicon;
(3) Divide the users in 4 strata according to their associated beliefs after the diffusion process, and perform a stratified sampling, obtaining up to 1500 user per strata.

We create our lexicon with words from Hatebase.org [1], and ADL's hate symbol database [20]. We choose words such as `holohoax`, `racial treason` and `white genocide` as they are less likely to be used in a non-hateful context. Furthermore, as we run the diffusion process later, we do not risk having a sample which is excessively small or biased towards some vocabulary. Notice that the difference here is that we use the lexicon as a starting point to select regions of the graph to be sampled, whereas other works sample directly through the lexicon [8, 19, 40].

We briefly present our diffusion model, as illustrated in Figure 2. Let A be the adjacency matrix of our retweeted induced graph $G = (V, E)$ where each node $u \in V$ represents a user and each edge $(u, v) \in E$ represents a retweet. We have that $A(u, v) = 1$ if u retweeted v. We create a transition matrix T by inverting the edges in A (as the influence flows from the retweeted user to the user who retweeted him or her), adding a self loop to each of the nodes and then normalizing each row in A so it sums to 1. This means each user is equally influenced by every user he or she retweets. We then associate a belief $p_i^{(0)} = 1$ to every user who employed one of the words in our lexicon, and $p_i^{(0)} = 0$ to all who didn't. Lastly, we create new beliefs $\mathbf{p}^{(t)}$ using the updating rule:

$$\mathbf{p}^{(t)} = T\mathbf{p}^{(t-1)} \tag{1}$$

Notice that the all the beliefs converge $p_i^{(t)}$ to the same value as $t \to \infty$, thus we run the diffusion process with $t = 2$. Notice also that $p_i^{(t)} \in [0, 1]$. With this real value associated with each user, we get 4 strata by randomly selecting up to 1500 users with p_i in the intervals $[0, .25)$, $[.25, .50)$, $[.50, .75)$ and $[.75, 1]$.

Annotating Hateful Users. We annotate 4,972 users as hateful or not using *Crowdflower*, a crowdsourcing service. The annotators were given the definition of hateful conduct according to Twitter's guidelines, and asked to annotate each user with the question:

> *Does this account endorse content that is humiliating, derogatory or insulting towards some group of individuals (gender, religion, race, nationality) or support narratives associated with hate groups (white genocide, holocaust denial, jewish conspiracy, racial superiority)?*

Annotators were asked to consider the whole webpage context rather than only individual publications or isolate words, and given examples of terms and codewords in ADLs hate symbol database. Each user was independently annotated by 3 annotators, and, if there was disagreement, he or she would be annotated by up to 5 annotators. In the end the annotators identified 544 hateful users.

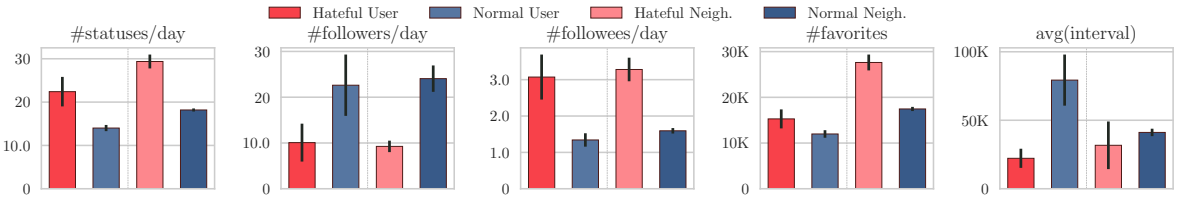

Figure 3: Average values for several activity-related statistics for hateful users, normal users, and users in the neighborhood of those. `avg(interval)` was calculated on the 200 tweets extracted for each user. Error bars represent 95% confidence intervals. The legend used in this graph is kept in the remainder of the paper.

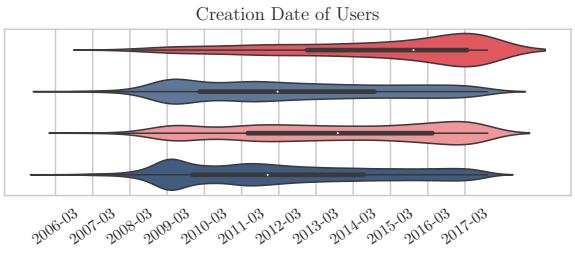

Figure 4: KDEs of the creation dates of user accounts. The white dot indicates the median and the thicker bar the first and third quartiles. Hateful users were created significantly later than their normal counterparts.

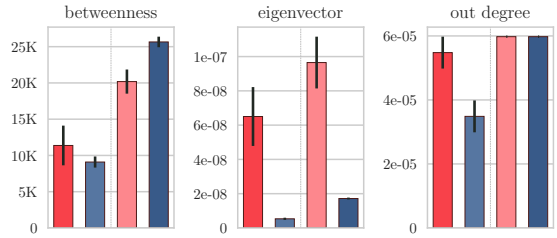

Figure 5: Median for network centrality metrics for hateful and normal users and their neighborhood calculated on the sampled retweet-induced graph.

4 CHARACTERIZING HATEFUL USERS

In this section we look at how hateful and normal users and their neighborhoods are different w.r.t. *profile attributes* provided by Twitter or inferred in the subgraph we sampled. Furthermore, we perform sentiment and lexical analysis on the *content* produced.

Creation Dates. We begin by analyzing the account creation date of hateful and non-hateful users, as depicted in Figure 4. Notice that the hateful users were created later than the normal ones (p-value < 0.001). A hypothesis for this difference is that hateful users are banned more often than normal ones. This resonates with existing methods for detecting accounts created to sell followers, where methods using the distribution of creation date have been successful [36]. We obtain similar results comparing the 1-neighborhood of such users, where the neighborhood of hateful users was also created more recently (p-value < 0.001).

User Activity. Other interesting metrics through which we can compare hateful and normal users, are the number of statuses, followers, followees and favorites a user has, and the interval in seconds between the tweets of each user. We show these statistics in Figure 3. We normalize the number of statuses, followers and followees by the number of days the users have since their account creation date. The results suggest that hateful users are "power users" in the sense that they tweet more, favorite more tweets by other people, and follow other users more (although they are less followed). We also show these statistics to the users in the 1-neighborhood of hateful and normal users, which in practice represents the users these groups retweeted. The analysis yields similar results when we compare the 1-neighborhood of hateful and normal users: neighbors of hateful users have more statuses per day, more followees per day more favorites, but the difference on the interval between tweets is smaller. It is hard to compare hateful/normal users and their neighborhood because of the distinct sampling methodology.

Network Centrality. We also analyze different measures of centrality for the users and their neighborhood, as depicted in Figure 5. The median hateful users and those in their neighborhood are more central in all measures when compared to their normal counterparts. This is an counter-intuitive finding, as hateful crimes, for example, have long been associated with "lone wolves", and antisocial people [3]. However, notice that, although the median for the centrality measurements for hateful user is bigger, the statistic for their average network centrality aren't. For example, none of the top 970 most central users according to eigenvector centrality are hateful).

Spam. It is interesting to consider the possible intersection between users that propagate hate speech and spammers, which have been widely studied. First, it is worth to notice that our methodology of data collection is robust against spammers, as spammers often exploit trending topics or popular hashtags to post URLs. As our data collection don't specifically look for these trending hashtags or topics, it is intuitive that this problem is lessened significantly. To confirm this intuition, we analyze metrics that have been used by previous work to detect spammers, such as the number of URLs/tweet,

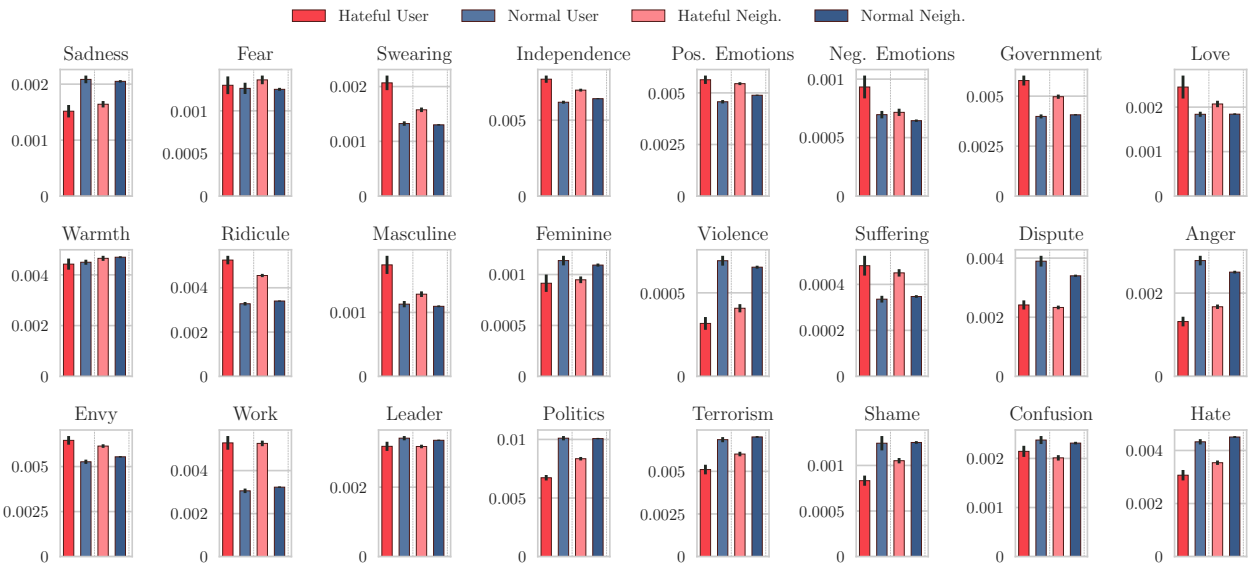

Figure 6: Average values for the relative occurrence of several categories in *Empath*. Notice that not all Empath categories were analyzed and that the to-be-analyzed categories were chosen before-hand to avoid spurious correlations. Error bars represent 95% confidence intervals.

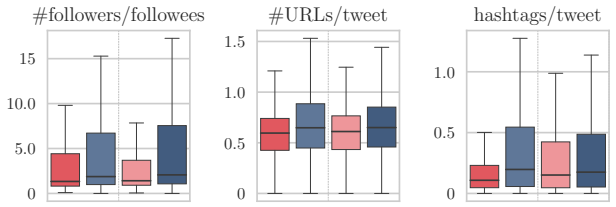

Figure 7: Boxplots for the distribution of metrics that indicate spammers. Hateful users and their neighborhood have slightly *less* followers per followee, *less* URLs per tweet, and *less* hashtags per tweet.

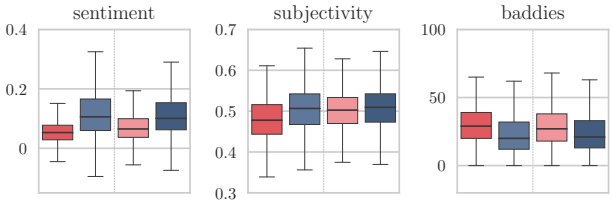

Figure 8: Boxplots for the distribution of sentiment, subjectivity and bad-words usage. Hateful users and their neighborhood are more negative, and use more bad words. Hateful users are less subjective.

and hashtags/tweet and the number of followers per followees [2]. This boxplot of these distributions is shown on Figure 7. We find that hateful users use, in average, *less* hashtags (p-value < 0.001) and less URLs (p-value < 0.001) per tweet than normal users. The same analysis holds if we compare the 1-neighborhood of hateful and non-hateful users (also with p-values < 0.001). Additionally, we also find that in average normal users have more followers per followees than hateful ones (p-value < 0.005), which also happens for their neighborhood (p-value < 0.001). This suggests that the hateful users are not spammers, and thus were probably annotated as hateful or suspended for abusive behavior. Notice that it is not possible to extrapolate this finding to all hateful users in Twitter, as maybe there are other types that spread hate speech tagging messages in popular hashtags or trending topics. Notice also that this doesn't necessarily mean that these accounts are not bots, although manual inspection by the authors suggests otherwise.

Lexical Analysis. We characterize hateful and normal users, as well as their neighborhood w.r.t. their content with *Empath* [11], as depicted in Figure 6. Our results are counter-intuitive. To begin with, hateful users use less words related to hate, anger, shame and terrorism, violence, and sadness (with p-values < 0.001), all of which are often taken as assumptions in the sampling process of other work intended to detect hateful tweets [8, 19]. A question that rises in this context is how sampling tweets based exclusively in a hate-related lexicon biases the sample of content to be annotated to a very specific type of user, which may not be representative of the average "hate-spreading" ones. This also reinforces the already stated claims that sarcasm and code-words may play a significantly role in defining such users [8, 21]. Categories of words more used by hateful users include positive emotions, negative emotions, suffering, work, love and swearing (with p-values < 0.001). This suggests the use of emotional vocabulary by hateful users (and those in their

Figure 9: Wordcloud for normal users. Notice that it shares several hashtags with the wordcloud associated with hateful users, such as `MAGA` and `Syria`.

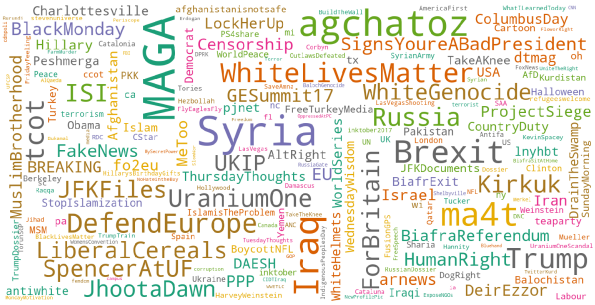

Figure 10: Word cloud for hateful users. Notice the inclusion of some hashtags associated with White Supremacist groups such as `WhiteGenocide`.

1-neighborhood). An interesting direction in that sense would be to analyze the sensationalism of statements made by hateful users when compared to normal ones, as it has been done in the context of *clickbaits*, catchy titles often associated with frivolous or fake news-pieces [7]. Overall, the non-triviality of the lexical characteristics of these groups of users reinforces the difficulties found in the NLP community to attack the problem of successfully detecting hate-speech [8].

Sentiment. Following on the finding that, according to *Empath*, hateful users use more negative and positive words, we explore the sentiment in the sentences they write using *VADER* [41], as depicted in Figure 8. We find that sentences written by hateful users are more negative, and are less subjectivee (p-value < 0.001). The neighborhood of hateful user is also more negative (p-value < 0.001), however not less subjective. We also analyze the distribution profanity per tweet in hateful and non-hateful users. The latter is obtained by matching all the words in Shutterstock's "List of Dirty, Naughty, Obscene, and Otherwise Bad Words" [2]. We find that hateful users and their neighborhood employ more profane words per tweet, also confirming the results from the analysis with *Empath*.

A qualitative look. Finally, we briefly present two qualitative insights on the content present in the user profiles we analyze. In Figures 9 and Figure 10 we display wordclouds containing the hashtags that were mostly used by hateful and non-hateful users. The wordcloud for hateful users contains some hashtags that have been associated with openly racist institutions or individuals such as American Renaissance [5]. Also, we can see that several hashtags are shared among both groups, such as `#Iraq` or `#MAGA`. Additionally, in Figure 11 we show Groyper, a picture of Pepe the Frog resting on his chin, which originated in the imageboard 4chan, and is known commonly used as avatar among the alt-right and the new right in social media [22]. An expressive number of the profiles identified as hateful by the annotators had Groyper (or some variation of Groyper) as a profile picture. These profiles are anonymous and tweet almost exclusively about politics, race and religion. Although we approach the problem of detecting hateful speech as a nuanced one, in the case of most of these profiles it is trivial to classify the vehiculated content according to the definition of hateful speech that we provided.

Suspended Accounts. Finally, we briefly analyze accounts that have been suspended three months after the data collection period in the 100 thousand users we collect. Most Twitter accounts are suspended due to spam, however as these accounts rarely get retweeted, they are harder to reach in the retweet induced graph. Thus, we have that other common reasons for suspension are abusive behavior and security issues with the account. We find the accounts that have been suspended among the 100, 386 collected accounts by making requests to Twitter's API. We use these suspended accounts as another source for potentially hateful behavior, as quantitative and qualitative analysis suggests they do not behave as spammers, and as they have a large intersection with the accounts labeled as hateful. Notice that these accounts may present other types of abusive behavior other than hate speech, such as offenses not based on attributes such as race, gender, etc.

Table 1: Percentage and absolute number of accounts that got suspended after three months

	Hateful	Normal	Others
Suspended Accounts	9.09% (55)	0.32% (14)	0.33% (314)

As depicted in Table 1, we find that 55 of the users classified as hateful by the crowdsourced annotators were banned in 3 months time, which corresponds to roughly 9% of all hateful users. In contrast, only 14 normal users were banned (0.32%), and for all 100 thousand users, 314 users were banned, corresponding to 0.33%. This result strengthens our findings, as we find that the annotations we performed seem to be somewhat in accordance to Twitters own moderation process. Interestingly, we collected the suspended accounts right before Twitter started to enforce new rules on violence, abuse, and hateful conduct, making exploring the differences between accounts that have been suspended *before* and *after* this change of policy a promising direction.

---

[2] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

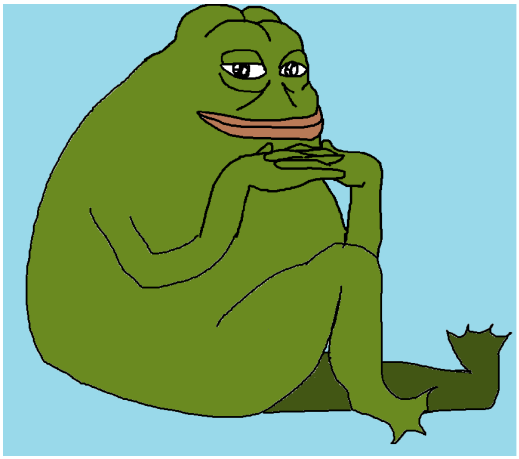

Figure 11: Groyper, an illustration of Pepe the Frog which was present in several hateful users identified, often in some variation.

5 RELATED WORK

We briefly review previous work on detecting and characterizing hate speech in OSNs. Tangent problems such as cyber-bullying and offensive language are not extensively covered, refer to [31]. We compare aspects of other methodology previously employed. It is important to notice that, for many of the works done in the context of OSNs, the main objective of the work we refer was to detect hate speech, whereas we emphasize characterization.

Many previous studies collect data by sampling OSNs with the aid of a lexicon with terms associated with hate speech [4, 8, 21, 40]. This may be succeeded by expanding this lexicon adding other co-occurring terms [40]. Other techniques employed include matching regular expressions [37], selecting features in tweets from users known to have reproduced hate speech [19]. We employ a random-walk-based methodology. Unlike previous work, our methodology uses a lexicon of hate-related words as a starting point to run a diffusion process. This diffusion process will give us a number of "closeness to hate-related words" associated with each user, which we use to perform a stratified sampling of the users to be annotated.

In the existing previous work on hate-speech detection, human annotators are used to label content. This labeling may be done by the researchers themselves [10, 19, 21, 40], selected annotators [14, 37], or crowd-sourcing services [4]. Hate-speech speech has been pointed out as a difficult subject to annotate on [29, 38]. We also employ *CrowdFlower* to annotate our data. Unlike previous work we provide annotators with the the entire profile of the user instead of individual tweets, this provides better context for the annotators [39].

Although most previous works focus on detection, there are some notable exceptions. Silva et. al [33], matches regex-like expressions on large datasets on Twitter and Whisper to characterize the targets of hate in online social networks. Also, Gerstenfeld et. al [13] analyze hateful websites characterizing their *modus operandi* w.r.t. monetization, recruitment, and international appeal.

6 DISCUSSION AND CONCLUSION

We present a first characterization of hate speech in Online Social Networks at a user-level granularity. We develop a methodology to sample Twitter which consists of obtaining a generic subgraph in Twitter, finding users who employed words in a lexicon of hate-related words and running a diffusion process based on DeGroot's learning model to sample for users in the neighborhood of these users. We then used *Crowdflower* to manually annotate 4, 972 users, of which 544 were considered to be hateful.

Our findings shed light on how hateful users are different from normal ones with respect to their user activity patterns, network centrality measurements, and the content they produce. Among our findings, we discover that the median hateful user is more central in the retweet network, more recently created, write more negative sentences and use lexicon associated with categories such as hate, terrorism, violence and anger *less* than normal ones. Furthermore, this analysis seem to also hold for the 1-neighborhood of the hateful and normal users.

Nevertheless, our analysis still has limitations that lead to interesting future research directions. Firstly, it is reasonable to question the definition of *hateful user*, in the sense that it is not clear what is the threshold an account has to violate to be considered hateful. Although we argue that classifying hateful users is easier than classifying hateful content, it is still a non-trivial task due to the subjectivity of the definition of hate-speech. Secondly, it is not clear whether the characterization (and possibly detection) of hateful *users* would solve all problems related to hate speech, as looking at this coarser-grained level of OSNs may make detecting users who only occasionally propagate hate speech harder. Thus, an interesting question in this scenario is *How much of the hate speech is produced by what percentage of users?* Another weakness of our characterization is that we only considered the behavior of such users on Twitter, and it is possible that this analysis does not hold in other widely used OSNs, such as Facebook or Instagram.

As future work, we want to detection of hateful users OSNs, a task which may be explored in different ways. A simple strategy would be to develop classification models based on the numerical attributes that are associated with each user and analyzed in this paper, and with representations for the text employed by the users, such as *word2vec* [24]. However, another exciting strategy would be to use the connections in the entire graph that we sampled in Twitter to create representations for each node (user). Interestingly, modern approaches allow each node to be linked to a vector of features [18], which suggests that we would be able to use both the content produced by the user as well as their positions in the network. If accomplished, such methods for detecting these misbehaving users could help the moderation teams of online social network to quickly identify and take the necessary measures against the hateful profiles.

ACKNOWLEDGEMENTS

This is work was supported by CNPq, CAPES, FAPEMIG, InWeb, MASWEB, INCT-Cyber, and ATMOSPHERE PROJECT. We would like to thank Nikki Bourassa, Ryan Budish, Amar Ashar and Robert Faris from the Berkman Klein Center for Internet and Society for their insightful suggestions.

REFERENCES

[1] Hate Base. 2017. Hate Base. (2017). https://www.hatebase.org/

[2] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6. 12.

[3] Jason Burke. 2017. The myth of the 'lone wolf' terrorist. (2017). https://www.theguardian.com/news/2017/mar/30/myth-lone-wolf-terrorist

[4] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (2016), 11.

[5] Southern Poverty Law Center. [n. d.]. Active Hate Groups in the United States in 2014. ([n. d.]). https://www.splcenter.org/fighting-hate/intelligence-report/2015/active-hate-groups-united-states-2014

[6] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. *Icwsm* 10, 10-17 (2010), 30.

[7] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv preprint arXiv:1703.04009* (2017).

[9] Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. *arXiv preprint arXiv:1605.03481* (2016).

[10] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*. 29–30.

[11] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[12] Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry* (1977), 35–41.

[13] Phyllis B Gerstenfeld, Diana R Grant, and Chau-Pu Chiang. 2003. Hate online: A content analysis of extremist Internet sites. *Analyses of social issues and public policy* 3, 1 (2003), 29–44.

[14] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 4 (2015), 215–230.

[15] Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–149.

[16] The Guardian. 2017. Google's bad week: YouTube loses millions as advertising row reaches US. (2017). https://www.theguardian.com/technology/2017/mar/25/google-youtube-advertising-extremist-content-att-verizon

[17] Pedro Calais Guerra, Roberto CSNP Souza, Renato M Assunção, and Wagner Meira Jr. 2017. Antagonism also Flows through Retweets: The Impact of Out-of-Context Quotes in Opinion Polarization Analysis. *arXiv preprint arXiv:1703.03895* (2017).

[18] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. *arXiv preprint arXiv:1706.02216* (2017).

[19] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks.. In *AAAI*.

[20] Anti Defamation League. 2017. ADL Hate Symbols Database. (2017). https://www.adl.org/education/references/hate-symbols

[21] Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the Hate Code on Social Media. *arXiv preprint arXiv:1703.05443* (2017).

[22] Know Your Meme. [n. d.]. Groyper. ([n. d.]). http://knowyourmeme.com/memes/groyper

[23] Know Your Meme. 2016. Operation Google. (2016). http://knowyourmeme.com/memes/events/operation-google

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] L Rainie, Janna Anderson, and Jonathan Albright. 2017. The future of free speech, trolls, anonymity and fake news online. *Pew Research Center, March* 29 (2017).

[26] Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. 2012. Sampling directed graphs with random walks. In *INFOCOM, 2012 Proceedings IEEE*. IEEE, 1692–1700.

[27] Bruno Ribeiro, Pinghui Wang, and Don Towsley. [n. d.]. On Estimating Degree Distributions of Directed Graphs through Sampling. ([n. d.]).

[28] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation.. In *EMNLP*, Vol. 13. 704–714.

[29] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).

[30] Fabio Sabatini and Francesco Sarracino. 2017. Online Networks and Subjective Well-Being. *Kyklos* 70, 3 (2017), 456–480.

[31] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain*. 1–10.

[32] Andrew Sellars. 2016. Defining Hate Speech. (2016).

[33] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media.. In *ICWSM*. 687–690.

[34] Eric Stein. 1986. History against Free Speech: The New German Law against the" Auschwitz" - And Other - "Lies". *Michigan Law Review* 85, 2 (1986), 277–324.

[35] Twitter. 2017. Hateful conduct policy. (2017). https://support.twitter.com/articles/20175050

[36] Bimal Viswanath, Muhammad Ahmad Bashir, Muhammad Bilal Zafar, Simon Bouget, Saikat Guha, Krishna P Gummadi, Aniket Kate, and Alan Mislove. 2015. Strength in numbers: Robust tamper detection in crowd computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 113–124.

[37] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[38] Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on Natural Language Processing and Computational Social Science*. 138–142.

[39] Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. 78–85.

[40] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.. In *SRW@ HLT-NAACL*. 88–93.

[41] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 129–136.