Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Characterizing and Detecting Hateful Users on Twitter

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

Abstract

Current approaches to characterize and detect hate speech focus on *content* posted in Online Social Networks (OSNs). They face shortcomings to get the full picture of hate speech due to its subjectivity and the noisiness of OSN text. This work partially addresses these issues by shifting the focus towards *users*. We obtain a sample of Twitter's retweet graph with $100,386$ users and annotate $4,972$ as hateful or normal, and also find 668 users suspended after 4 months. Our analysis shows that hateful/suspended users differ from normal/active ones in terms of their activity patterns, word usage and network structure. Exploiting Twitter's network of connections, we find that a node embedding algorithm outperforms content-based approaches for detecting both hateful and suspended users. Overall, we present a user-centric view of hate speech, paving the way for better detection and understanding of this relevant and challenging issue.

Introduction

A growing body of work aims to understand and detect hate speech by creating representations for *content* in Online Social Networks (OSN) and then classifying these tweets or comments as hateful or not, drawing insights along the way (Greevy and Smeaton 2004; Warner and Hirschberg 2012). However, in OSNs, texts are often not self-contained, and are packed with informal language, spelling errors, special characters and sarcasm (Dhingra et al. 2016; Riloff et al. 2013). Besides that, hate speech itself is highly subjective, reliant on temporal, social and historical context, and occurs sparsely (Schmidt and Wiegand 2017).

Fortunately, the data in posts, tweets or messages is not the only signal we may use to study hate speech in OSNs. These are often linked to a *profile* which also conveys important information. Studying hate on a *user-level* rather than *content-level* enables the characterization of hateful users' activities and connections, and the usage of the very structure of the social network by detection techniques.

In this paper we characterize and detect hate-speech on Twitter on a *user-level*. We collect a dataset of $100,386$ users along with up to 200 tweets for each with a random-walk-based crawler. We then select $4,972$ to manually annotate through crowdsourcing using a diffusion-based methodology. We obtain 544 users labeled as hateful, and also find 668 users that have been suspended by Twitter approximately 4 months after the data collection.

Our analysis shows that hateful users differ from normal ones in terms of their activity patterns, word usage and network structure. Similar results are obtained when comparing the neighbors of hateful *vs* neighbors of normal users and also suspended *vs* active users. We observe that hateful users are densely connected, and thus formulate the hate speech detection problem as a task of semi-supervised learning over a graph. We find that a node embedding algorithm, which exploits the retweet network, outperforms content-based approaches for the detection of hateful (95% *vs* 88% AUC) and suspended users (93%*vs* 88% AUC) [1].

Data Collection

Most previous work on detecting hate employs a lexicon-based data collection (Davidson et al. 2017; Waseem and Hovy 2016). However, this methodology is biased towards direct hate speech, struggling with code-words (Magu, Joshi, and Luo 2017) or the lack of offensive words (Davidson et al. 2017). We propose an alternative methodology.

We represent the connections among users in Twitter using the retweet network (Cha et al. 2010), and sample the graph with the *DURW* algorithm, which unbiasedly estimates the out-degree distribution of nodes (Ribeiro, Wang, and Towsley 2010). We collect a sample of Twitter's retweet graph T with $100,386$ users and $2,286,592$ edges along with the 200 most recent tweets for each user.

We then select a subsample to be annotated. We:

1. Create a lexicon of words that are mostly used in the context of hate speech. This is unlike other work (Davidson et al. 2017) as we do not consider words that are employed in a hateful context but often used in other contexts in a harmless way (*e.g.* `n*gger`).
2. Run a diffusion process on the graph based on DeGroot's Learning (Golub and Jackson 2010), assigning an initial belief $p_i^0 = 1$ to each user u_i who employed the words in the lexicon; and iteratively updating the beliefs with the rule $\mathbf{p}^t = T\mathbf{p}^{t-1}$. This prevents our sample from being excessively small/biased to some vocabulary.

[1] code/data: https://github.com/manoelhortaribeiro/HatefulUsersTwitter

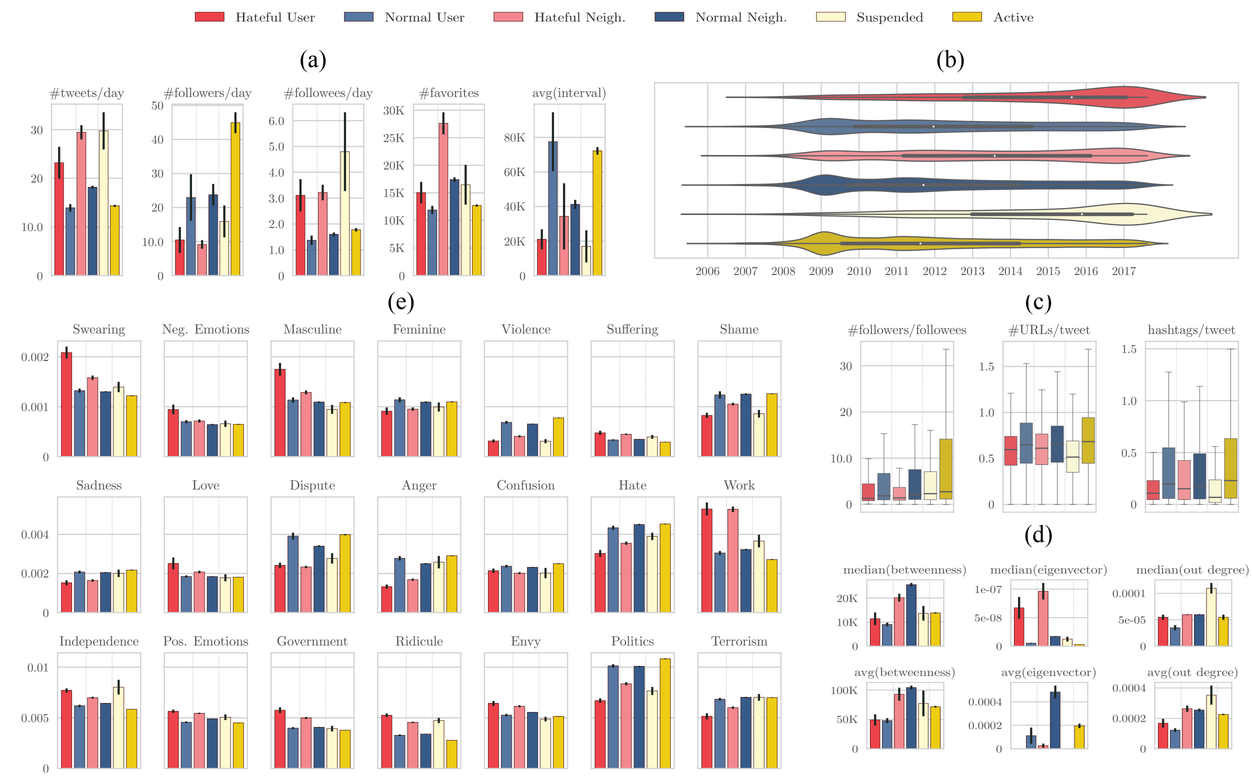

Figure 1: **(a)** Average values for several activity-related statistics for hateful users, normal users, users in the neighborhood of those, and suspended/active users. Error bars represent 95% confidence intervals. **(b)** KDEs of the creation dates of user accounts. **(c)** Boxplots for the distribution of metrics that indicate spammers. **(d)** Network centrality metrics for hateful and normal users, their neighborhood, and suspended/non-suspended users calculated on the sampled retweet graph. **(e)** Average values for the occurrence of words in *Empath* categories.

3. Divide the users in 4 strata according to their associated beliefs after the diffusion process (users with p_i in the intervals $[0, .25)$ $[.25, .5)$ $[.5, .75)$ $[.75, 1]$), and perform a stratified sampling, obtaining up to 1500 user per strata to be manually annotated.

We then annotate $4,972$ users given their profile as hateful or not using *CrowdFlower*. Annotators were given Twitter's hateful conduct guidelines and asked, for each user, if *the account endorsed content that is humiliating, derogatory or insulting towards some group of individuals or supported narratives associated with hate groups*. Each user was independently annotated 3 times, and, if there was disagreement, up to 5. We ended up identifying $4,428$ normal and 544 hateful users. The data was collected between the 1st and 7th of Oct/17. We also obtain all the users who got suspended up to 14/Jan/18 (668).

Characterizing Hateful Users

We analyze how hateful and normal users differ w.r.t. their activity, vocabulary and network centrality. We also analyze the neighbors of hateful/normal in the retweet graph, and suspended/active users to reinforce our findings. We compare those in pairs as the sampling mechanism for each of the populations differs. We argue that each of these pairs is a proxy for hateful speech, and inspecting the three increases the robustness of our analysis.

Hateful users are power users. We analyze the number of tweets, followers, followees and favorite tweets a user has, and the interval in seconds between their tweets. We show these statistics in Figure 1**(a)**. We normalize the number of tweets, followers and followees by the number of days the users have since their account creation date. Our results suggest that hateful users are "power users" in the sense that they tweet more, in shorter intervals, favorite more tweets by other people and follow other users more (p-values < 0.01). The analysis yields similar results when we compare the 1-neighborhood of hateful and normal users, and when comparing suspended and active accounts.

Hateful users have newer accounts. The account creation date of users is depicted in Figure 1**(b)**. Hateful users were created later than normal ones (p-value < 0.001). A hypothesis for this difference is that hateful users are banned more often due to Twitter's guidelines infringement. We obtain similar results w.r.t. the 1-neighborhood of such users, where the hateful neighbors were also created more recently (p-value < 0.001), and also when comparing suspended and active accounts (p-value < 0.001).

Hateful users don't behave like spammers. We investigate whether users that propagate hate speech are spammers. We analyze metrics that have been used by previous work to detect spammers, such as the numbers of URLs per tweet, of hashtags per tweet and of followers per followees (Benevenuto et al. 2010). The boxplot of these distributions is shown on Figure 1**(c)**. 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. We obtain similar results when comparing the 1-neighborhood of hateful and non-hateful, or suspended and active users. We also find that, in average, normal users have more followers per followees than hateful ones (p-value < 0.05), which also happens for their neighborhood (p-value < 0.05). This suggests that the hateful and suspended users in the sample do not use systematic and programmatic methods to deliver their content.

The median hateful user is more central. We analyze different measures of centrality for users, as depicted in Figure 1**(d)**. The median hateful user is more central in all measures when compared to their normal counterparts. This is a counter-intuitive finding, as hateful crimes have long been associated with "lone wolves" (Burke 2017). We observe similar results when comparing the median eigenvector centrality of the neighbors of hateful and normal users, as well as suspended and active users. In the latter pair, suspended users also have higher median out degree. When analyzing the average for such statistics, we observe the average eigenvector centrality is higher for the opposite sides of the previous comparisons. This happens because some very influential users distort the value: for example, the 970 most central users according to the metric are normal.

Hateful users use non-trivial vocabulary. We characterize users w.r.t. their content with *Empath* (Fast, Chen, and Bernstein 2016), as depicted in Figure 1**(e)**. Hateful users use *less* words related to hate, anger, shame and terrorism, violence, and sadness when compared to normal users (with p-values < 0.001). This raises the question of how sampling tweets based in a hate-related lexicon biases the sample towards a very specific type of "hate-spreading" user. Categories of words more *used* by hateful users include positive emotions, negative emotions, suffering, work, love and swearing (with p-values < 0.001), suggesting the use of emotional vocabulary. When we compare the neighborhood of hateful and normal users and suspended vs active users, we obtain very similar results. Overall, the non-triviality of the vocabulary of these groups of users reinforces the difficulties found in the NLP approaches to sample, annotate and detect hate speech (Davidson et al. 2017; Magu, Joshi, and Luo 2017).

Hateful users are densely connected. Finally, we analyze the frequency at which hateful and normal users, as well as suspended and active users, interact within their own group and with each other. Table 1 depicts the probability of an node of a given type retweeting other type of node. We find that 41% of the retweets of hateful users are to other hateful users, which means that they are 71 times more likely to retweet another hateful user, considering the occurrence of hateful users in the graph. We observe a similar phenomenon with suspended users, which have 7% of their retweets directed towards other suspended users. As suspended users correspond to only 0.68% of the users sampled, this means they are approximately 11 times more likely to retweet other suspended users. The high density of connections among hateful and suspended users suggest a strong modularity. We exploit this, along with user-level features attributes to robustly detect these users.

Node Type	(%)	Node Type	(%)
🔴 → 🔴	41.50	🔴 → 🔵	13.10
🔵 → 🔵	15.90	🔵 → 🔴	2.86
⚪ → ⚪	7.50	⚪ → 🟡	92.50
🟡 → 🟡	99.35	🟡 → ⚪	0.65

Table 1: Occurrence of the edges between hateful 🔴 and normal 🔵 users, and between suspended ⚪ and active 🟡 users. Results are normalized w.r.t. to the type of the source node, as in: $P(\text{source type}\rightarrow\text{dest type}|\text{source type})$. Notice that the probabilities do not add to 1 for hateful and normal users as we don't present the statistics for non-annotated users.

Detecting Hateful Users

As we consider users and their connections in the network, we can use information that is not available for models which operate on the granularity level of tweets or comments to detect hate speech. We consider two sets of features: *(i)* `user`: features such as number of statuses, followers, followees, favorites, and centrality measurements such as betweenness, eigenvector centrality and the in/out degree of each node. *(ii)* `glove`: off-the-shelf 300-dimensional GloVe's vector (Pennington, Socher, and Manning 2014) averaged across all words in a given tweet, and subsequently, across all tweets a user has.

Using these, we experimentally compare Gradient Boosted Trees (`GradBoost`), known to perform very well when the number of instances is not very large, and a model aimed specifically at learning in graphs, GraphSage (Hamilton, Ying, and Leskovec 2017a) (`GraphSage`). Interestingly, the latter approach is semi-supervised, and allows us to use the neighborhood of the users we are classifying even though they are not labeled, exploiting the modularity between hateful and suspended users we observed. The algorithm creates low-dimensional embeddings for nodes, given associated features (unlike other node embeddings, such as `node2vec` (Grover and Leskovec 2016)). Moreover, it is inductive — which means we don't need the entire graph to run it. For additional information on node embeddings methods, refer to (Hamilton, Ying, and Leskovec 2017b).

Experimental Settings. We run the algorithms trying to detect both hateful and normal users, as annotated by the crowdsourcing service, as well as trying to detect which users got suspended. We perform a 5-fold cross validation for the two proposed approaches (`GradBoost` and `GraphSage`), and accounted for the class imbalance (of approximately 1 to 10) in the loss function. We keep the same ratio of positive/negative classes in both tasks, which,

		Hateful/Normal			Suspended/Active		
Model	Features	Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
`GradBoost`	`user+glove`	84.6 ± 1.0	52.0 ± 2.2	88.4 ± 1.3	81.5 ± 0.6	48.4 ± 1.1	88.6 ± 0.1
	`glove`	84.4 ± 0.5	52.0 ± 1.3	88.4 ± 1.3	78.9 ± 0.7	44.8 ± 0.7	87.0 ± 0.5
`GraphSage`	`user+glove`	$\mathbf{90.9 \pm 1.1}$	$\mathbf{67.0 \pm 4.1}$	$\mathbf{95.4 \pm 0.2}$	$\mathbf{84.8 \pm 0.3}$	$\mathbf{55.8 \pm 4.0}$	$\mathbf{93.3 \pm 1.4}$
	`glove`	90.3 ± 1.9	65.9 ± 6.2	94.9 ± 2.6	84.5 ± 1.0	54.8 ± 1.6	93.3 ± 1.5

Table 2: Prediction results and standard deviations for the two proposed settings: detecting hateful users and detecting suspended users. The semi-supervised node embedding approach performs better than state-of-the-art supervised learning algorithms in all the assessed criteria, suggesting the benefits of exploiting the network structure to detect hateful and suspended users.

in practice, means we used the 4972 annotated users in the first setting (where approximately 11% were hateful) and, in the second setting, selected 6680 users from the graph, including the 668 suspended users, and other randomly sampled active users from the graph. Notice that, as we are dealing with a binary classification problem, we may control the trade-off between specificity and sensitivity by varying the positive-class threshold. We report the area under the ROC curve (AUC) the Accuracy and the F1-Score.

Results. The results of our experiments are shown in Table 2. We find that the node embedding approach using the features related to both users and the *GloVe* embeddings yields the best results for all metrics in the two considered scenarios. Using the features related to users makes little difference in many settings, yielding, for example, exactly the same AUC, and very similar Accuracy/F1-Score in the Gradient Boosting models trained with the two sets of parameters. However, the usage of the retweet network yields promising results, especially because we observe improvements in both the detection of hateful users and of suspended users, which shows the performance improvement occurs independently of our annotation process.

Conclusion

We present a user-centric view of hate speech, paving the way for better detection and understanding of this relevant and challenging issue. Our characterization sheds light on how hateful users differ from normal ones with respect to their user activity patterns, network centrality measurements, and the content they produce. We show that these differences can be exploited to robustly detect such users. We expand our characterization and detection methodology to *suspended users*, obtaining similar results.

Acknowledgements

We would like to thank Nikki Bourassa, Ryan Budish, Amar Ashar and Robert Faris from the BKC at Harvard for their insightful suggestions. This work was partially supported by CNPq, CAPES and Fapemig, as well as projects InWeb, INCT-Cyber, MASWEB, BigSea and Atmosphere.

References

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *CEAS*.

Burke, J. 2017. The myth of the 'lone wolf' terrorist. *The Guardian*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. *ICWSM*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *arXiv:1703.04009*.

Dhingra, B.; Zhou, Z.; Fitzpatrick, D.; Muehl, M.; and Cohen, W. W. 2016. Tweet2vec: Character-based distributed representations for social media. *arXiv:1605.03481*.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *CHI*.

Golub, B., and Jackson, M. O. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*.

Greevy, E., and Smeaton, A. F. 2004. Classifying racist texts using a support vector machine. In *SIGIR*.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017a. Inductive representation learning on large graphs. *arXiv:1706.02216*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017b. Representation learning on graphs: Methods and applications. *arXiv:1709.05584*.

Magu, R.; Joshi, K.; and Luo, J. 2017. Detecting the hate code on social media. *arXiv:1703.05443*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Ribeiro, B.; Wang, P.; and Towsley, D. 2010. On estimating degree distributions of directed graphs through sampling. *University of Massachusetts CMPSCI Technical Report UM-CS-2010-046*.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP*. ACL.

Warner, W., and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In *LSM*, 19–26. ACL.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW @ HLT-NAACL*.