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,yurisantos,virgilio,meira}@dcc.ufmg.br
Universidade Federal de Minas Gerais, Brazil

Hate Speech Beyond Content

- We propose characterizing hate speech on a user-level granularity.
- This allow us to explore dimensions such as one's activity and connections.
- Other approaches struggle with subjectivity, noisy text, and codewords [2,4].

Data Collection

A. Sampled 100,386 users from Twitter's retweet graph using a random-walk based approach.

B. Collected users suspended 3 m after the data collection (670).

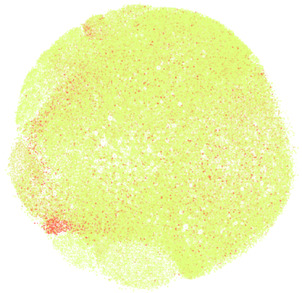
C. Created a lexicon of words mostly used in the context of hate speech.

D. Ran a diffusion process on the graph, marking users who employed the lexicon as infected.

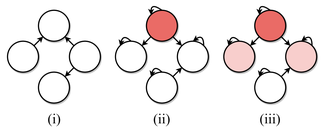
E. Selected users with different associated values after the diffusion process to annotate 4,972 users, out of which 544 were considered hateful.

F. Annotators were asked:

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



F-1. Network of 100,386 users.



F-2 Depiction of our diffusion process.

Characterizing Hateful Users

- We compare neighbors of hateful/normal & suspended/active users (in pairs as the sampling mechanism for each differ).
- We observe many similarities among the three pairs. We find that:

(a) Hateful Users are Power Users: they tweet more, in shorter intervals, favorite more tweets by other people and follow other users more.

(b) Hateful Users have Newer Accounts: they tweet more, in shorter intervals, favorite more tweets by other people and follow other users more.

(c) Hateful Users Don't Behave Like Spammers: they have less hashtags and URLs per tweet and have a smaller follower/followee ratio.

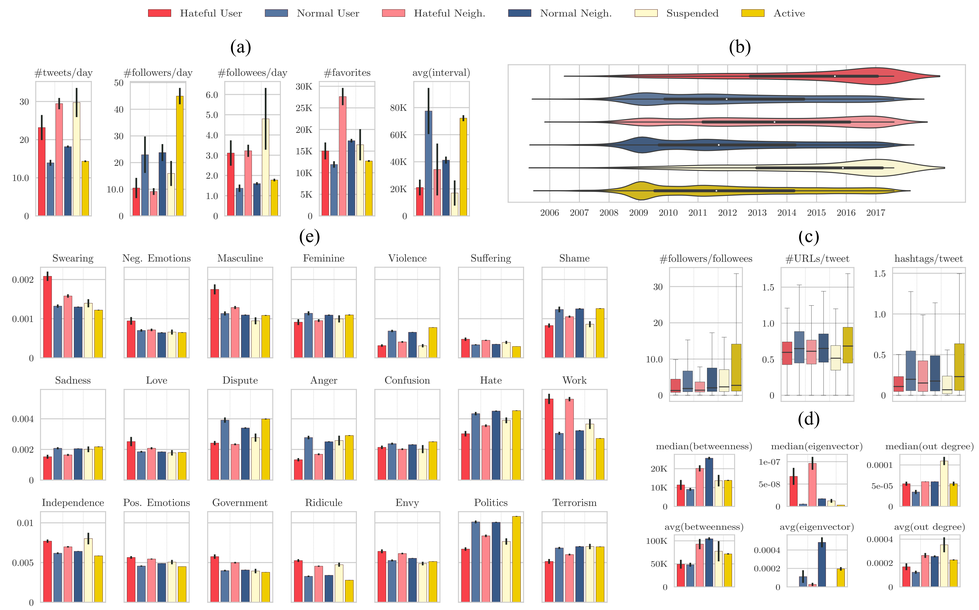
(d) The Median Hateful User is More Central: the median hateful user is more central in all measures when compared to their normal counterparts. Most influential users aren't hateful.

(e) Hateful Users Use Non-Trivial Vocabulary: they use less words related to hate, anger, shame and terrorism, violence, and sadness than normal users and more words related to pos./neg. emotions, work, love and swearing.

(T-1) Hateful users are more likely be connected: they are 71 times more likely to retweet another hateful user. Suspended users are 11 times more likely to retweet other suspended users.

Detecting Hateful Users

- Previous detection approaches use content-exclusive features [1-3,5], we explore activity and network centrality related features (`user`) in addition to word embeddings (`glove`).
- We use GraphSage [6], a inductive node embedding algorithm, and compare the performance with a Gradient Boosting Classifier, as depicted in table **T-2.**
- Using activity and network related features in the supervised learning algorithm yields statistically insignificant results. The semi-supervised node embedding approach performs better than Gradient Boosting, suggesting the benefits of exploiting the network structure to detect hateful and suspended users.



F-3. (a) Average values for several activity-related statistics for hateful users, normal users, users in the neighborhood of those, and suspended/active users. **(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 relative occurrence of several categories in Empath.

Node Type	(%)	Node Type	(%)
● → ●	41.50	● → ●	13.10
● → ●	15.90	● → ●	2.86
○ → ○	7.50	○ → ○	92.50
○ → ○	99.35	○ → ○	0.65

T-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. Notice that the probabilities do not add to 1 in hateful and normal users as we don't present the statistics for non-annotated users.

		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

T-2. Prediction results and standard deviations for the two proposed settings: detecting hateful users and detecting suspended users.

References

[1] *A survey on hate speech detection using natural language processing. Schmidt, A. and Wiegand, M. 5th ACL workshop on NLP for Social Media*
[2] *Automated Hate Speech Detection and the Problem of Offensive Language, Davidson, T. and Warmsley, D. and Macy, M. and Weber, I. ICWSM 2017*
[3] *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Waseem, Z. and Hovy, D. SRW @ HLT-NAACL 2016.*
[4] *Detecting the Hate Code on Social Media. Magu, Rijul and Joshi, Kshitij and Luo, Jiebo. ICWSM 2017*
[5] *Deep learning for hate speech detection in tweets. Badjatiya, P and Gupta, S and Gupta, M and Varma, V. WWW 2017*
[6] *Inductive Representation Learning on Large Graphs. Hamilton, William L and Ying, Rex and Leskovec, Jure NIPS 2017*