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

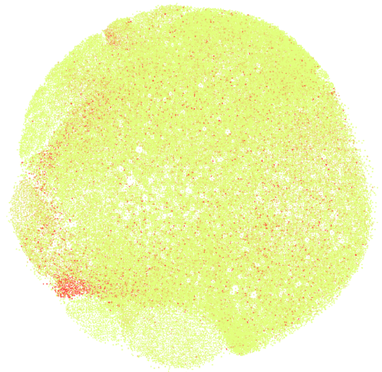
Hate Speech Beyond Content

- 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].

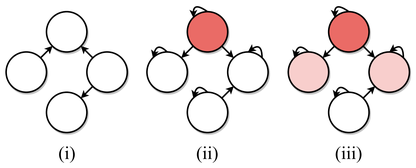
Data Collection

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.



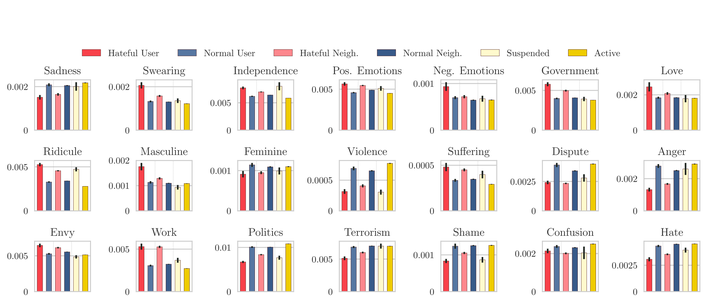
F-2 Depiction of our diffusion process.

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**

Hateful Users Use Non-Trivial Vocabulary

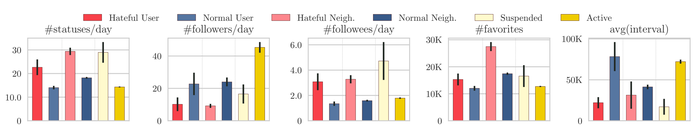
- Hateful users use **less** words related to hate, anger, shame and terrorism, violence, and sadness than normal users.
- Hateful users use **more** words related to positive/negative emotions, work, love and swearing than normal users.



F-6. Average values for the relative occurrence of several categories in *Empath*. Error bars represent \$95\%\$ confidence intervals.

Hateful Users are “Power Users”

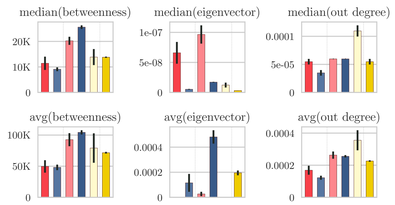
- Hateful users tweet more, in shorter intervals, favorite more tweets by other people and follow other users more.



F-3. Average values for activity-related statistics for hateful/normal users, users in the neighborhood of those, and suspended/active users.

The Median Hateful User is More Central

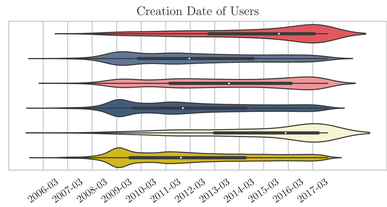
- The median hateful user is more central in all measures when compared to their normal counterparts. Most influential users aren’t hateful.



F-4. Network centrality metrics for hateful/normal users, their neighborhood, and suspended/active users. Calculated on the retweet graph.

Hateful Users Have Newer Accounts

- Hateful users tweet more, in shorter intervals, favorite more tweets by other people and follow other users more.



F-5. KDEs of the creation dates of user accounts. The white dot indicates the median and the thicker bar indicates the first and third quartiles.

A Peek into The Future: Detection

- Previous detection approaches use content-exclusive features [1-3,5], we explore activity and network centrality.
- We also explore using GraphSage [6] a node embedding.

		Hateful/Normal			Suspended/Active		
Model	Features	Accuracy	Recall	AUC	Accuracy	Recall	AUC
GradBoost	user+glove	84.6 ± 1.0	76.7 ± 2.4	88.4 ± 1.3	81.5 ± 0.6	78.9 ± 1.7	88.6 ± 0.1
	glove	84.4 ± 0.5	77.2 ± 2.1	88.4 ± 1.3	78.9 ± 0.7	77.7 ± 1.6	87.0 ± 0.5
AdaBoost	user+glove	69.1 ± 2.4	84.6 ± 1.9	85.5 ± 1.4	70.1 ± 0.1	84.4 ± 3.6	84.3 ± 0.5
	glove	69.1 ± 2.5	84.8 ± 1.8	85.5 ± 1.4	69.7 ± 1.0	83.0 ± 0.3	82.7 ± 0.1
GraphSage	user+glove	$\mathbf{90.9 \pm 1.1}$	84.6 ± 6.0	$\mathbf{95.4 \pm 0.2}$	$\mathbf{84.8 \pm 0.3}$	$\mathbf{85.6 \pm 5.4}$	$\mathbf{93.3 \pm 1.4}$
	glove	90.3 ± 1.9	$\mathbf{85.1 \pm 7.6}$	94.9 ± 2.6	84.5 ± 1.0	85.5 ± 3.9	93.3 ± 1.5

T-1. Performance of different classifiers to detect hateful *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*