Peer to Peer Hate: Hate Speech Instigators and Their Targets

Mai ElSherief, Shirin Nilizadeh,* Dana Nguyen, Giovanni Vigna, Elizabeth Belding
University of California, Santa Barbara
*CyLab, CMU Silicon Valley
{mayelsherif, dananguyen, vigna, ebelding}@ucsb.edu, shirin.nilizadeh@sv.cmu.edu

Abstract

While social media has become an empowering agent to individual voices and freedom of expression, it also facilitates anti-social behaviors including online harassment, cyberbullying, and hate speech. In this paper, we present the first comparative study of hate speech instigators and target users on Twitter. Through a multi-step classification process, we curate a comprehensive hate speech dataset capturing various types of hate. We study the distinctive characteristics of hate instigators and targets in terms of their profile self-presentation, activities, and online visibility. We find that hate instigators target more popular and high profile Twitter users, and that participating in hate speech can result in greater online visibility. We conduct a personality analysis of hate instigators and targets and show that both groups have eccentric personality facets that differ from the general Twitter population. Our results advance the state of the art of understanding online hate speech engagement.

Introduction

Social media has become a ubiquitous, powerful communication tool. However, while it has enabled rich, quick information sharing and conversation, it has also facilitated anti-social behavior including online harassment, trolling, cyberbullying, and hate speech. In a Pew Research Center study[1], 60% of Internet users had witnessed offensive name calling, 25% had seen someone physically threatened, and 24% witnessed sustained harassment of an individual.

In this paper, we focus on speech that denigrates a person because of their innate and protected characteristics, which is also known as *hate speech*. While there is no consensus on the definition of hate speech, prior work has shown that people are primarily bullied for their *perceived or actual* ethnicity, behavior, physical characteristics, sexual orientation, class or gender (Silva et al. 2016). Targeting a community or individual because of their immutable or prominent characteristics slowly eradicates feelings of safety and security (Hamm 1994; Levin and MacDevitt 2013). Prior studies have focused on online hate speech detection (Schmidt and Wiegand 2017) and characterization, e.g., effect of banning hate speech (Chandrasekharan et al. 2018); on-the-ground events that are triggered by hate speech (Williams and Burnap 2015; Wired 2016; Benesch 2014); and semi-organized raids by instigators to cripple hate speech detection technology (Hine et al. 2017). Despite this work, little is known about online hate speech actors, including hate speech instigators and targets.

We present the first comparative study of online hate speech instigators and targets. We curate a dataset of 27,330 hate speech Twitter tweets and extract 25,278 instigator and 22,287 target accounts. Prior work has presented evidence that social media can be used to obtain valuable data that incorporates facets of the virtual and physical worlds of bullying (Xu et al. 2012). We choose Twitter because it provides a platform for open discourse and a cross-section of the general public, with 328 million monthly active users in 2017 (Statista 2017). Our work seeks to answer the following research questions:

RQ1: How do hate instigator and target account characteristics and online visibility differ from each other and from generic Twitter account holders?

RQ2: Are there key personality differences between hate speech instigators, targets and general Twitter users?

Due to the lack of public hate speech datasets that include labeled roles of instigators and targets, we curate our own dataset for what we coin "*Peer to peer*" hate speech. This paper presents the following contributions:

- We present the first comparison of hate instigators, targets and general Twitter users in terms of profile self-presentation, Twitter visibility, and personality traits.
- We provide a compressed lexicon of Hatebase (the world's largest hate expression repository) for hate speech researchers, comprised of 51 terms likely to result in hate speech content across eight different hate classes. We outline a method of semi-automated classification that could be used for directed explicit hate speech data curation. We curate a dataset of 27,330 hate speech tweets, which we make publicly available for other researchers.[2]
- We examine the visibility of Twitter users through multivariant regression models and controlling for variables that can impact visibility measures.

[1] http://www.pewinternet.org/2014/10/22/online-harassment/

[2] The lexicon and the dataset are available here: https://github.com/mayelsherif/hate_speech_icwsm18

Our study yields multiple important findings. First, hate targets often have older accounts while instigators often have younger accounts. Compared to general users, both instigators and targets are more active in terms of becoming friends with others, posting tweets, and populating profile content. Targets include 60% and 40% more verified accounts than instigators and general users, respectively. Even when controlling for variables that can impact visibility measures, we find that higher visibility and participation in hate are correlated. More visible Twitter users (with more followers, retweets and lists) are more likely to become targets of hate. Finally, instigators and targets share some personality traits such as suspiciousness, low emotional awareness, and high anger and immoderation, which differ from personality traits of the general Twitter user population.

Related Work

Anti-social behavior. In 1997, the use of machine learning was proposed to detect classes of abusive messages (Spertus 1997). Cyberbullying has been studied on numerous social media platforms, *e.g.,* Twitter (Silva et al. 2016) and YouTube (Dinakar et al. 2012). Other work has focused on detecting personal insults and offensive language (Burnap and Williams 2014).

Hate speech characterization. The characterization and correlation of hate speech with contributing factors has recently received attention. Factors include on-the-ground "trigger" events, *e.g.,* terrorist attacks (Williams and Burnap 2015), crime (Wired 2016), and news (Hine et al. 2017).

Most closely related to our work are (Chatzakou et al. 2017b; 2017a; Cheng et al. 2017; Silva et al. 2016; Waseem and Hovy 2016). Chatzakou *et al.* (Chatzakou et al. 2017a) study the users of tweets with the #Gamergate hashtag. Similar to our results, they found that these users tend to have more friends and followers, and are generally more engaged than random users. Chatzakou *et al.* (Chatzakou et al. 2017b) study the properties of bullies and aggressors and employ supervised machine learning to classify Twitter users into four classes: bully, aggressive, spam, and normal. In contrast to their dataset, our dataset is more diverse and not biased towards specific types of hate speech. Moreover, we compare the characteristics of hate instigators and the targets of hate from multiple perspectives and show that, even when controlling for features that capture the activity level of the users, both hate instigator and target users are more likely to get attention on Twitter, *i.e.,* they obtain more followers, are retweeted and listed more.

Alternatively, Cheng et al. find that prior negative mood and the context of the discussion can combine to double participants' baseline engagement in trolling behavior. While the authors only used sentiment analysis to investigate mood, we incorporate a full analysis of the Big Five personality traits. In addition, we study the personality traits of both instigators and targets and compare results to a random sample of general Twitter users. Silva *et al.* (Silva et al. 2016) identified hate target groups in terms of their class and ethnicity on Twitter and Whisper by searching for sentence structures similar to "I <intensity> hate <targeted group>." However, we identify the actual accounts of hate targets on Twitter, *i.e.,* those that are explicitly mentioned by hate instigators. Therefore, our analysis provides a unique lens to analyze characteristics of target accounts.

Preliminaries

Waseem *et al.* outline a typology of abuse language that differentiates between language directed towards a specific individual or entity (*Directed*) versus a general group of individuals who share a common characteristic, e.g., ethnicity or sexual orientation (*Generalized*) (Waseem et al. 2017). Another dimension is whether the abusive language is *explicit*, e.g., contains racial, sexist or homophobic slurs, or *implicit*, which is harder to determine without adding contextual variables to the content. In this work, we study instances of *directed* hate speech that occur between two Twitter accounts. We define the following entities:

- A **hate tweet** is an explicit directed tweet that contains one or more hate speech terms used against a Twitter account holder. An example from our dataset is: "*@usr n*gger f*ck u igger n*gger n*gger n*gger.*"[3] This tweet is explicit because of the word "n*gger;" it is directed because it targets a specific account (@usr).[4]
- A **hate instigator (HI)** is a Twitter account that posts one or more hate tweets.
- A **hate target (HT)** is a Twitter account targeted by a hate tweet and explicitly mentioned in the tweet using the mention sign (@), e.g., *usr* in our example. We note that role labels are not mutually exclusive in our dataset; a HI account may be a HT in another hate tweet.

Data and Methods

Despite the existence of a body of work dedicated to detecting hate speech (Schmidt and Wiegand 2017), accurate hate speech detection is still extremely challenging (CNN Tech 2016). A key problem is the lack of a commonly accepted benchmark corpus for the task. Each classifier is tested on a corpus of labeled comments ranging from a hundred to several thousand. Another option for collecting a dataset is filtering comments based on hate terms and annotating them. This is challenging because (i) annotation is time consuming and the percentage of hate tweets is very small relative to the total; and (ii) there is no consensus on the definition of hate speech (Sellars 2016). Some work has distinguished between profanity, insults and hate speech (Davidson et al. 2017), while other work has considered any insult based on the intrinsic characteristics of the person (e.g. ethnicity, sexual orientation, gender) to be hate speech related (Warner and Hirschberg 2012).

This annotation process can become even harder for role labeling, i.e., annotating actors as instigators, targets, bystanders (Xu et al. 2012). This is particularly challenging for social networking APIs that do not provide the whole thread of the conversation but only a random sample of comments, as in the case of the Twitter Streaming API. In this

[3]We replace select vowels with the star (*) character in obscene language.

[4]We anonymize all user mentions by replacing them with *@usr*.

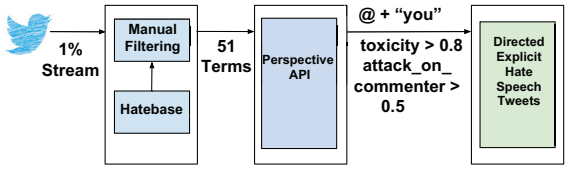


Figure 1: Flowchart of the filtering process used to obtain our dataset.

work, we adopt a definition of hate speech inspired by Facebook's community standards (Facebook 2016) and Twitter's hateful conduct policy (Twitter 2016) as "*direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease*." To mitigate the aforementioned challenges, we collect our own explicit Twitter hate speech dataset. We describe our semi-automated detection approach for directed explicit hate speech in the following subsections.

Data Collection

(1) Key phrase-based dataset (HS-1%): We adopt a multi-step classification approach. First, we use Twitter's Streaming API[5] to procure a 1% sample of Twitter's public stream from January 1st, 2016 to July 31st, 2017. We began by inspecting hate speech keyphrases in the Hatebase repository[6], the world's largest online repository of structured, multilingual, usage-based hate speech[7]. Online users can contribute to Hatebase by adding new derogatory words or phrases, their meaning, and language. Hatebase asks users who add terms to classify the term under one or more of the following hate categories: archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation (SexOrient). We use Hatebase as a lexical resource to retrieve English hate terms, broken down as: 42 archaic terms, 57 class, 7 disability, 427 ethnicity, 13 gender, 147 nationality-related, 38 religion, and 9 related to sexual orientation. After careful inspection and five iterations of keyword scrutiny, we removed keyphrases that resulted in tweets with uses distinct from hate speech or phrases that were extremely context sensitive. For example, the word "pancake" appears in Hatebase, but clearly can be used in benign contexts. Since our goal was a high quality dataset, we only included key phrases that were highly likely to indicate hate speech. The result is 8, 8, 2, 12, 4, 11, 4, and 2 keyphrases for the above, respective, hate speech classes. Due to the sheer volume of Twitter data, our main focus is to curate a relevant and accurate hate speech dataset with minimal amount of noise.

Despite the qualitative inspection of the keyphrases, when we used the resultant keyphrases to filter tweets from the 1% public stream, non-hate speech tweets remained in our dataset. To mitigate the effects of obscure contexts and stance on the filtering process, we were in need of a hate speech classifier that could remove non-hate speech tweets. Consider the following two tweets:
(a): "*@usr_1 i'll tear your limbs apart and feed them to the f*cking sharks you n*gger*"
(b): "*@usr_2 what influence?? that you can say n*gger and get away with it if you say sorry??*.
While both of these tweets contain the word "n*gger", the first tweet (a) is pro-hate speech where the hate instigator is attacking *usr_1*; the second tweet (b) is anti-hate speech in which the tweet author denounces the comments of *usr_2*. Thus stance detection is vital to consider when classifying hate speech tweets. To mitigate the effects of obscure contexts and stance with respect to hate speech on the filtering process, we used the Perspective API[8] developed by Jigsaw and the Google Counter-Abuse technology team, the model for which is comprehensively discussed in (Wulczyn, Thain, and Dixon 2017).[9]

The Perspective API contains different models of classification including: toxicity, attack of commenter, inflammatory, and obscene, among others. When a request is sent to the API with specific model parameters, a probability value [0, 1] is returned for each model type. For our datasets, we focus on two models: `toxicity` and `attack_on_commenter`. The `toxicity` model is a convolutional neural network trained with word-vector inputs. It measures how likely a comment will make people leave a discussion. The `attack_on_commenter` model measures the probability a comment is an attack on a fellow commenter and is trained on a New York Times dataset tagged by their moderation team. After inspecting the `toxicity` and `attack_on_commenter` scores for the tweets filtered by the Hatebase phrases, we found that a threshold of 0.8 for `toxicity` scores and 0.5 for `attack_on_commenter` scores yielded a high quality dataset.

As a final step to ensure that the resultant tweets attacked a specific Twitter user, we took the remaining tweets in our hate dataset and retained only those tweets that both mention another account (@) and that contain second person pronouns (e.g., "you", "your", "u", "ur"). The use of second person pronouns has been found to occur with high prevalence in directed hostile messages (Spertus 1997). The result of applying these filters is a high precision hate speech dataset of 27,330 tweets in which HIs use explicit Hatebase expressions against HTs. Figure 1 depicts the filtering process along with our workflow.

(2) General dataset (Gen-1%): To provide a larger context for interpretation of our analyses, we compare data from the HS-1% dataset with a random sample of all general Twitter accounts. To create this dataset, we use the Twitter Streaming API to obtain a 1% sample of tweets posted per day within the same 18 month collection window and extract the union set of users who posted them. We then remove ac-

[5]Twitter Streaming APIs: https://dev.twitter.com/streaming/overview

[6]Hatebase: https://www.hatebase.org/

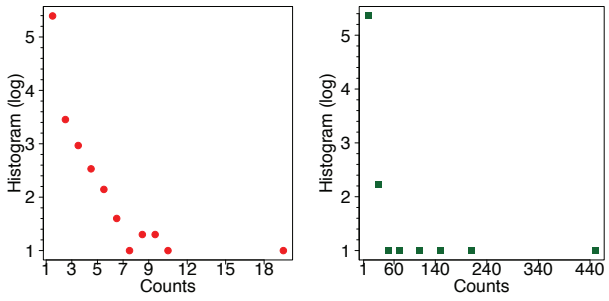
[7]We refer to hate speech terms as keyphrases, keywords, hate terms and hate expressions, interchangeably.

[8]Conversation AI source code: https://conversationai.github.io/

[9]We also experimented with classifiers including (Davidson et al. 2017) but found Perspective API to be empirically better.

	Total Unique Users		Suspended		Deleted	
HS Type	HI	HT	HI (%)	HT (%)	HI (%)	HT (%)
Archaic	169	169	8.3	11.2	4.1	4.1
Class	849	837	10.0	7.3	4.9	4.4
Disability	8,044	7,930	11.8	6.7	5.7	4.3
Ethnicity	2,073	2,045	18.8	11.3	6.6	5.2
Gender	13,195	13,340	9.4	5.7	5.6	4.7
Nationality	78	79	9.0	11.4	6.4	3.8
Religion	45	47	13.3	19.1	13.3	2.1
SexOrient	3,638	3,584	15.3	9.0	6.9	6.0
HS-1%	25,278	22,857	12.8	8.3	6.5	5.7
Gen-1%	60,000		5.2		3.2	

Table 1: Suspended and deleted accounts for all datasets.



(a) Tweets by instigators (b) Tweets against targets

Figure 2: Frequency of hate tweets in HS-1%.

counts appearing in the HS-1% dataset, and randomly sample 60K of the remaining users. To mitigate the bias towards more active users, we sample from the union set of users to ensure equiprobable selection of all users, regardless of activity level. While we try our best to remove all the bias, we acknowledge the possibility that this set might include some HIs and HTs. However, later our results show that this bias is likely to have have little impact because we observe significant differences between characteristics of HIs and HTs compared to the general dataset.

Table 1 shows the number of users in each of our datasets. In total, our dataset includes 25,278 hate instigators and 22,857 targets. The table shows the quantity of hate tweets for different hate classes.

The number of keywords used for identifying each class of hate can have an impact on the number of detected HIs and HTs. However, we observe that some classes with fewer keywords, such as *gender*, *disability* and *sexual orientation*, with 4, 2 and 2 keywords, have a higher contribution to our dataset, with 52%, 32% and 14% of HIs. This shows the prevalence of these hate keywords on Twitter.

Table 1 also shows the percentages of suspended and deleted accounts. The Twitter API returns an error message when the user account is suspended or the user is not found. According to Twitter, account suspensions occur when the account is spam, its security is at risk, or it is engaged in abusive tweets or behaviors. Twitter accounts that are not found (deleted) occur when the user does not exist. This error could arise for a variety of reasons: the user deactivated their account, the account was permanently deleted after thirty days of deactivation, etc. We label users that no longer exist as *deleted*. On average, suspended accounts comprise 12.8% of instigators, 8.3% of targets, and 5.2% of general Twitter users. Additionally, on average, deleted accounts comprise 6.5% of instigators, 5.7% of targets, and 3.2% of general Twitter accounts. Our findings show that instigators and targets are more likely to have their accounts suspended or deleted than general Twitter users, with instigators as the most likely.

Across each hate class, approximately 5% of accounts are deleted. The only exception is the *Religion* class, where 13% of hate instigator accounts are deleted. However, this may be the result of the small sample from this class. Interestingly, it seems Twitter is more successful in detecting hate related to *Ethnicity*, *SexOrient* and *Religion* as these categories have the highest number of suspended instigator accounts.

Many account holders in HS-1% either post more than one hateful tweet, or are hate targets more than once. Further, we identify 2,077 (approximately 5%) accounts that are both hate instigators and targets. Figure 2a illustrates the logarithmic histogram for the number of hate tweets posted by each instigator account. In our HS-1% dataset, about 10% of instigator accounts have posted more than one hate tweet. Figure 2b illustrates the histogram representing the number of hate tweets against other accounts. Approximately 11% of accounts are mentioned in more than two tweets, while two specific accounts are mentioned in 449 and 210 hate tweets.

Human-centered dataset evaluation. We evaluate the quality of our final dataset by incorporating human judgment using Crowdflower. We provided annotators with a class balanced random sample of 1000 tweets.[10] To aid annotation, all annotators were provided a set of precise instructions. This included the definition of hate speech according to the social media community (Facebook and Twitter) and examples of hate tweets selected from each of our eight hate speech categories. Then, for each tweet, we asked annotators two questions: (1) whether the tweet is hate speech, and (2) whether the tweet is a direct attack towards the account mentioned in the tweet. To improve the quality of responses, before assigning a task to annotators, we asked them five test questions with already known responses. If they could not answer at least 80% of these questions correctly, we identified them as unreliable annotators and removed them from the task. Each tweet was labeled by at least three independent Crowdflower annotators.

Using the majority vote, we found that annotators labeled 97.8% of the tweets as hate speech and 94.3% of tweets as an attack towards the mentioned account. We then evaluated the inter-annotator reliability by measuring the agreement percentage of annotators for each of the questions. We found that the agreement percentage for the first question is 92.8%, and for second question is 82.6%. These results shows that our hate speech dataset is reliable with minimal noise.

Measures

We adopt several measures based on prior work to answer our research questions. To compare the account characteristics of HIs and HTs, we investigate whether users have a

[10]We used a random sample of 1000 tweets to keep the monetary cost manageable.

	Gen-1% users				HIs				HTs			
Statistic	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median
Followers count	932	0	4,589,177	93	1,358	0	1,006,790	259	229,676	0	102,008,153	857
Friends count	408	0	243,937	160	663	0	1,012,412	239	1,897	0	1,698,640	396
Tweets count	4,384	0	570,550	545	14,160	0	4,321,652	3,266	29,559	1	3,644,240	10,902
Listed count	8	0	10,118	0	13	0	7,855	2	755	0	616,271	9
Retweet counts	3	0	13,220	0	30	0	27,390	2	623	0	304,900	10
Account age (years)	3.73	0.09	10.99	3.33	3.67	0.09	10.66	3.22	4.40	0.09	11.37	4.16
len. description (chars)	45	0	164	28	53	0	164	37	63	0	164	49
Profile image	0.95	0	1	NA	0.97	0	1	NA	0.99	0	1	NA
Profile URL	0.23	0	1	NA	0.24	0	1	NA	0.40	0	1	NA
Geo location	0.33	0	1	NA	0.39	0	1	NA	0.51	0	1	NA
Location	0.53	0	1	NA	0.61	0	1	NA	0.69	0	1	NA
Timezone	0.40	0	1	NA	0.52	0	1	NA	0.68	0	1	NA
Verified	0.003	0	1	NA	0.002	0	1	NA	0.12	0	1	NA
	$N = 60,000$				$N = 25,278$				$N = 22,857$			

Table 2: Descriptive statistics of our datasets.

profile image, set a geo-location and a timezone, whether the account is verified, and the length of the profile description. We study the number of tweets and retweets, friends, followers, and whether the account is enlisted. Similar to Nilizadeh *et al.* (Nilizadeh et al. 2016), we differentiate accounts by *perceived*, as opposed to *actual*, user characteristics. This is because we can only study how an account holder chooses to represent him/herself, i.e., through a profile photo, and cannot determine their actual characteristics.

We predict user gender by extracting first names and comparing them with those listed in the 1900 – 2013 U.S. Census (Mislove et al. 2011; Nilizadeh et al. 2016). We leverage the IBM Watson Personality Insights API (IBM Bluemix Docs 2015b) to quantify the Big Five personality traits for HIs and HTs. The API has been used in prior studies to correlate personality traits with information-spreading (Lee et al. 2014) and targeted advertising (Chen et al. 2015).

Analysis

RQ1: Account Characteristics

Our first objective is to understand the differences of self presentation through profile configurations, activity level, and interaction with other users. To study profile presentation, we analyze whether profile image, location, and timezone are provided by the user; whether the user has enabled the geo-location to be posted along with their tweets; whether the account is verified by Twitter; and the length of profile description in characters. For user activity level, we analyze number of tweets, friends, followers, lists, and retweets. The last three of these indicate how Twitter users interact with an account and are used as visibility measures (Nilizadeh et al. 2016; Sharma et al. 2012).

All characteristics can be extracted from the meta-data provided with the tweets, except the retweet count. For every user, we count the number of times the user's tweets are reposted in our 1% dataset. Although the obtained retweet counts only represent a subset of the actual retweets, they provide useful insight when comparing different samples.

We determine the gender of users by extracting first names and comparing them with first names listed in the U.S. Census dataset obtained from 1900 – 2013 (Mislove et al. 2011). Some first names are gender-neutral, such as "Pat." Similar to other work (Mislove et al. 2011), if a name has a female-to-male ratio larger than 0.95 or smaller than 0.05, we label it as female or male; other names are labeled as 'gender ambiguous'. We are able to extract first names for 53% of HIs, 55% of HTs and 56% of general users. HIs use pseudonyms more than others, which can be an indication of desire to hide their identities. 25%, 23% and 8% of users in the Gen-1% dataset; 35%, 10% and 8% of users in the instigator dataset; and 35%, 12% and 8% of users in the instigator dataset are male, female and gender ambiguous, respectively. Instigator and target datasets include 10% more male and 13% fewer female users than the Gen-1% dataset, which implies that *users with female account names are less engaged in hate discussions.*

Table 2 statistically describes the users in our Gen-1% and HS-1% datasets. Since the distribution of most characteristics is skewed, in addition to mean, the table also shows the min, max and median of values. The table illustrates multiple differences between user types. The t-tests for account age (by year) suggest that, on average, the accounts for HTs are older than those of HIs ($\mu = 4.40$, vs. $\mu = 3.67$) ($t = 32.18, p < 0.001$) and generic random users ($\mu = 4.40$, vs. $\mu = 3.73$) ($t = 32.91$, $p < 0.001$). Also, the accounts for HIs are younger than those of general random users ($\mu = 3.67$ vs. $\mu = 3.73$) ($t = 3.33$, $p < 0.001$). We observe that compared to random users, HIs and HTs are more active in becoming friends with others, posting tweets, and providing more content on their profiles.

The t-tests for profile description length (in characters) show that, on average, the descriptions provided by HTs are longer than those for HIs ($\mu = 63$, vs. $\mu = 53$) ($t = 20.14$, $p < 0.001$). The descriptions provided by hate targets and instigators are longer than those of generic random users ($\mu = 63$, vs. $\mu = 45$) ($t = 40.04$, $p < 0.001$), ($\mu = 53$, vs. $\mu = 45$) ($t = 19.56$, $p < 0.001$). These results may suggest that both HIs and HTs are more willing to present themselves.

Table 3 shows the results of Chi-square tests for the binary variables. In general, HTs reveal more information on their profiles; they are more likely to add image, URL, location and timezone to their profiles compared to both HIs and general Twitter users. There is only one exception where the difference between the distribution of geo-location for HIs and that of HTs is not significant ($p = 0.06$).

Twitter verifies accounts that are of public interest. When accounts are verified, a blue badge appears next to the user's

		HT vs. HI		Gen-1% vs. HT		Gen-1% vs. HI	
	df	X^2	p	X^2	p	X^2	p
Profile image	1	7633	***	672	***	4901	***
Profile URL	1	325	***	1858	***	3546	***
Geo location	1	3.53	0.06	1937	***	1801	***
Location	1	1606	***	1389	***	66	***
Timezone	1	1389	***	4444	***	797	***
Verified	1	99	***	6226	***	4789	***
Gender (name)	1	1318	***	1230	***	21	***
Invalid image	1	2,088,900	***	1,221	***	4,827,400	***
Detected face	1	1,138,200	***	505	***	1,821,700	***
Multiple faces	1	282,530	***	127	***	368,000	***
One face (Male)	1	289,160	***	24,493	***	224,900	***
One face (Female)	1	270,580	***	197,900	***	933,780	***

Note: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 3: Pearson's Chi square tests.

	U (HT vs. HI)	U (Gen-1% vs. HT)	U (Gen-1% vs. HI)	p
Followers	321,900K	183,400K	504,620K	***
Tweets	294,930K	190,920K	445,380K	***
Friends	278,670K	316,970K	586,540K	***
Lists	305,450K	221,840K	503,890K	***
Retweets	304,560K	139,270K	369,650K	***

Table 4: Mann-Whitney U tests.

name on their profile.[11] Interestingly, when comparing HIs and HTs, we observe that HTs include significantly more high profile and established users; 12% belong to verified accounts. However, HIs themselves are less likely to have verified accounts, even compared to random general users.

Next, we examine the activity and visibility levels of account holders. We compare these variables by using Mann-Whitney U tests, because they do not follow a normal distribution. These results are provided in Table 4. Interestingly, HTs have more friends and post more tweets than both HIs and general users. They also have higher visibility and influence; their median numbers of followers and retweets are larger than those of both HIs and general users.

Twitter's 'List' feature allows users to organize others by creating topical user lists. If some users are known for something, *e.g.,* are computer scientists, then they might be listed by others in "Computer Scientists" list. Organizing Twitter users into lists helps track tweets from those in the list. Our results show that targets of hate are listed more often.

Figure 3 compares the distribution of the activity and visibility characteristics of HIs and HTs with those from the Gen-1% dataset. This figure shows CCDF plots for variables that exhibit heavy-tailed distributions. Figure 3a shows that HTs on average have more followers than both HIs and general Twitter users, while the distribution of followers count for HIs is more similar to that of general Twitter users. Specifically, the difference between HTs and others is more significant for visibility measures including followers, lists and retweet counts.

Visibility: We next examine the visibility of HIs and HTs by controlling for variables that can have an impact on the visibility measures. For example, older accounts have had more time to accumulate followers; following many others usually yields more followers by sheer reciprocity; and posting many tweets can increase the chances to be

[11]Request to verify an account:
https://support.twitter.com/articles/119135#

	Followers count				
	Poisson	0.25 Qrt.	0.5 Qrt.	0.75 Qrt.	1.00 Qrt.
HT	2.68***	0.41***	0.10***	0.05***	2.36***
IRRs	14.64	1.51	1.11	1.05	10.60
	Lists count				
HT	1.93***	0.04	0.08***	0.06***	1.59***
IRRs	6.92	1.036	1.08	1.06	4.92
	Retweet count				
HT	4.06***	2.57***	4.18***	3.76***	3.35***
IRRs	57.94	13.00	65.01	42.98	28.53

Table 5: HTs vs. All Poisson Regressions.

	Followers count	*Lists count*	*Retweet count*
HT (IRRs)	2.03*** (7.65)	1.59*** (4.90)	3.15*** (23.32)

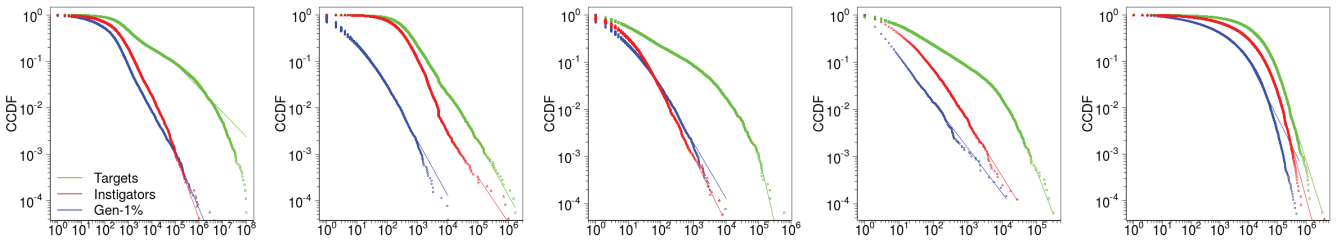
Table 6: HTs vs. HIs Poisson Regressions.

noticed. Thus, we incorporate the following control variables in our models: account age, number of tweets, number of friends, and profile characteristics such as URL, location, image, length of user description, timezone and verified, as well as perceived user gender. We control for profile characteristics and gender because user self-presentation can affect the way people perceive them, and therefore, can have an impact on visibility measures (Ridgeway 2001; Nilizadeh et al. 2016).

We select three dependent variables as the main measures of online visibility on Twitter: 'number of followers', 'retweets,' and 'lists.' We apply multiple multivariate regression models and present the results from our Poisson regression model. Linear and negative binomial regression models show qualitatively consistent results, although a couple did not converge.

Since our dependent variables exhibit a skewed distribution, examining the whole population may not capture more nuanced patterns (Yu, Lu, and Stander 2003). For example, in Table 2, we observe that a hate target account holder has more than 100M followers and this user alone can impact the overall and average statistical results. Thus, we adopt the quartile regression technique to analyze our dataset in each quartile. We divide the data into quartiles based on each dependent variable and apply multivariate regression models. Although we include control variables in all models, for brevity, we omit them from the result tables; full tables are available upon request. We add followers count as a control for the retweets and lists count models because more followers may result in being retweeted and listed more. We add lists count as a control for the retweets count model because being listed by many people may result in being retweeted more. We report Incident Rate Ratios (IRRs), the exponentiated coefficients of Poisson regressions, which allow us to compare the rates of variables between HIs, HTs, and general users.

Table 5 shows the results of Poisson regression comparing HTs vs. the union of HIs and general users. The first column shows the result for the entire sample such that HTs have significantly more followers, are listed and retweeted



(a) Followers count (b) Friends count (c) Listed count (d) Retweets count (e) Tweets count

Figure 3: Comparison of account characteristics of HIs, HTs, and general users.

	Followers count	*Lists count*	*Retweet count*
HI (IRRs)	0.46*** (1.59)	0.49*** (1.62)	1.98*** (7.26)

Table 7: HIs vs. Gen-1% Poisson Regressions.

more than all other users ($p < 0.001$). Particularly, for followers, lists and retweet counts, the HTs have IRRs 14.64, 6.92 and 57.94 times of those of the union of HIs and general users. Table 6 illustrates that these findings hold even when HTs are compared only with HIs ($p < 0.001$). *These results suggest that regardless of user activity level, profile self-presentation, and gender, more visible Twitter users (with more followers, lists, and retweets) are more likely to become target of hate.*

Table 7 demonstrates the results of models for HIs vs. general users. The coefficients for both overall and quartiles models are positive and larger than one, which indicate that HIs are positively associated with being visible.

In Table 5, quartile regression reveals that the overall and average results are not just the effects of most visible users, and in each quartile, the HTs are more visible than HIs and general users. Although the effect of HTs (IRR) increases as one moves from the least visible to most visible users, in almost all quartiles values are larger than one. For brevity, we do not report the results of models per quartiles in Tables 6 and 7 although the interpretation of their results is consistent with those reported for Table 5.

Comparing the IRR results with those in Tables 5 and 6 shows that the differences between the HTs and HIs are significantly higher than those of HIs and general users. *These results also suggest that participating in hate speech and being more visible and popular are related; even when controlling for all mentioned independent variables, both HIs and HTs are more popular and visible than general users.*

RQ2: Personality Traits

To study the key differences between the personalities of HIs, HTs, and the general population, we use the Twitter REST API to fetch tweet traces of users. A Twitter user can share content on their profile in three different ways: an original tweet, a reply to a tweet written by another user, or a redistribution of a tweet written by another account (retweeting). Retweets do not necessarily indicate content endorsement but suggest content to be viewed by the retweeter's network. Since retweeting content might not reflect the author's point of view, we only include original tweets and replies as part of our personality analysis. We attempt to fetch the most recent 2000 tweets (excluding retweets) for each account. We use IBM Watson Personality Insights API[12] for our personality analysis. Since the Personality Insights API requires a minimum of 600 words to obtain statistically significant result estimates, we discard any accounts that do not satisfy this requirement. After discarding suspended and deleted accounts, accounts with statistical insignificance, and accounts with languages other than English, we were able to fetch tweets for a total of 17,951 unique HIs, 17,553 unique HTs, and 12,900 unique general users (pulled from Gen-1%).[13] We use the general users personality results as a means of account sample representation on Twitter. The word count distribution is ($\mu = 11,045.6, \sigma = 7,230.5$) for HI accounts, ($\mu = 12,316.1, \sigma = 7,308.7$) for HT accounts, and ($\mu = 8,108.2, \sigma = 7,288.7$) for accounts in Gen-1%.

The IBM Watson Personality API infers personality characteristics from textual information based on an open-vocabulary approach (IBM Bluemix Docs 2015b). The API's machine learning algorithm is trained using scores obtained from surveys conducted among thousands of users along with data from their Twitter feeds. The API provides scores [0, 1] that reflect the normalized percentile score for the characteristic. We analyze the results of the *Big Five* personality model, the most widely used model for generally describing how a person engages with the world. The model includes five primary dimensions: Agreeableness, Conscientiousness, Extraversion, Emotional range, and Openness. The Big Five personality traits, their associated facets, and how to interpret them are defined in detail in (IBM Bluemix Docs 2015a).

To quantify the difference between the continuous distributions of different personality aspects, we compute the Hellinger distance (Tanton 2005). The Hellinger distance between two measures P and Q represented by two distributions $f(x)$ and $g(x)$, respectively, is defined as:

[12] https://www.ibm.com/watson/services/personality-insights/

[13] All sampling errors in our results are less than 0.1.

	Medians			HI vs. HT		HI vs. Gen-1%		HT vs. Gen-1%		Hellinger distances		
Personality facet	HI	HT	Gen-1%	U	p	U	p	U	p	HI-HT	HI-Gen-1%	HT-Gen-1%
Agreeableness	0.06	0.1	0.4	134,790K	***	47,512K	***	61,130K	***	0.11	0.37	0.27
Openness	0.49	0.51	0.5	152,400K	***	114,760K	0.18	115,840K	***	0.03	0.03	0.04
Emotional range	0.18	0.22	0.38	142,360K	***	77,917K	***	87,490K	***	0.08	0.22	0.15
Conscientiousness	0.02	0.05	0.31	128,370K	***	35,667K	***	55,020K	***	0.18	0.46	0.31
Extraversion	0.23	0.31	0.47	149,410K	***	83,693K	***	88,067K	***	0.04	0.17	0.13

Note: *$p < 0.05$ **< 0.01 ***< 0.001

Table 8: Scores and Hellinger distances for the Big Five personality traits of HIs, HTs and general users.

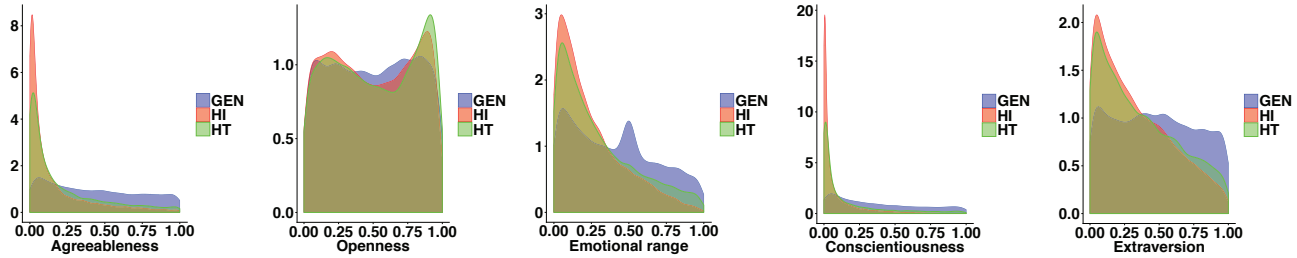


Figure 4: Distribution of scores for the Big Five personality traits.

$$H(P,Q) = \sqrt{\frac{1}{2}\int (\sqrt{f(x)} - \sqrt{g(x)}\,)^2\, dx}\,, \qquad (1)$$

where $H(P,Q) \in [0,1]$. The minimum distance of 0 is achieved when P and Q exhibit exactly the same distributions; the maximum distance of 1 is achieved when P assigns probability zero to every set to which Q assigns a positive probability, and vice versa. Table 8 depicts the pairwise distribution distances between HIs and HTs (HI-HT), and the distance between the HI and HT distributions and the general users, (HI-Gen-1%) and (HT-Gen-1%), respectively. We also report the results of the Mann-Whitney U tests.

HIs and HTs personalities differ from general users: For all the personality traits depicted in Table 8, the Hellinger distance of (HI-HT) is always less than or equal to (HI-Gen-1%) and (HT-Gen-1%). This indicates that HIs and HTs have more similar personalities to each other than general users. This is also shown for each personality trait's median. With the exception of Openness, the median for HIs personality facets is closer to the median of HTs than Gen-1%.

Both HIs and HTs exhibit lower Agreeableness than general users. Lower Agreeableness scores are often associated with suspicious and antagonistic behaviors (Toegel and Barsoux 2012). Our results indicate that HIs and HTs are more self-focused, contrary, proud, cautious of others, and can compromise morality.

While Figure 4 shows that the distributions for HIs, HTs, and general users are close (with a median of approximately 0.5), when we investigate Openness, we find discrepancies in the lower level facets: Adventurousness, Emotionality, and Imagination. Both HIs and HTs exhibit lower scores for Emotionality and Adventurousness, and higher Imagination scores, in comparison to the general users. Moreover, HIs and HTs have similar distributions for Artistic Interests ($p = 0.24$) and Liberalism ($p = 0.98$). These results indicate that HIs and HTs are less emotionally aware and less adventurous with a wild imagination (lower preference to facts), and more authority challenging behavior, in comparison to the general users.

For Emotional range, HIs and HTs have lower scores than general users across all facets. HIs have slightly lower scores, but still statistically significant, than HTs. The high Emotional range scores indicate that HIs and HTs are more fiery, prone-to worry, melancholy, hedonistic, and susceptible to stress. Cheng *et al.* observe that negative mood increased a user's probability to engage in trolling, and that anger begets more anger (Cheng et al. 2017). It seems that Emotional range facets such as Anxiety, Depression, Immoderation, and Self-consciousness are embodied more in the tweets of HIs and HTs but further work is needed to directly correlate these parameters with hate speech and online trolling.

For Conscientiousness, HIs and HTs generally have lower scores than general users. Consistently, HTs score slightly higher, but still statistically significant, than HIs. Our results suggest that HIs and HTs tend to disregard rules and obligations, as indicated by low dutifulness scores, and would rather take action immediately than spend time deliberating a decision, as indicated by low Cautiousness scores. As for Extraversion, HIs and HTs tend to have lower scores of Activity-level, Friendliness, and Cheerfulness but higher scores for Excitement seeking, in comparison to general users. Our results indicate that HIs and HTs are inclined to be less sociable, less assertive, and more solemn.

HIs and HTs tend to share personality facets: It is possible that the personality facets for HIs and HTs could contribute to the problem of hate speech. Our results show that indeed the personalities of HIs and HTs are much closer to each other than to the general users. Moreover, our results agree with prior work conducted for victims of bullying. Prior studies, in workplaces and schools, have shown that bullying victims tend to show depression and helplessness as a result of bullying (Price et al. 1994). Moreover victims are described as lacking social skills, tending to show emo-

tions, e.g., crying easily (Schwartz, Dodge, and Coie 1993), and are likely to experience anxiety, loneliness, and hyperactivity (Camodeca et al. 2003; Johnson et al. 2002). Our work also agrees with studies that show that bullies and victims share a wide range of bully-typifying personality traits such as machiavellianism, narcissism, psychoticism, and aggression, and that bullies and victims could exchange roles (Linton and Power 2013). Interestingly, in this work we have shown that these personality signals have been mirrored from the physical world and now have a presence in the digital world as well.

Discussion and Conclusion

Hate mitigation and counter speech. Successful counter speech is a direct response to hateful comments aimed at influencing discourse and behavior (Benesch 2014; Benesch et al. 2016). Recently, Munger showed that counter speech using automated bots can reduce instances of racist speech if instigators are sanctioned by a high-follower white male (Munger 2017). If AI-powered counter speech bots are widely deployed (Forbes 2017), a research challenge would then be how we can design these bots to achieve maximum impact. Prior work has shown that people respond more positively to messages tailored to their personality (Hirsh, Kang, and Bodenhausen 2012). For instance, Myszkowski and Storme correlated Openness with product design and found that individuals with low openness scores respond to product appearance and, conversely, high openness individuals tend to focus on product aspects (Myszkowski and Storme 2012). Our personality analyses could be used to design next generation counter speech bots of increased effectiveness. Moreover, our personality results show that 50% of HIs and HTs score above 0.53 for the Openness to change personality facet, which may imply that counter speech could be successfully used to decrease hate speech.

Profile-based data collection. Most common methods of data collection use hate terms and trained classifiers to classify new content as hateful or benign. Another method employs bootstrapping, which is used in (Xiang et al. 2012) to obtain training data by classifying Twitter accounts as either "good" or "bad" based on usage of offensive terms. All tweets from "bad accounts" are marked as hate speech instances. Our results could be incorporated through the use of personality scores as features to classify users. Alternatively, a user could be represented as a vector of personality facets and then compared to values for hate speech accounts. This could be especially useful for content curation for cases when the instigator is likely to engage in hate speech more than once (Xiang et al. 2012; Chatzakou et al. 2017b) or as features for early instigator identification (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) and implicit hate speech detection.

Critique of methodology and limitations. There are limitations to our methodology and findings. Recent studies (Tufekci 2014; Morstatter et al. 2013) discuss common issues associated with social media analysis and the sample quality of the Twitter Streaming API. Our analysis focused on explicit hate speech and relied on keyword-based methods, which have been shown to miss instances of hateful speech (Saleem et al. 2016). However, while we cannot claim to have captured a complete representation of hate speech on Twitter, as our starting point for tweet filtering was based on a set of hate terms from Hatebase, our primary objective was to investigate hate speech instigator and target accounts with a high precision dataset. We believe that our careful curation methodology achieved this end goal.

Conclusion. We have presented the first comparative study of hate speech instigators, targets, and general Twitter users. We have outlined a semi-automated classification approach for curation of directed explicit hate speech. Our analysis yields a number of interesting and unexpected findings about actors of hate speech. For example, we found that hate instigators target more visible users and that participating in hate commentary is associated with higher visibility. We also showed that hate instigators and targets have unique personality characteristics that may contribute to hate speech such as anger, depression, and immoderation. We hope that our results can be used as meta-information to improve hate speech classification, detection and mitigation to combat this increasingly pervasive problem.

References

Benesch, S.; Ruths, D.; P Dillon, K.; Mohammad Saleem, H.; and Wright, L. 2016. Considerations for Successful Counterspeech. Technical report, Evaluating Methods to Diminish Expressions of Hatred and Extremism Online as part of The Kanishka Project of Public Safety Canada.

Benesch, S. 2014. Countering Dangerous Speech to Prevent Mass Violence during Kenya's 2013 Elections. Technical report, United States Institute of Peace.

Burnap, P., and Williams, M. L. 2014. Hate Speech, Machine Classification and Statistical modeling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making. *Internet, Policy and Politics Conference*.

Camodeca, M.; Goossens, F. A.; Schuengel, C.; and Terwogt, M. M. 2003. Links between Social Information Processing in Middle Childhood and Involvement in Bullying. *Aggressive behavior* 29(2):116–127.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2018. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *CSCW'18*.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017a. Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter. In *HT'17*.

Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Vakali, A. 2017b. Mean Birds: Detecting Aggression and Bullying on Twitter. In *WebSci'17*.

Chen, J.; Haber, E. M.; Kang, R.; Hsieh, G.; and Mahmud, J. 2015. Making Use of Derived Personality: The Case of Social Media Ad Targeting. In *ICWSM'15*.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *CSCW'17*.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *ICWSM'15*, 61–70.

CNN Tech. 2016. Twitter Launches New Tools to Fight Harassment. https://goo.gl/AbYbMv.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM'17*.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Facebook. 2016. Controversial, Harmful and Hateful Speech on Facebook. https://goo.gl/TWAHdr.

Forbes. 2017. Fighting Social Media Hate Speech With AI-Powered Bots. https://goo.gl/79u6Yd.

Hamm, M. S. 1994. Conceptualizing Hate Crime in a Global Context. *Hate crime: International perspectives on causes and control* 173–194.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM'17*.

Hirsh, J. B.; Kang, S. K.; and Bodenhausen, G. V. 2012. Personalized Persuasion: Tailoring Persuasive Appeals to Recipients' Personality Traits. *Psychological science* 23(6):578–581.

IBM Bluemix Docs. 2015a. Personality Models. https://goo.gl/KzGgWa.

IBM Bluemix Docs. 2015b. The Science behind the Service. https://goo.gl/6SPwfr.

Johnson, H. R.; Thompson, M. J.; Wilkinson, S.; Walsh, L.; Balding, J.; and Wright, V. 2002. Vulnerability to Bullying: Teacher-reported Conduct and Emotional Problems, Hyperactivity, Peer Relationship Difficulties, and Prosocial Behaviour in Primary School Children. *Educational Psychology* 22(5):553–556.

Lee, K.; Mahmud, J.; Chen, J.; Zhou, M.; and Nichols, J. 2014. Who Will Retweet This?: Automatically Identifying and Engaging Strangers on Twitter to Spread Information. In *ACM IUI'14*.

Levin, J., and MacDevitt, J. 2013. *Hate Crimes: The Rising Tide of Bigotry and Bloodshed*.

Linton, D. K., and Power, J. L. 2013. The Personality Traits of Workplace Bullies are Often Shared by Their Victims: Is there a Dark Side to Victims? *Personality and Individual Differences* 54(6):738–743.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the Demographics of Twitter Users. In *ICWSM'11*.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM '13*.

Munger, K. 2017. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior* 39(3):629–649.

Myszkowski, N., and Storme, M. 2012. How Personality Traits Predict Design-driven Consumer Choices. *Europe's Journal of Psychology* 8(4):641–650.

Nilizadeh, S.; Groggel, A.; Lista, P.; Das, S.; Ahn, Y.-Y.; Kapadia, A.; and Rojas, F. 2016. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. In *ICWSM'16*.

Price, J.; Sloman, L.; Gardner, R.; Gilbert, P.; and Rohde, P. 1994. The Social Competition Hypothesis of Depression. *The British Journal of Psychiatry* 164(3):309–315.

Ridgeway, C. L. 2001. Gender, Status, and Leadership. *Journal of Social issues* 57(4):637–655.

Saleem, H. M.; Dillon, K. P.; Benesch, S.; and Ruths, D. 2016. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety*.

Schmidt, A., and Wiegand, M. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *SocialNLP'17: Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*.

Schwartz, D.; Dodge, K. A.; and Coie, J. D. 1993. The Emergence of Chronic Peer Victimization in Boys' Play Groups. *Child development* 64(6):1755–1772.

Sellars, A. 2016. Defining Hate Speech. Technical report, Berkman Klein Center for Internet and Society at Harvard University.

Sharma, N. K.; Ghosh, S.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Inferring Who-is-Who in the Twitter Social Network. In *WOSN'12: Proceedings of the 2012 ACM Workshop on Online Social Networks*.

Silva, L. A.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the Targets of Hate in Online Social Media. In *ICWSM'16*.

Spertus, E. 1997. Smokey: Automatic Recognition of Hostile Messages. In *AAAI'97*.

Statista. 2017. Twitter, Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions). https://goo.gl/HE6K3E.

Tanton, J. 2005. *Encyclopedia of Mathematics*. Facts On File.

Toegel, G., and Barsoux, J.-L. 2012. How to Become a Better Leader. *MIT Sloan Management Review* 53(3):51.

Tufekci, Z. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and other Methodological Pitfalls. In *ICWSM'14*.

Twitter. 2016. Hateful Conduct Policy. https://goo.gl/NxR4sR.

Warner, W., and Hirschberg, J. 2012. Detecting Hate Speech on the World Wide Web. In *ACL'12: Proceedings of the 2nd Workshop on Language in Social Media*.

Waseem, Z., and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL Student Research Workshop*.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *arXiv preprint arXiv:1705.09899*.

Williams, M. L., and Burnap, P. 2015. Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology* 56(2):211–238.

Wired. 2016. Inside Google's Internet Justice League and its AI-Powered War on Trolls. https://goo.gl/Nvf6ZA.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *WWW'17*.

Xiang, G.; Fan, B.; Wang, L.; Hong, J.; and Rose, C. 2012. Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In *CIKM'12*.

Xu, J.-M.; Jun, K.-S.; Zhu, X.; and Bellmore, A. 2012. Learning from Bullying Traces in Social Media. In *NAACL'12*.

Yu, K.; Lu, Z.; and Stander, J. 2003. Quantile Regression: Applications and Current Research Areas. *Journal of the Royal Statistical Society* 52(3):331–350.