Characterizing Usage of Explicit Hate Expressions in Social Media

Mainack Mondal[a], Leandro Araújo Silva[b], Denzil Correa[c] and Fabrício Benevenuto[b]

[a]University of Chicago, Chicago, IL, USA; [b]Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; [c]Max Planck Institute for Software Systems, Germany

ARTICLE HISTORY
Compiled June 21, 2018

ABSTRACT

Social media platforms provide an inexpensive communication medium that allows anyone to publish content and anyone interested in the content can obtain it. However, this same potential of social media provide space for discourses that are harmful to certain groups of people. Examples of these discourses include bullying, offensive content, and hate speech. Out of these discourses hate speech is rapidly recognized as a serious problem by authorities of many countries. In this paper, we provide the first of a kind systematic large-scale measurement and analysis study of explicit expressions of hate speech in online social media. We aim to understand the abundance of hate speech in online social media, the most common hate expressions, the effect of anonymity on hate speech, the sensitivity of hate speech and the most hated groups across regions. In order to achieve our objectives, we gather traces from two social media systems: Whisper and Twitter. We then develop and validate a methodology to identify hate speech on both of these systems. Our results identify hate speech forms and unveil a set of important patterns, providing not only a broader understanding of online hate speech, but also offering directions for detection and prevention approaches.

KEYWORDS
hate speech, anonymity, social media, Whisper, Twitter, pattern recognition

1. Introduction

Online social media sites today allow users to freely communicate at nearly marginal costs. Increasingly users leverage these platforms not only to interact with each other, but also to share news. While the open platforms provided by these systems allow users to express themselves, there is also a dark side of these systems. Particularly, social networks have become a fertile ground for inflamed discussions that usually polarize 'us' against 'them', resulting in many cases of insulting and offensive language.

Another important aspect that favors such behavior is the level of anonymity that some social media platforms grant to users. As example, "Secret" (*Apple Pulls Secret App in Brazil After Judge's Request*, 2014) was created, in part, to promote free and anonymous speech but became a mean for people to defame others while remaining

CONTACT Mainack Mondal. Email: mainack@uchicago.edu
Leandro Araújo Silva. Email: leandro@dcc.ufmg.br
Denzil Correa. Email: denzil@mpi-sws.org
Fabrício Benevenuto. Email: fabricio@dcc.ufmg.br

anonymous. Secret was banned in Brazil for this very reason and shut down in 2015 [1]. There are reports of cases of hateful messages in many other social media independently of the level in which the online identity is bonded to an offline identity – e.g., in Whisper (Silva, Mondal, Correa, Benevenuto, & Weber, 2016), Twitter (Sanchez & Kumar, 2011), Instagram (Hosseinmardi et al., 2015), and Facebook (Kwan & Skoric, 2013).

With this context, it is not surprising that most existing efforts are motivated by the impulse to detect and eliminate hateful messages or hate speech (Agarwal & Sureka, 2015; Bartlett, Reffin, Rumball, & Williamson, 2014; Gitari, Zuping, Damien, & Long, 2015; Stephens, 2013; Ting, Chi, Wu, & Wang, 2013). These efforts mostly focus on specific manifestations of hate, like racism (Chaudhry, 2015). While these efforts are quite important, they do not provide the big picture about the problem in the current popular social media systems. More importantly, these efforts do not provide any understanding about the root causes of online hate speech and do not provide any insight on how to deal with the underlying offline hate.

In this paper we take a first step towards better understanding online hate speech. In this work, we mainly focus on explicit expression of hate as hate speech. Our effort consists of characterizing how explicitly posted hate messages are spread in common social media. We aim to understand how hate speech manifests itself under different dimensions such as its targets, the identity of the haters, geographic aspects of hate contexts. Particularly, we focus on the following research questions.

What is hate speech about? We want to understand not only which are the most common hated groups of people, but also what are the high level categories of hate targets in online hate speech.

How does online and offline hate correlate? Since hate speech is a problem that plagues the offline world for long time, does the online hate speech correlated with its offline counterpart?

What role does anonymity play on hate speech? Is anonymity a feature that exacerbates hate speech or are social media users not worried about expressing their hate under their real names? What fraction of haters use their personal names in social media?

How do users perceive different categories of hate speech? Since online hate speech might be directed towards a wide category of hate targets, do online users perceive some categories of hate speech as more sensitive than others?

How does hate speech vary across geography? Does hate speech targets vary across countries? And within states of a country, like US? Are there categories of hate speech that are uniformly hated and others that are hated only in specific regions?

Answering these questions is crucial to help authorities (including social media sites) for proposing interventions and effectively deal with hate speech. To find answers, we gathered one-year data from two social media sites: Whisper and Twitter. Then, we propose and validate a simple yet effective method to detect hate speech using sentence structure and using this method construct our hate speech datasets. Using this data, we conduct the first of a kind characterization study of hate speech along multiple different dimensions: hate targets, the identity of haters, geographic aspects of hate and hate context. Our results unveil a set of important patterns, providing not only a broader understanding of hate speech, but also offering directions for detection and

[1] http://www.bbc.com/news/technology-32531175

prevention approaches.

The rest of the paper is organized as follows: Next, we briefly discuss related efforts in this field. Then, we present our whisper and Twitter datasets and our approach to identify and measure hate speech in them. The next sections provide a series of analysis results that answer our research questions stated before. We conclude the paper discussing some potential implications of our findings.

2. Related work

We start with reviewing existing work on hate speech along two dimensions.

2.1. Understanding hate speech

Hate speech has been an active research area in the sociology community (Delgado & Stefancic, 2004). Particularly, Massaro (Massaro, 1990) claims that some forms of hate speech are far from being solved in our society, especially those against black people and women. A recent effort from (Waldron, 2012) discusses proposals for regulations to suppress hate speech. They highlight the importance of such regulations as part of commitment to human dignity and to inclusion and respect for members of vulnerable minorities. Their proposed regulations also argue about the security of hate targets in an environment polluted by hate speech. Very recently a UNESCO supported study (Gagliardone, Gal, Alves, & Martinez, 2015) reviewed the problem of growing hate speech. They suggested that platforms like Facebook and Twitter have primarily adopted only a reactive approach to deal with hateful messages reported by their users, but they could do much more with the huge data available to them. In our work, we concentrate our efforts towards exactly this direction.

2.2. Detecting hate speech in online media

In recent years, there has been a number of studies which focus on computational methods to find hate speech in social media. Faris et al. (Faris, Ashar, Gasser, & Joo, 2016) reviewed these techniques and identified approaches which range from computational to legal or sociological (interviews and lab studies). Currently multiple researchers try to detect hate speech using manual inspection or a mix of crowdsourced labeling and machine learning techniques (Agarwal & Sureka, 2015; Bartlett et al., 2014; Gitari et al., 2015; Greevy & Smeaton, 2004; Reis et al., 2015; Stephens, 2013; Ting et al., 2013; Warner & Hirschberg, 2012).Their basic framework consists of creation of a corpus which contains a set of known hate keywords and then manually label that corpus to construct a training dataset with hate posts and non-hate posts. Then they use this corpus as training dataset to build automated systems (via machine learning approaches) to detect hate speech. Overall, these types of approaches have two shortcomings. Firstly, it is hard to detect new hate targets using hate keywords. Secondly, manual labeling, although useful, but is not scalable if we want to understand and detect hate speech at large scale. A very recent work by Chandrasekharan et al. (Chandrasekharan et al., 2017) extracted hate keywords by comparing posts from hate driven communities with posts from non-hate communities. However, their work is platform (Reddit) and hate type (shaming overweight people, hate against African Americans) specific. In short, although these efforts offer advances in this

field, computational methods to detect hate speech are in a nascent stage.

Most of these prior efforts focus on detecting online hate speech. Differently, our research goal is to use computational techniques to *understand* the social phenomena of online hate speech. Our approach, based on sentence structure, provides a reasonably accurate data set to answer our research questions. Our strategy also allows us to identify a number of explicit hate speech targets (or communities), which directly complements (and benefits) the existing keyword search based semi-automated approaches.

A preliminary version of the present work has been published recently (Mondal, Silva, & Benevenuto, 2017). The present work improves and extends the findings in that earlier work. For instance, in section 6 we compare the volume of online and offline hate speech to better understand if online hate speech is in some way different than offline hate speech. Furthermore, in section 8 we compare the sensitivity of different categories of hate speech. We compute a crowdsourced metric called Anonymity Sensitivity score (AS-score) (Correa, Silva, Mondal, Benevenuto, & Gummadi, 2015) and show that content sensitivity varies significantly across categories. In our new analysis we note that "criminal" hate speeches, which corresponds to offline hate speech categories recognized by governments, are more likely to have higher AS-scores.

3. Datasets

Now we briefly describe our methodology to gather data from two popular online social media sites: Whisper and Twitter.

3.1. Collecting data from Whisper

Whisper is a popular anonymous social media site, launched in March 2012 as a mobile application. Whisper users post short text anonymous messages called "whispers" in this platform. In other words, whispers do not contain any identifiable information. An initial username is randomly assigned to users by Whisper, but it is not persistent i.e., users can change their usernames at any point of time. In addition, multiple users may choose to use the same username. Within a short span of time Whisper has become a very popular anonymous social media with more than 2.5 billion page-views, higher than even some popular news websites like CNN (Gannes, 2013). Within 2013 Whisper reached more than 2 million users and 45% of these users post something every day (Griffith, 2013). Statistics published by Whisper mention that 70% of their users are women, 4% have age under 18 years, and most of the Whisper users belong to the age group 17-28.

Whisper represents a valuable venue for studying online hate speech. In fact, recent works (Correa et al., 2015; Wang et al., 2014) suggest that Whisper offers an interesting environment for the study of online hate speech. These efforts show that users present a disinhibition complex in Whisper due to the anonymity. Since in an anonymous environment, people are more likely to shed their hesitation and disclose more personal information in their communications (Joinson, 2001). This anonymous nature of whispers combined with its popularity make Whisper an ideal candidate for our study.

Whisper users can only post messages via mobile phones, however Whisper has a read only web interface. In order to collect data from Whisper we employ a similar methodology as (Wang et al., 2014). We gather our dataset for one year (from 6th

June, 2014 to 6th June 2015) via the "Latest" section of the Whisper website which shows a stream of publicly posted latest whispers. Each downloaded whisper contains the text of the whisper, location, timestamp, number of hearts (favorites), number of replies and username.

Overall, our dataset contains 48.97 million whispers. We note that the majority (93%) of whispers are written in English. For the next sections we focus only on whispers in English as our approach to identify hate speech is designed for the English language. Moreover, we found that, 65% of these posts have a location associated to them. These locations are represented with unique place IDs (assigned by Whisper). We used the Whisper system to find a mapping between all possible values of latitude longitude (provided by us) and these place IDs. Using this mapping we ascertain exact location of **27.55 million whispers**. This dataset of more than 27 million whispers constitutes our final Whisper dataset used in the next sections.

3.2. Collecting data from Twitter

Since we want to study general hate speech in the online world, along with Whisper we also collected and analyzed data from Twitter, as it is one of the most popular social media sites today with more than 300 million monthly active users. The main difference between Whisper and Twitter is that users post in Twitter non-anonymously. The posts in Twitter are called tweets, and each tweet is associated with a persistent user profile which contains identifiable information. We found that, in spite of the non-anonymity, there are recent evidences of hate speech in Twitter (Chaudhry, 2015) and decided that it is useful to include Twitter in our study for a more inclusive analysis.

We collected the 1% random sample of all publicly available Twitter data using the Twitter streaming API (team, 2017) for a period of 1 year – June 2014 to June 2015. In total we collected 1.6 billion tweets (posts in Twitter) during this period. Some of the tweets also contained fine grained location information like whispers. However, one limitation for this Twitter dataset is that this addition of location is not enabled by default in Twitter. Thus, only a comparatively small fraction (1.67%) of Tweets have location information. Due to this limitation, we refrain from reporting results from Twitter in our location-based analysis due to insufficient location information later in this paper. Just like Whisper, we also used only English tweets, resulting in a dataset containing **512 million tweets** (32% of our crawled dataset). This dataset of more than 512 million tweets constitute our final Twitter dataset.

4. Measuring Hate Speech

Before presenting our approach to measure online hate speech, first we need to clarify what we mean by hate speech or hateful messages in this work. We note that, hate speech lies in a complex nexus with freedom of expression, group rights, as well as concepts of dignity, liberty, and equality (Gagliardone et al., 2015). For this reason, any objective definition (i.e. that can be easily implemented in a computer program) can be contested. In this work, we define hate speech as *an offensive post, motivated, in whole or in a part, by the writer's bias against an aspect of a group of people.*

Under our definition, all online hate speech might not necessarily be criminal offenses, but they can still harm people. The offended aspects can encompass offline

hate crimes[2], based on race, religion, disability, sexual orientation, ethnicity, or gender. However, they might also include behavioral and physical aspects that are not necessarily crimes. We do not attempt to separate organized hate speech from a rant as it is hard to infer individuals' intentions and the extent to which a message will harm an individual.

4.1. Using sentence structure to detect hate speech

Most existing efforts to measure hate speech require knowing the hate key words or hate targets apriori (Kwok & Wang, 2013). Differently, we propose a simple yet very effective method for identifying hate speech in social media posts which is in agreement with our definition of hate speech and which properly allows us to answer our research questions. Our key idea is the following: If some user posts about their hateful emotions in a post, e.g. "I really hate black people", then there is little ambiguity that it is a hate speech. In other words, we can leverage the sentence structure to detect hate speeches with high precision very effectively. Although our strategy do not identify all the existing hate speech in social media, however it still provides us a good and diverse set of hate speeches to perform analysis presented in this study.

Our expression to find hate speech: Based on our key idea, we construct the following basic expression (i.e., a sentence template) to search in social media posts:

$$I < intensity >< userintent >< hatetarget >$$

The components of this expression are explained next. The subject "I" means that the social media post matching this expression is talking about the user's (i.e., post writer's) personal emotions. The verb, embodied by the <user intent> component specifies what the user's intent is, or in other word how he feels. Since we are interested in finding hate in social media posts, we set the <user intent> component as "hate" or one of the synonyms of hate collected from an online dictionary[3]. We enumerate our list of synonyms of hate in the appendix. Some users might try to amplify their emotions expressed in their intent by using qualifiers (e.g., adverbs), which is captured by the <intensity> component. Note that a user might decide to not amplify their emotions and this component might be blank. Further the intensity might be negative which might disqualify the expression as a hate speech, for e.g., "I don't hate X". To tackle this, we manually inspect the intent expressions found using our dataset and remove the negative ones. We list expressions and words used as the <intensity> component in appendix as well. The final part of the expression is related to the hate targets, i.e., who is on the receiving end of hate.

Table 1 shows the top ten hate expressions formed due to the <intensity> component in conjunction with synonyms of hate. Although the simple expression "I hate" accounts for the majority of the matches, we note that the use of intensifiers was responsible for 29.5% of the matches in Twitter and for 33.6% in Whisper.

Determining hate targets: A simply strategy that searches for the sentence structure I <intensity> <user intent> <any word> results in a number of posts that do not actually contain hate speech against people, i.e. "I really hate owing people fa-

[2] `https://www.fbi.gov/about-us/investigate/civilrights/hate_crimes`

[3] http://www.thesaurus.com/browse/hate/verb

Twitter	% posts	Whisper	% posts
I hate	70.5	I hate	66.4
I can't stand	7.7	I don't like	9.1
I don't like	7.2	I can't stand	7.4
I really hate	4.9	I really hate	3.1
I fucking hate	1.8	I fucking hate	3.0
I'm sick of	0.8	I'm sick of	1.4
I cannot stand	0.7	I'm so sick of	1.0
I fuckin hate	0.6	I just hate	0.9
I just hate	0.6	I really don't like	0.8
I'm so sick of	0.6	I secretly hate	0.7

Table 1. Top ten hate intent in Twitter and Whisper.

vors", which is not in agreement with our definition of online hate speech. Thus, to focus on finding hate against groups of people, we additionally design two templates for filtering correct hate target tokens.

(1) Our first template for our <hate target> token is simply *"<one word> people"*. Thus, hate targets like "black people" or "mexican people" will match this template. This template for <hate target> captures the scenario when hate is directed towards a group of people. However, we observe that this template gives some false positives like "I hate following people". Thus, to reduce false positives we create a list of exclusion words for this particular hate target template. They include words like following, all, any or watching. The full list of such exclusion words is in the appendix.

(2) Naturally, not all hate targets might not contain the term "people". To account for this general nature of hate speech we take the help of Hatebase [4]. It is world's largest online crowdsourced repository of structured, multilingual, usage-based hate words. So we crawled Hatebase on September 12, 2015 to create a comprehensive list of hate targets. There are 1,078 hate words in Hatebase spanning 8 categories: archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation. However, each word in Hatebase is associated with an offensivity score. The score varies from 0 (not offensive) to 100 (most offensive). We take only the hate words from Hatebase with offensivity greater than 50[5], and use those words as template for <hate target> tokens in our sentence pattern. Note that, the usage of words from hatebase in this second template is inspired by earlier work which leveraged particular hate keywords for finding hate speech. However, unlike prior work, we use these keywords as part of our predefined sentence structure, explicitly putting these hate keywords into the context of hate speech while building our template.

Overall, our strategy identified **20,305 tweets** and **7,604 whispers** containing hate speech. We present the top hate targets (by% occurrence in posts) from Twitter and Whisper that we found using our methodology in Table 2. It shows racist hate words like "Black people", "White people" or "Nigga" are the most significant hate targets. We further checked how many of these hate messages are detected by our two different templates for hate target. Overall, the template with "people" finds more

[4]http://www.hatebase.org/

[5]There are 116 such hate words in Hatebase

Twitter		*Whisper*	
Hate target	**% posts**	**Hate target**	**% posts**
Nigga	31.11	Black people	10.10
White people	9.76	Fake people	9.77
Fake people	5.07	Fat people	8.46
Black people	4.91	Stupid people	7.84
Stupid people	2.62	Gay people	7.06
Rude people	2.60	White people	5.62
Negative people	2.53	Racist people	3.35
Ignorant people	2.13	Ignorant people	3.10
Nigger	1.84	Rude people	2.45
Ungrateful people	1.80	Old people	2.18

Table 2. Top ten targets of hate in Twitter and Whisper.

hate speech than using the words from Hatebase, accounting for 65% of the Twitter dataset and 99% of the Whisper dataset. One possible explanation for this difference is that Whisper operators might already filtering out some of the offensive words from Hatebase.

Limitation of our detection methodology: We acknowledge that our methodology aims for high precision while collecting hate speech and thus misses hate speech which does not conform to our sentence structure. However, we actually aimed to identify a diverse set of posts (not only race or gender based) which are truly spewing hate for further analysis, so we found our method acceptable. We also allowed a bit manual intervention to increase the precision (e.g., exclusion keywords). Moreover, our work may suffer from the biases that any work that rely on gathering online social media data currently suffers (Morstatter, Pfeffer, & Liu, 2014).

4.2. Evaluating our hate speech detection method

Next, we evaluate the accuracy of hate speech detection for our approach. Specifically, we wanted to ascertain if our detected posts can be labeled as hate speech by human judgment. Since human labeling is resource-consuming we decided upon labeling a subset of our detected posts. To that end we randomly sampled 50 posts from each of Twitter and Whisper from the set of posts which matched our language structure based expression. Finally, we end up with total of 100 posts (0.35% of all detected posts) which matched our language structure based expression.Then one of the authors manually verified whether these 100 posts can be really classified as hate speech by human judgment. We found that that 100% of both the whispers and tweets can be classified as hate speech, where the poster expressed their hate against somebody.

It is important to highlight that our methodology was not designed to capture *all* of the hate speech that in social media. In fact, detecting online hate speech is still an open research problem. Our approach aimed at building a high precision dataset that allowed us to simply answer our research questions. We plan to release these datasets by the time of the publication of this work.

4.3. Categorizing hate targets

For better understanding of the hate targets we manually categorize them in hate categories. For example, the term “black” should be categorized as race and “gay”

Categories	Example of hate targets
Race	nigga, nigger, black people, white people
Behavior	insecure people, slow people, sensitive people
Physical	obese people , short people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

Table 3. Hate categories with example of hate targets.

as sexual orientation. In order to decide the hate categories we take inspiration from the hate categories of Hatebase (mentioned earlier). We also consider categories reported by FBI for hate crimes. We end up with nine hate categories. We also add an "other" category for any non-classified hate targets. The final hate categories and some examples of hate targets for each category is shown in Table 3.

Since manual classification of hate targets into categories are resource consuming, we aim to categorize only the top hate targets that cover most of the hate speech in our data. Twitter and the Whisper datasets contain 264 and 242 unique hate targets respectively, and there is high overlap between the hate targets from Twitter and Whisper. We manually label the most popular 178 hate targets into categories, which accounts to more than 97% for both Twitter and Whisper hate speeches. We took a brief look at a random sample of the niche hate targets outside these popular hate targets (e.g., "brown people", "insensitive people", "neger") and they appear to contain less-used hate words which still belong to the same hate categories that we identified. However, due to the under-representation of these hate targets in our dataset it is hard to use them in statistically valid analysis and we instead focus on the popular hate targets. We will explore these hate categories and associated hate speech further in the next section.

5. Types of Online Hate Speech

Twitter		***Whisper***	
Categories	**% posts**	**Categories**	**% posts**
Race	48.73	Behavior	35.81
Behavior	37.05	Race	19.27
Physical	3.38	Physical	14.06
Sexual orientation	1.86	Sexual orientation	9.32
Class	1.08	Class	3.63
Ethnicity	0.57	Ethnicity	1.96
Gender	0.56	Religion	1.89
Disability	0.19	Gender	0.82
Religion	0.07	Disability	0.41
Other	6.50	Other	12.84

Table 4. The hate categories observed in hate speech from Twitter and Whisper.

We start with observing which categories of hate are most prevalent in our experimental platforms – Twitter and Whisper. The results are shown in Table 4. The hate categories are sorted by the number of hate speech in these categories (except for the non-classified hate targets, which we put in the other category). We made two interesting observations from this table. First, for both Twitter and Whisper the top 3 hate categories are the same – Race, behavior, and physical. However, in Twitter these categories cover 89% of the tweets, whereas in Whisper they cover only 69% of all the whispers related to hate. One potential explanation for this difference may be that, Whisper already filters very aggressive hate words, like those from the Hatabase. We also note that, for these categories in both Twitter and Whisper, there is also hate speech as a response to hate, e.g., "I hate racist people". However, such types of hate are not expressed in a high number of posts, and hate with negative connotation is much more common.

Secondly, we observe that out of the top 3 hate categories for both Twitter and Whisper, the categories "behavior" and "physical aspects" are more about *soft* hate targets, like fat people or stupid people. This observation suggests that perhaps many of the online hate speech are targeted towards groups of people, that are not generally captured by the documented offline hate speech (which considers hate speech based on race, nationality or religion). To dig further into this issue, next we contrast the difference between online and offline hate.

6. Online and offline hate speech

To compare online hate with offline hate we use a database from FBI about the hate crimes all over USA for the year 2013[6] and 2014[7]. The database provided us hate categories and the number of reported hate crime incidents in each category. Note that FBI does not include physical and behavioral-related hate speeches, even though they can be as harmful as the ones from other categories they are not considered crime.

Category	**% FBI$_{2013}$**	**% FBI$_{2014}$**	**% tweets**	**% whispers**
Race	49.28	48.30	48.73	19.27
Sexual orientation	20.21	18.68	1.86	9.32
Religion	16.92	17.06	0.07	1.89
Ethnicity	11.36	12.29	0.57	1.96
Disability	1.37	1.44	0.19	0.41
Gender	0.87	2.23	0.56	0.82
Behavior	-	-	37.05	35.81
Class	-	-	1.08	3.63
Physical	-	-	3.38	14.06

Table 5. Hate speech in each category in two different social media in comparison with hate crimes reported by FBI.

FBI reported total 5,928 and 5,479 hate crime incidents in 2013 and 2014 respectively. Since the raw numbers of FBI-reported hate crimes are different from the hate posts that we identified, in this section we decide to compare the percentages (and not raw number) of these hate crimes that fall in each category in the (offline) data from FBI and our (online) hate speech data. The result is shown in Table 5. Note that the

[6] https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-2013-hate-crime-statistics
[7] https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-2014-hate-crime-statistics

top hate targets that are considered crimes, both for twitter and Whisper, correspond to the most frequent forms of hate crimes in USA. Surprisingly, the proportion of crimes based on sexual orientation or religion is far higher than the perceived online hate speech, both in Twitter and Whisper.

Dataset	**FBI_{2013}**	**FBI_{2014}**	**Twitter**	**Whisper**
FBI_{2013}	1.0000	0.9988	0.9035	0.9525
FBI_{2014}	0.9988	1.0000	0.9115	0.9451
Twitter	0.9035	0.9115	1.0000	0.9109
Whisper	0.9525	0.9451	0.9109	1.0000

Table 6. Confusion matrix showing pairwise Pearson product-moment correlation coefficient for our four datasets. The coefficient is calculated between the percentage of hate across the hate categories common in both online and offline world.

We try to correlate the percentage of hate for the hate categories common in both online and offline world. We use Pearson product-moment correlation coefficient for the analysis. The coefficient varies from -1 to +1, where a value close to +1 signified very high similarity. Table 6 shows the confusion matrix from our analysis (all p values are <0.05, indicating that there is a low probability that this correlation occurred due to random chances). Note that, the Pearson product-moment correlation coefficients among the categories for Twitter, Whisper, and FBI data are all higher than 0.9, indicating that these categories are all very highly correlated. Thus, overall, the hate speech perceived online is highly correlated with hate crimes in USA, indicating that our detected explicit hate expressions in online world is an indicator of situation in the offline world.

We also note that hate speech based on physical or behavioral features, which are not considered crimes in most jurisdictions, are very common in social media. This suggests that detecting hate speech in online world can be helpful specially to identify new forms of hate that are not easily perceived in the offline world. For example, hate against fat people or poor people may not be considered a crime in many places, but it can be frequent and harmful. Its detection in certain regions is the first step for any sort of intervention. The implication of our observation is: finding and detecting hate speech in online world might require different mechanisms than offline world. Next, we will focus on the effect on identities on the hating behavior in online world.

7. Anonymity and hate speech

Early social psychology research found a number of evidences that the feeling of anonymity strongly influences one's behavior. Particularly, people tend to be more aggressive in situations in which they feel they are anonymous (Zimbardo, 1969). Thus, in this section we aim to investigate the effects of anonymity on online hate speech. Specifically, we investigate the amount of users that unveil personal names as part of their identities across different categories of online hate speech. Our hypothesis is that more sensitive categories of hate speech, like those associated to offline hate crimes, tend to be posted by a large fraction of users that do not use a personal name as part of their Twitter profiles (we exclude Whisper from this section as it is already anonymous).

Detecting personal names: Our approach consists of using a lexicon lookup approach to detect if the name provided by the Twitter account can be considered a

Category	% Tweets posted anonymously (without personal names)
Random tweets	40%
Race	55%
Sexual Orientation	54%
Physical	49%
Behavior	46%
Other	46%

Table 7. Percentage of tweets posted through accounts without common personal names (i.e., anonymously) across categories of hate speech.

common personal name. Since Facebook, another large social media site has a 'Real Name' policy, we exploit the names provided by Facebook users to build our personal name database. We use a Facebook dataset[8] containing 4.3 million unique first names and 5.3 million unique last names as lexicon. In order to reduce noise, we removed first/last names that appear lesser than five times in the Facebook dataset. We call a name provided by a Twitter account as personal if the name matches two or more tokens in our lexicon. In other words, we posit that a personal name must have at least two tokens as names used in the real world (equivalent to Facebook's first and last name policy). We ensure a clean matching by eliminating tokens from Twitter account names that contain stopwords or those that belong to WordNet (Miller, 1995), a database that contains common English words. We evaluate this system independently and discover the accuracy (F1 score) to be 78% for detecting names of real people. Using this method, we identify the fraction of hate speech that is posted by *not* using a personal name, i.e., anonymously.

Correlation between anonymity and hate speech: Table 7 shows the percentage of tweets posted using anonymous accounts across top hate speech categories. We also consider a set of 1,000 tweets, randomly sampled from all tweets posted in 2014-15, which we use as baseline for comparison. We make two observations: Firstly, the percentage of users posting hate speech not using personal names i.e, anonymously is more than a random set of tweets. Secondly, more hate speech concerning race or sexual orientation is posted anonymous compare to when users post softer categories of hate, i.e., Behavior and Physical. Our findings suggest that weak forms of identity (i.e., anonymity) fuels more hate in online media systems and the use of anonymity varies with the type of hate speech. Next, we will investigate if along with the usage of anonymous identities the public perception of sensitivity also varies with different types of hate speech.

8. Hate speech sensitivity

In this section, we investigate how people perceive hate speeches; in other words, according to social norms, what hate categories are more likely to be considered sensitive. In order to measure this perception, we leverage the Anonymity Sensitivity Score (explained below) proposed by Correa at al. (Correa et al., 2015).

[8]https://blog.skullsecurity.org/2010/return-of-the-facebook-snatchers

8.1. Measuring content sensitivity via AS Score

Generally online social media sites treat anonymity as a binary concept and systems are designed to cater to either anonymous or non-anonymous content. However, research in behavioral psychology suggests that anonymity is *subjective* in nature (Pinsonneault & Heppel, 1997). Prior work (Correa et al., 2015) have shown that the anonymous sensitivity transcends binary notions and has different levels. They proposed a crowdsourcing based methodology to measure sensitivity of content based on public perception. In order to measure the sensitivity of content we took a similar approach. We setup the following experiment in the crowdsourcing platform CrowdFlower: we randomly pick ten posts from each of the ten hate speech categories from section 5, and ten non-hate speech posts (for baseline), from each social media (Twitter and Whisper). We ask ten CrowdFlower workers to annotate each of these $10 \times (10 + 1) \times 2 = 220$ posts as anonymous or non-anonymous. We do not reveal the origin of the tweets or whispers to CrowdFlower workers and each worker can assign at most 50 questions (CrowdFlower restriction).

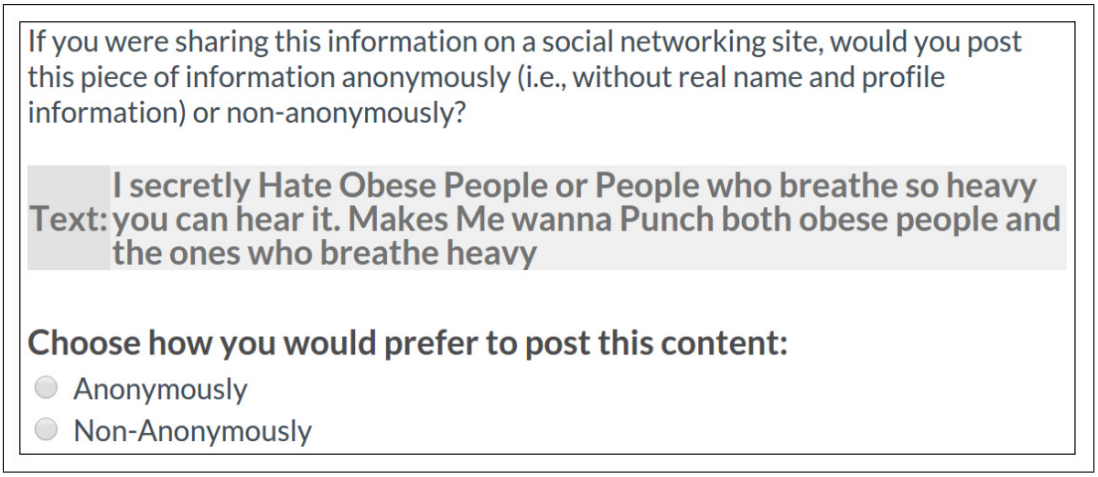

Figure 1. Screenshot of a survey question inside CrowdFlower platform.

Note that we used CrowdFlower in our experiment whereas the earlier study (Correa et al., 2015) used Amazon Mechanical Turk (AMT). We chose CrowdFlower mainly due to the inherent data quality assurance and noise elimination mechanisms of CrowdFlower. Furthermore, related research has shown that the performance of AMT and CrowdFlower workers are quite comparable (Finin et al., 2010).

Representativeness of annotators: Since we aim to measure the *public* perception of hate speech sensitivity, representativeness of CrowdFlower workers is one potential concern. CrowdFlower provides the country of the annotators and we find that 14.75% of the annotators are from India, the other 85.25% are spread into 41 countries, which shows that our workers are from multiple countries and hence multiple cultures.

Computing AS Score: In our annotated data, each text (tweet or whisper) is marked by 10 CrowdFlower workers as anonymous or non-anonymous – see figure 1. As proposed in (Correa et al., 2015), the fraction of CrowdFlower workers who annotate each text as anonymous is a probabilistic estimate of the fraction of users that would consider the text as anonymous. This probabilistic estimate is the Anonymity Sensitivity Score (or *AS Score*) for that particular text. Formally, the *AS Score* for a given piece

of text is the probability that users would consider this text as anonymous. Formally if a CrowdFlower worker w_j annotating a text t_i then,

$$AS\ Score_{t_i}^{w_j} = \begin{cases} 0, & \text{when } w_j \text{ marks } t_i \text{ non-anonymous} \\ 1, & \text{when } w_j \text{ marks } t_i \text{ anonymous} \end{cases} \quad (1)$$

AS Score	**Example**
0.0	Feels good to have my family back :)
0.1	I really can't stand stupid people.
0.2	I'm really sick of immature people.
0.3	I hate racist people. I never get their sense about races.
0.4	I can not stand sexist people
0.5	I don't like Bipolar people they annoy me!
0.6	I'm not sick, I'm just allergic to cunts like you.
0.7	Literally I cannot right now. I hate gay people :—
0.8	I hate ugly people
0.9	I hate religious people. Keep YOUR beliefs to YOURselves, douche bags
1.0	I fucking hate poor people

Table 8. Examples posts from our CrowdFlower experiment with different *AS Scores*.

Table 8 shows examples of messages from our CrowdFlower experiment for different *AS Scores*. We can see that the message *"I don't like Bipolar people they annoy me!"* was labeled as anonymous by 5 CrowdFlower workers, and as non-anonymous by the remaining 5 workers, giving an *AS Score* of 0.5 to this text. We note that pieces of content with higher *AS Scores*, i.e. which higher number of CrowdFlower workers annotate as anonymity sensitive, tend to be more controversial and intuitively require more anonymity. Behavioral studies in psychology have also shown that anonymity leads people to reveal sensitive content (Suler, 2004).

Category	**Fleiss Kappa**
Race	0.218
Ethnicity	0.214
Sexual orientation	0.213
Class	0.204
Gender	0.188
Disability	0.166
Religion	0.158
Other	0.129
Physical	0.122
Behavior	0.109

Table 9. Fleiss Kappa score for each hate category. Higher values indicate higher agreement.

User agreement: Since, we computed AS score with multiple annotators, we also checked the inter-annotator agreement between the CrowdFlower annotators using Fleiss Kappa score. The result is in Table 9 which shows that the score varies from 0.109 to 0.218, indicating slight to fair agreement. However, the score also varies with categories; in fact the table shows that Fleiss kappa scores for offline hate-related categories is higher (indicating more agreement).

8.2. Analyzing the AS-scores

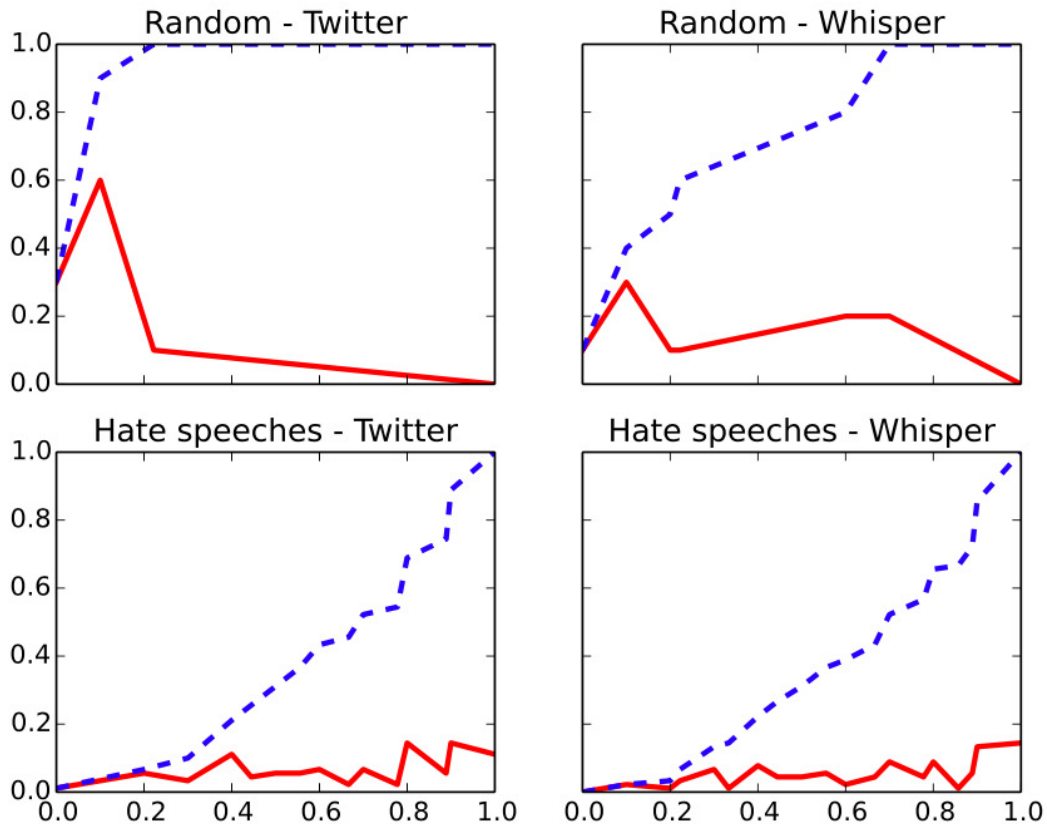

Figure 2. Cumulative Distribution Function (CDF) and Probability Distribution Function (PDF) of the AS Scores of hate speeches on Twitter and Whisper. CDF is in blue dotted lines and PDF is in red solid lines. We show separate figures for random posts and hate speeches, each for Twitter and Whisper.

In this section, we examine the differences in anonymity sensitivity of content across hate speech categories. Our experiment helped us quantify the anonymity sensitivity of each post, and from chapter 5 we have the hate speech category it is related to. Now, we combine results from both the aforementioned CrowdFlower experiment and the category assignment. Figure 3 shows the distribution of this score for all categories via a violin-plot.

In order to compare anonymity sensitivity of hate speech content, we inspect *AS Score* probability distributions. Figure 2 shows the cumulative and probability distributions of *AS Score* for anonymous and non-anonymous media (for hate speech posts and similar number of random posts). Earlier work (Correa et al., 2015) had already shown that AS Score distribution in tweets are more concentrated in lower values, while AS Score from whispers is distributed over the whole spectrum. We re-checked this observation for hate speech posts as well as random posts using the Mann-Whittney-U test (Mann & Whitney, 1947). This test helped us to ascertain if there is a statistically significant difference between the AS-scores of Twitter and Whisper posts. We found that, for AS-scores from random posts in Whisper and Twitter Mann-Whittney-U test shows statistically significant difference ($p < 0.05$). However for hate speech posts the difference is not statistically significant between Twitter and Whisper ($p = 0.35$).

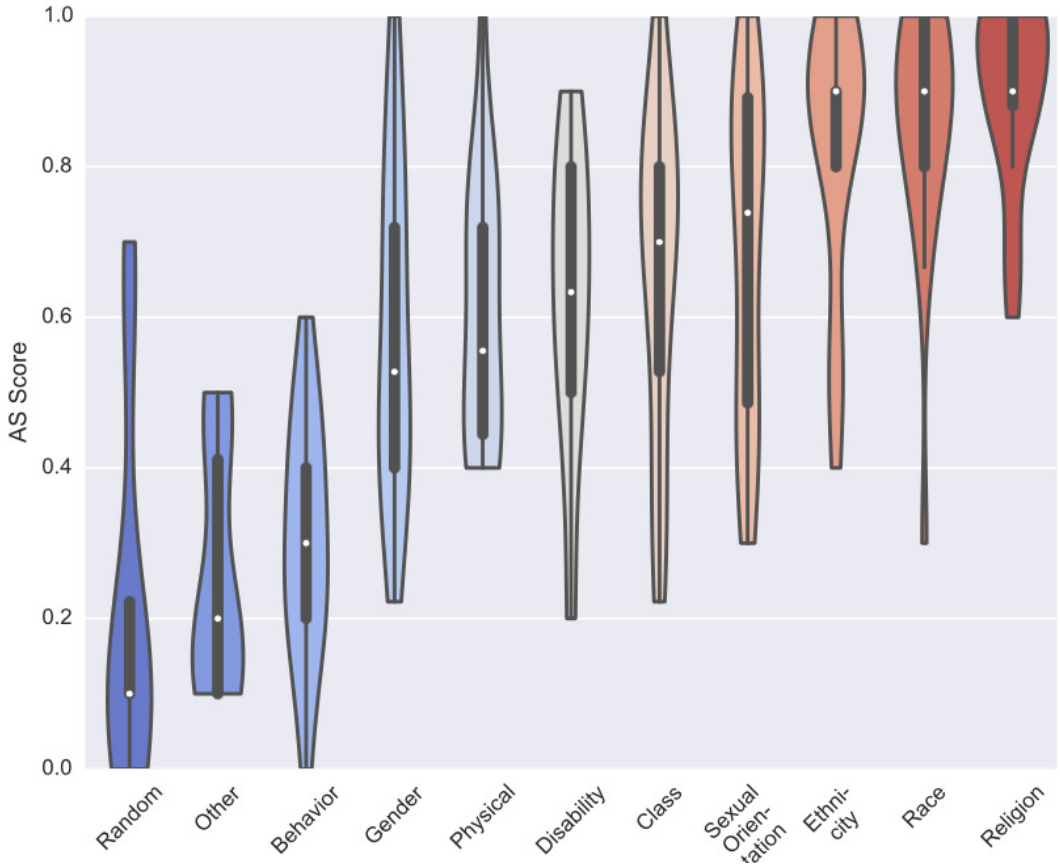

Figure 3. Violin-plot distribution of AS Scores of hate speeches for each category.

Furthermore, Figure 3 shows that content sensitivity varies significantly across categories and different categories contain content with different levels of anonymity sensitivity, and hate speeches from hate categories like Religion or Sexual orientation are more likely to "desire" anonymity than say Behavior based hate speech.

9. The Geography of Hate Speech

Next, we explore the correlation of geography and hate speech. In this section we very briefly present our key findings. We encourage interested readers to check out our earlier work (Mondal et al., 2017) for detailed results. For this analysis, we focus solely on whisper data as the amount of Twitter with geographic information is not significant. We start by comparing hate speech in different countries.

9.1. Hate speech across countries

Recall that our approach to measure hate speech only considered posts in English. Thus, unsurprisingly, US, Canada, and UK are top three countries in our Whisper dataset; they are responsible for 80%, 7%, and 5% of total hate speech in Whisper respectively. We focus our inter nation comparative analysis on these three countries.

We note that, hate towards people based on behavior is the most dominant hate

category in all three countries. Hate based on physical aspects of individuals also appear in the top 3 positions for the three countries. Interestingly hate based on race in US is higher (20%) in comparison with Canada (13%) and UK (13%). On the other hand, hate speech related to sexual orientation in UK (14%) is almost two times higher than in US (8%) and Canada (7%).

Our observation suggests that, monitoring hate in online social media can help authorities to strategically detect and prevent different types of hate speech in different countries.

9.2. Hate speech within a country

We also analyzed the hatespeech within a specific country, namely US. We found that users from southern US states post more hate speech based on race and sexual orientation. Whereas users from west part of US post more hate speech based on physical features. Furthermore, we found that hate speech from categories that are not related to crimes such as behavior, and physical features are more uniformly distributed across all the states. However, hate speech on crime related topics, such race, sexual orientation and class is more skewed across states. Our inter country analysis suggests that local actions and interventions for specific types of hate speech in specific locations (even within a country) is necessary.

10. The context of hate speech

Finally, we investigate other sentences that appear together with hate speech. Our goal is to better understand the sentences associated with hate speech. We noted that 65% of the messages in our whisper dataset and 80% in our Twitter dataset contains extra (part of) sentences following a detected hate pattern (i.e, the part that matched our hate expression). We call these parts of sentences the context of hate speech.

We filter out the context by grouping all of our detected hate speech, and removing the parts of sentences that matched our hate expression (along with the hate target) for each hate speech. The resulting sentences give us context. For example, in the sentence "I hate racist people, their point of view is medieval", we extract the (partial) sentence "their point of view is medieval" as context.

Figure 4 shows a WordTree [9] visualization built from our contexts for the root *I* and *they* (i.e. "I hate fat people, they..."), using as input our aforementioned analysis. The visualization shows phrases that branch off from this root expression (hate speech) across all hate speeches of our dataset. A larger font size means that the word occurs more often. We can note that the words *I, I, they, who, and, but* are quite popular. Among them, 'I' and 'they' emphasize personal nature of hate whereas 'but' soften the hate the users express in hate speech. Due to space limitations, next we chose the suffixes *I* and *they* to further analyze.

Figure 5 zooms on the WordTrees for these two suffixes. We can make two important observations from them. Firstly, we noted that part of these sentences simply attempt to intensify the hate expressed against a group of people. Second, and more interestingly, these phrases provide evidence that many users tend to justify their hate against others, especially in Twitter. We believe that the analysis of these particular sentences might be a valuable source of information to better understanding the root

[9] `https://www.jasondavies.com/wordtree/`

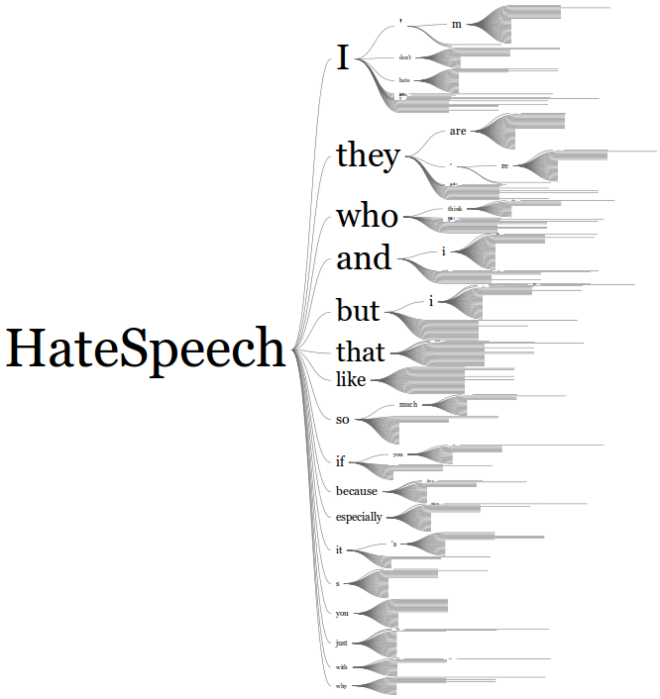

Figure 4. Word tree for our contexts of hate speeches.

causes of hate in some regions and places.

11. Conclusion

The fight against online hate speech is beginning to reach a number of concerned parties, ranging from governments, private companies and Internet Service Providers to a growing number of active organizations and affected individuals. Our measurement study on online hate speech provides an overview of how this important problem of spewing hate manifests online. Our effort consists of studying generic online hate speech according to four dimensions: the main targets of online hate speech, correlation with anonymity, the geography of hate speech and the context of hate speech.

Among our main findings, we highlight the importance of having real names associated with posts to reduce the hate in the online world. More important, we show that the hate speech in the online world reflects the hate in the offline world. This suggests that detecting and monitoring hate speech in the online world can be one step forward to the solution for detection and early prevention of hate speech and hate related crimes in the offline world.

Acknowledgements

This research was supported in part by the Alexander von Humboldt Foundation, CNPq, CAPES, and Fapemig.

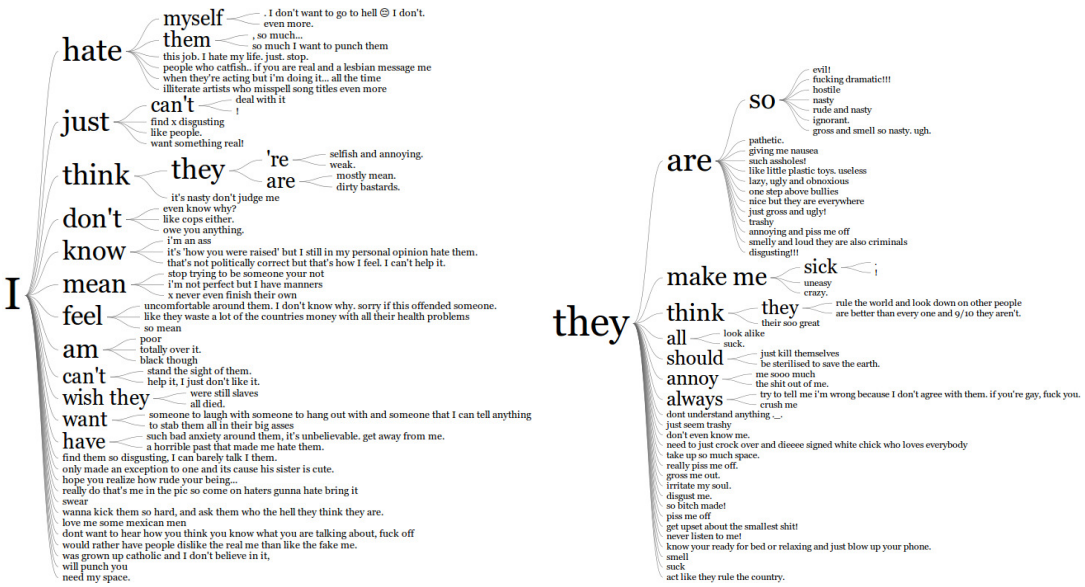

Figure 5. Word tree for context of hate speech starting with the words *I* and *They*.

Appendix

List of synonyms of "hate": *do not like, abhor, despise, detest, loathe, scorn, shun, abominate, anathematize, contemn, curse, deprecate, deride, disapprove, disdain, disfavor, disparage, execrate, nauseate, spurn, am allergic to, am disgusted with, am hostile to, am loath, am reluctant, am repelled by, am sick of, bear a grudge against, cannot stand, down on, feel malice to, have an aversion to, have enough of, have no use for, look down on, do not care for, object to, recoil from, shudder at, spit upon*

List of words used as <intensity> token: *absolute, absolutely, actually, already, also, always, bloody, completely, definitely, do, especially, extremely, f*cking, fckin, fkn, fr, freakin, freaking, fucken, fuckin, fucking, fuckn, generally, genuinely, honestly, honesty, jus, just, kinda, legitimately, literally, naturally, normally, now, officially, only, passively, personally, proper, really, realy, rlly, rly, secretly, seriously, simply, sincerely, so, sometimes, sorta, srsly, still, strongly, totally, truly, usually*

List of words to exclude from the first hate word pattern: *about, all, any, asking, disappointing, everyone, following, for, having, hearing, how, hurting, is, it, letting, liking, many, meeting, more, most, my, myself, on, other, seeing, sexting, some, telling, texting, that, the, them, these, this, those, watching, wen, what, when, when, whenever, why, with, you*

References

Agarwal, S., & Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *Proceedings of the 11th international conference on distributed computing and internet technology (icdcit'15).*

Apple Pulls Secret App in Brazil After Judge's Request. (2014, August). `http://mashable.com/2014/08/22/secret-app-brazil/`.

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2014). *Anti-social media.* DEMOS.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 31:1–31:22.

Chaudhry, I. (2015). #hashtagging hate: Using twitter to track racism online. *First Monday*, *20*(2).

Correa, D., Silva, L., Mondal, M., Benevenuto, F., & Gummadi, K. P. (2015). The many shades of anonymity: Characterizing anonymous social media content. In *Proceedings of the 9th international aaai conference on weblogs and social media (icwsm'15).*

Delgado, R., & Stefancic, J. (2004). *Understanding words that wound.* Westview Press.

Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). *Understanding harmful speech online* (Tech. Rep.). Berkman Klein Center for Internet & Society Research. (`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882824`)

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the naacl workshop on creating speech and text language data with amazon's mechanical turk.*

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* UNESCO.

Gannes, L. (2013, August). *On making our digital lives more real.* http://allthingsd.com/20130802/im-so-over-oversharing-on-making-our-digital-lives-more-real/.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

Greevy, E., & Smeaton, A. F. (2004). Classifying Racist Texts Using a Support Vector Machine. In *Proceedings of the 27th annual international acm sigir conference on research and development in information retrieval.*

Griffith, E. (2013, May). *With 2 million users, "secrets app" whisper launches on android.* http://pando.com/2013/05/16/with-2-million-users-secrets-app-whisper-launches-on-android/.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*(2), 177–192.

Kwan, G. C. E., & Skoric, M. M. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior*, *29*(1), 16–25.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the aaai conference on artificial intelligence (aaai'13).*

Mann, H., & Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Massaro, T. M. (1990). Equality and freedom of expression: The hate speech dilemma. *William and Mary Law review*, *32*(2), 211–265.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 25th acm conference on hypertext and social media (ht'17).*

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased?: Assessing the representativeness of twitter's streaming api. In *Proceedings of the 23rd international conference on world wide web.*

Pinsonneault, A., & Heppel, N. (1997). Anonymity in group support systems research: A new conceptualization, measure, and contingency framework. *Journal of Management Information Systems*, *14*(3), 89–108.

Reis, J., Benevenuto, F., de Melo, P. O. V., Prates, R., Kwak, H., & An, J. (2015). Breaking

the news: First impressions matter on online news. In *International conference on web and social media (icwsm)*.

Sanchez, H., & Kumar, S. (2011). Twitter bullying detection. *ser. NSDI*, *12*.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In *International conference on web and social media (icwsm)*.

Stephens, M. (2013). *The geography of hate map.* http://users.humboldt.edu/mstephens/hate/hate_map.html.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321–326.

team, T. (2017). *The streaming apis.* `https://dev.twitter.com/streaming/overview`.

Ting, I.-H., Chi, H.-M., Wu, J.-S., & Wang, S.-L. (2013). An approach for hate groups detection in facebook. In *Proceedings of the 3rd international workshop on intelligent data analysis and management (iadm'13)*.

Waldron, J. (2012). *The harm in hate speech.* Harvard University Press.

Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., & Zhao, B. Y. (2014). Whispers in the dark: Analyzing an anonymous social network. In *Proceedings of the 2014 conference on internet measurement conference (imc'14)*.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the 2nd workshop on language in social media (lsm'12)*.

Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. *Nebraska Symposium on Motivation*, *17*, 237–307.

Author Biographies

Mainack Mondal *is a postdoctoral researcher at the University of Chicago in Chicago, IL, USA. His research interests are in networked systems, with an emphasis on user privacy. Mondal has a PhD from Max Planck Institute for Software Systems, Germany. Contact him at mainack@uchicago.edu.*

Leandro Araújo Silva *has a MSc in computer science from Universidade Federal de Minas Gerais (UFMG). Currently Silva is a data scientists at Kunumi, working on solutions based on deep learning.*

Denzil Correa *is a Data Scientist at Bayer Pharma AG, Germany. Currently, he uses computational techniques to help discover new medicine and bring them faster to the market. Previously, he worked on computational methods to understand human behavior from online social data.*

Fabrício Benevenuto *is associate professor in the Computer Science Department of Universidade Federal de Minas Gerais (UFMG). He was elected an affiliated member of the Brazilian Academy of Science and also received a prestigious scholarship from Humboldt foundation through which he was visiting faculty at MPI-SWS (2017-2018). His research interests are in social computing and sentiment analysis related projects.*