A Measurement Study of Hate Speech in Social Media

Mainack Mondal
MPI-SWS
Germany
mainack@mpi-sws.org

Leandro Araújo Silva
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
leandro@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Social media platforms provide an inexpensive communication medium that allows anyone to quickly reach millions of users. Consequently, in these platforms anyone can publish content and anyone interested in the content can obtain it, representing a transformative revolution in our society. However, this same potential of social media systems brings together an important challenge—these systems provide space for discourses that are harmful to certain groups of people. This challenge manifests itself with a number of variations, including bullying, offensive content, and hate speech. Specifically, authorities of many countries today are rapidly recognizing hate speech as a serious problem, specially because it is hard to create barriers on the Internet to prevent the dissemination of hate across countries or minorities. In this paper, we provide the first of a kind systematic large scale measurement and analysis study 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 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.

CCS CONCEPTS

• **Information systems** → *Social networks*; • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

KEYWORDS

hate speech; anonymity; social media; Whisper; Twitter; pattern recognition

ACM Reference format:
Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of HT '17, Prague, Czech Republic, July 04-07, 2017,* 10 pages.
https://doi.org/http://dx.doi.org/10.1145/3078714.3078723

HT '17, July 04-07, 2017, Prague, Czech Republic
© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-4708-2/17/07…$15.00
https://doi.org/http://dx.doi.org/10.1145/3078714.3078723

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 platform provided by these systems allow users to express themselves, there is also a dark side of these systems. Particularly, these social media sites have become a fertile ground for inflamed discussions, that usually polarize 'us' against 'them', resulting in many cases of insulting and offensive language usage.

Another important aspect that favors such behavior is the level of anonymity that some social media platforms grant to users. As example, "Secret" was created, in part, to promote free and anonymous speech but became a mean for people to defame others while remaining 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 [25], Twitter [24], Instagram [15], and Facebook [17].

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 [1, 2, 12, 26, 29]. These efforts mostly focus on specific manifestations of hate, like racism [3]. While these efforts are quite important, they do not attempt to provide a big picture of the problem of hate speech in the current popular social media systems. Specifically providing a broad understanding about the root causes of online hate speech was not main focus of these prior works. Consequently, these prior works also refrain from suggesting broad techniques to deal with the generic offline hate underlying online hate speech.

In this paper, we take a first step towards better understanding online hate speech. Our effort consists of characterizing how hate speech is spread in common social media, focusing on understanding 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.

What role does anonymity play on hate speech? Is anonymity a feature that exacerbate 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 does hate speech vary across geography? Does hate speech targets vary across countries? And within states of a country, like

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

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 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 review existing work on hate speech along three dimensions.

2.1 Understanding hate speech

Hate speech has been an active research area in the sociology community [9]. Particularly, [19] claims that some forms of hate speech are far from being solved in our society, specially those against black people and women. Hate speech originating from such prejudices are quite abundant and authorities have created standard policies to counter it. However, there has been a multitude of undesirable social consequences of these very policies (e.g., incivility, tension, censorship, and reverse discrimination) due to the suppression of haters and the protection of hate targets. Over time, this tension has driven the evolution of standard policies to regulate hate speech.

A very recent study [10], supported by UNESCO, reviews the growing problem of online hate speech with the advent of internet from a legal and social standpoint. They pointed out that platforms like Facebook and Twitter have primarily adopted only a reactive approach to deal with hatespeech reported by their users, but they could do much more. More specifically, their study reports "*These platforms have access to a tremendous amount of data that can be correlated, analyzed, and combined with real life events that would allow more nuanced understanding of the dynamics characterizing hate speech online*". Our work is motivated by this vision.

Even before the popularity of social networks, the problem of racism and hate detection was already a research theme in computer science. Back in 2004, there has been efforts that attempt to identify hateful webpages, containing racism or extremism [13]. Nowadays, there has been a multitude of related problems under investigation in social media systems [5, 23]. However, these approaches does not give a data driven global view of hate speech in online media today, we aim to bridge this gap.

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. [3] reviews three different recent studies that aim to detect the presence of racism or offensive words on Twitter. They point out that, while simple text searches for hate words in tweets represent a good strategy to collect hate speech data, it also creates a problem: the context of the tweets is lost. For instance, the word "Crow" or "Squinty" is a racial slur in United Kingdom, but it can also be used in multiple different non-hate related contexts. Multiple researchers try to solve this problem using manual inspection or a mix of crowdsource labeling and machine learning techniques [1, 2, 12, 26, 29, 31]. Their basic framework consists of creation of a corpus which contain a set of known hate keywords. This corpus is then manually annotated to construct a training dataset which contains positive and negative hate posts. Finally, they learn from this training dataset to build automated systems (via machine learning approaches) for detecting 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. Aside from leveraging text based features researchers also explored other features like leveraging user history [8] or even community detection [6]; These techniques can be used in addition to the text based features. Although all these efforts offer advances in this field, it is safe to say that computational methods to detect hate speech currently 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.

2.3 Hate speech and anonymity

The problem of hate speech inspired a growing body of work in effectively detecting such speeches on various social media platforms. However so far these efforts focused on either non-anonymous social media platforms, like Twitter or Facebook [18, 29], or on radical forums and known hate groups [31]. However there is an interesting and unexplored middle ground in between—Anonymous social media like Whisper or Secret. These media sites are recently becoming quite popular within normal users. These platforms do not require any account or persistent identity to post on their sites. Recent efforts [7, 30] reviewed content posted on such forums in depth. They found that users post more sensitive content on such forums and a significant fraction of such posts are confessions about their personal lives. Existing efforts in sociology [22, 27] already pointed out that in the presence of anonymity, humans show a disinhibition complex. In other words, the posters might be much less inhibited and express their otherwise suppressed feeling or ideas on anonymous social media sites. Thus, intuitively, in the presence of anonymity one will expect to find the presence of hate speech from a diverse set of users who are not radicalized, but they have certain prejudices which otherwise they will not express in their posts. Based on this intuition we made an effort to investigate Whisper, an anonymous social media system, in our analysis. We hope to provide a more inclusive picture about hate speech in social media in that way.

In a preliminary short paper [25], we attempted to correlate hate crimes incident with hate speech in Whisper and Twitter. In this paper, we used the same methodology to gather data from Twitter and Whisper, but we provide a much wider and deeper understanding of hateful messages in these systems.

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 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 at the same time. 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 [11]. Within 2013 Whisper reached more than 2 million users and 45% of these users post something every day [14]. 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 [7, 30] 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 [16]. 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 [30]. 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—one of the most popular social media sites today.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 [3] 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 [28] 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 [10]. 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 individual intentions and the extent to which a message will harm an individual.

4.1 Using sentence structure for hate speech detection

Most existing efforts require knowing the hate key words or hate targets apriori [18] for detecting hate speech. Differently, we propose a simple yet very effective method for identifying hate speech

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

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 does not identify all the existing hate speech in social media (signifying possibly low recall), 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 these cases, 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.

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.

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 favors", which is not in agreement with our the 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.

First template: The 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, in order 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.

Second template: 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 (provided by hatebase). The score varies from 0 (not offensive) to 100 (most offensive). We take only the hate words from Hatebase with offensivity score greater than fifty[5], and use those words as template for <hate target> tokens in our sentence pattern. Experimenting with other thresholds of offensivity score is part of our future work.

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.

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,

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

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

[5] There are 116 such hate words in Hatebase

the template with "people" finds more 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 [4].

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 (i.e., have possibly low recall). 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 further (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 [21].

4.2 Evaluating our detection method

Next, we evaluate the accuracy of hate speech detection for our approach. To that end, we performed a simple experiment: We randomly sampled 50 posts from each of Twitter and Whisper 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.

4.3 Categorizing hate targets

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.

For better understanding of the hate targets, two of the authors manually categorize them in hate categories (one author first categorized the targets and another author independently reviewed the categories and hate targets to ensure correctness of the categorization). For example, the term "black" should be categorized as race and "gay" 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. Our Twitter and 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 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 percentage of hate speech in these categories (except for the non-classified hate targets, which we put in the other category). We made two observations from Table 4. Firstly, for both Twitter and Whisper the top three hate categories (by percentage of hate targets, not counting "other" category) 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. As mentioned earlier, one potential explanation for this difference may be that, Whisper already filters very aggressive hate words, like those from the hatabase [4]. 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 usage of "I 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).

6 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 [32]. 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 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[6] 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 [20], 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: In Table 5 we show the percentage of tweets posted using anonymous accounts across top hate speech categories. We also consider a set of random tweets, unrelated to hate speech, 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.

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

Table 5: Percentage of tweets posted through accounts without common personal names (i.e., anonymously) across categories of hate speech.

7 THE GEOGRAPHY OF HATE SPEECH

Next we explore the correlation of geography and hate speech. 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.

7.1 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.

What are the top hate categories across countries? Table 6 shows the ranked breakdown of hate speech posted by users from these countries across hate speech categories. We make a few interesting observations from this breakdown. 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 three 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%). We further checked the exact hate targets posted by users from these countries.

What are the top hate targets across countries? In Table 7, we notice country specific biases on the usage of hate targets, which helps to explain the observed differences in the hate categories across countries. In US, there is a clear bias towards hate against black people, as it is the most popular hate target in hate speeches from US, accounting alone for 11% of the hate speech. Hate speech against white people only ranked 6th in hate targets from US and accounts only for 5% of hate speech. This discrepancy tends to be smaller in Canada and UK, where both of hate targets, Black and White people appear around the ranks 4th to 6th and account for about 4% to 6% of the hate speech. Interestingly, we also observed hate against specific groups of haters in these countries. For instance, racist people are one of the top 10 hate targets in all of these three countries. Similar type of bias can be noted for sexual orientation. We note that, in UK, hate against gays appears in second place, whereas hate against homophobic people is ranked 11th in the list of hate targets of UK (nor shown in table). Furthermore, in hate based on physical aspects, Fat people appear with high frequency, being the most popular hate target for Canada and UK and on the behavior category, hate against fake, stupid, selfish, and rude people are quite common across countries.

These results not only highlight the different forms in which hate speech manifests itself in different countries, but it also identifies country specific biases in the hate speech. Our observation suggests

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

Rank	***US***		***Canada***		***UK***	
	Hate category	%	**Hate category**	%	**Hate category**	%
1	Behavior	36	Behavior	39	Behavior	26
2	Race	20	Physical	17	Physical	21
3	Physical	14	Other	15	Sexual orientation	14
4	Other	13	Race	13	Race	13
5	Sexual orientation	8	Sexual orientation	7	Other	12
6	Class	4	Class	3	Class	4
7	Religion	1	Ethnicity	3	Religion	4
8	Ethnicity	1	Religion	2	Ethnicity	4
9	Gender	1	Gender	1	Gender	1
10	Disability	0	Disability	1	Disability	0

Table 6: Top hate categories for the countries with most posts: US, Canada and United Kingdom.

Rank	***US***		***Canada***		***UK***	
	Hate target	%	**Hate target**	%	**Hate target**	%
1	Black people	11	Fat people	11	Fat people	17
2	Fake people	10	Stupid people	9	Gay people	10
3	Stupid people	8	Fake people	6	Stupid people	7
4	Fat people	8	Black people	6	Black people	5
5	Gay people	7	Gay people	5	White people	4
6	White people	5	White people	4	Rude people	4
7	Ignorant people	4	Rude people	4	Fake people	4
8	Racist people	4	Racist people	3	Ignorant people	4
9	Old people	2	Selfish people	3	Religious people	3
10	Rude people	2	Old people	3	Racist people	3

Table 7: Top hate targets in US, Canada and United Kingdom.

that, monitoring hate in online social media can help authorities to strategically detect and prevent different types of hate speech in different countries. Next, we analyze hate speech within US.

7.2 Hate speech within a country

We start with measuring the volume of hate speech in each US state. As whisper is a young social media platform with uneven adaptation within US, the raw volume of hate speech from a state might be biased by simply the total volume of posts uploaded from that state. Thus we measure the relative amount of hate speech contributed by each state by dividing the state level actual percentage of hate speech with a state level expected percentage of hate speech. The state level expected percentage of hate speech for a state is simply the state level percentage of Whisper messages from that state in our Whole Whisper dataset posted from US.

Comparison of volumes of hate speech posted by US states: Interestingly, Figure 1 shows that west and northeast state users tend to have less relative amount of online hate speech. To further explore this trend we divided US into regions and focus on hate speech at region level[7]. We calculate the relative amount of hate speech for each region (similar to that of each state). Notably users from west and northeast states tend to have less relative amount of hate speech (relative amount 0.91 and 0.93 respectively), compare to the users from midwest and southern states (relative amount 1.07 and 1.05 respectively). To better understand these differences, we take three top categories of hate speech – race, physical, and sexual orientation. Then, we check what normalized volume of hate speech from these categories are posted from different US regions. Figure 2 shows the heat map of the volume of hate speech three hate categories across U.S. regions. The presented values are normalized by total volume of hate speech in each respective region. We can note that users from southern states post more hate speech based on race and sexual orientation. Whereas users from west post more hate speech based on physical features.

How concentrated are hate speech from different hate categories across US states? Finally, we focus on measuring the extent to which certain kinds of hate appear more or less spread across different U.S states. To do that, we define *hate entropy*, which is effectively the information entropy of the distribution of hate speech against a target category over the different regions. Hence, higher values of hate entropy denote target categories whose speeches are spread more uniformly across several US states, while lower entropy values signal hate speech more concentrated in a few regions.

Table 8 shows these entropy values for top hate categories in our US hate speech data. We note that categories that are related to crimes such as other, behavior, and physical features are more uniformity distributed across all the states. On the other hand, hate speech on crime related topics, such race, sexual orientation and

[7] We adopted the following division of US States across regions http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

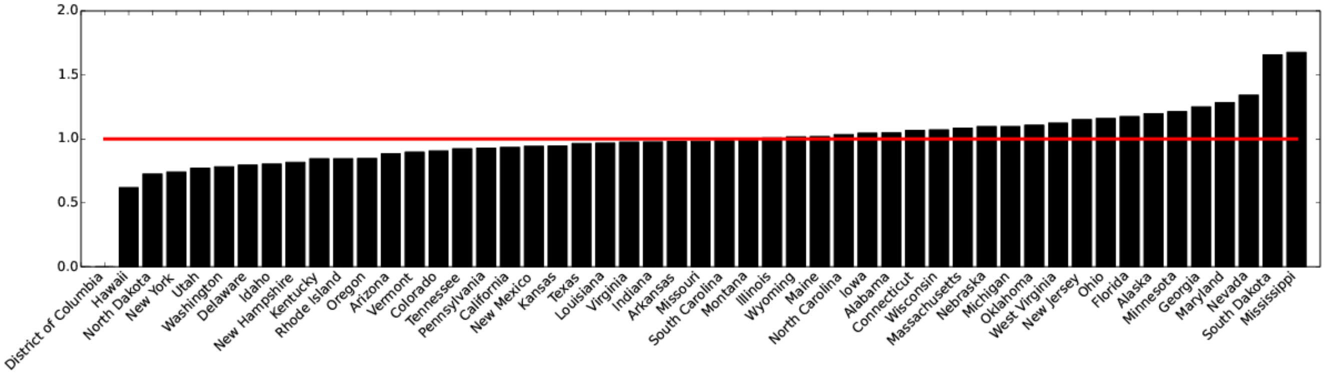

Figure 1: Relative amount of hate speech posted by users from different US states.

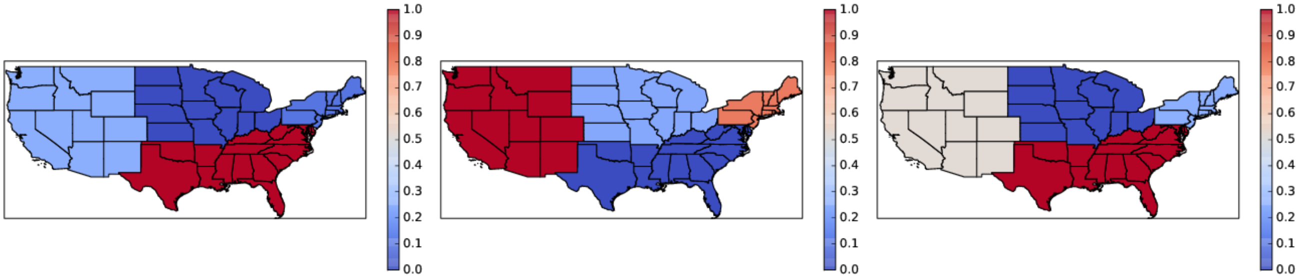

Figure 2: Heat map of US regions showing the distribution of normalized volume of hate speech for three hate categories: Race (left), Physical (center), and Sexual orientation (right).

Category	Hate entropy across US states
Other	5.1334
Behavior	5.0575
Physical	4.9550
Race	4.9313
Sexual orientation	4.8423
Class	4.4207

Table 8: Hate entropy across U.S. states for different hate categories.

class present lower values of entropy, indicating that the distribution of popularity of these forms of hate speech is more skewed across states. This observation further suggests that local actions and interventions for specific types of hate speech in specific locations (even within a country) is necessary.

8 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 (which provide the context for 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).

Thus to filter out the context we take our detected hate speeches, and for each hate speech remove the parts of sentences that matched our hate expression (along with the hate target). The resulting sentence gives us the context for that hate speech. For example, in the hate speech "I hate black people, their point of view is subhuman", we extract the (partial) sentence "their point of view is subhuman" as context for this hate speech.

Figure 3 shows a WordTree [8] 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 4 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,

[8] https://www.jasondavies.com/wordtree/

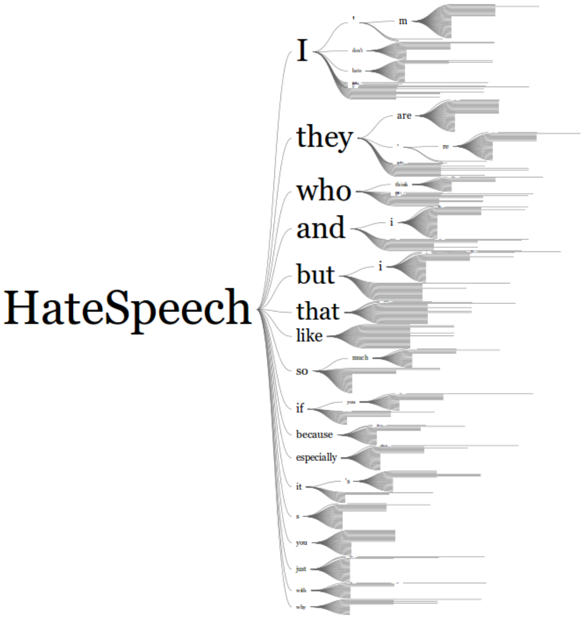

Figure 3: Word tree for our contexts of hate speeches.

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 causes of hate in some regions and places.

9 IMPLICATIONS

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. On a broad level our findings have three important implications:

Improving current keyword monitoring systems: A key aspect of a hate speech detection algorithm is that it must be able to classify messages in near real-time, as the longer the hateful message stays online, the larger is its damage to individuals. For example, Facebook, Google, and Twitter have recently agreed a deal with Germany under which they would remove hate speech posted on their websites within only 24 hours [9]. Moreover, recently, Google even promised marketers to find online hate speech to make sure advertisements are not shown alongside such content [10]. In this context, a key challenge is to identify new hate targets constantly. Our measurement methodology in this work is different from the ones used in previous efforts since we primarily used sentence structures to detect different types of hate speech instead of specific keywords. An interesting side effect is that our study is unique in the sense that it unveils a number of explicit hate targets (i.e. keywords associated to hate speech). Our effort even unveils new forms of online hate that are not necessarily crimes, but can still be harmful to people. We hope that our dataset and methodology can help monitoring systems and detection algorithms to identify novel keywords related to hate speech as well as inspire more elaborated mechanisms to identify online hate speech. To that end, building a hate speech detection system leveraging our findings is also part of our future work. In fact, such a system could be easily leveraged to notify the haters (users who post hate speech) and help them to better understand what harm they are causing.

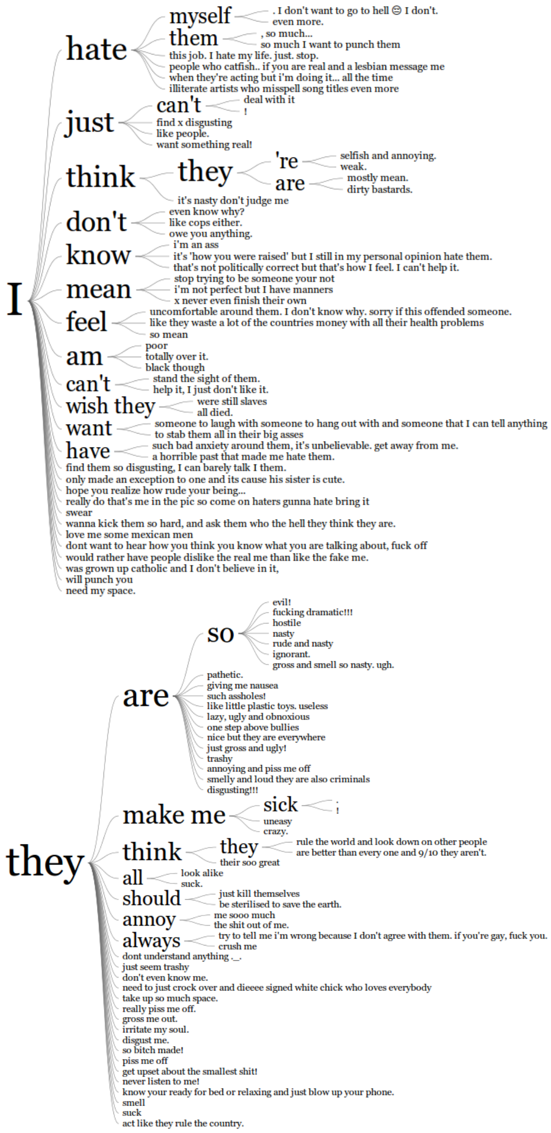
I
hate
myself
. I don't want to go to hell 😑 I don't.
even more.
them
, so much...
so much I want to punch them
this job. I hate my life. just. stop.
people who catfish.. if you are real and a lesbian message me
when they're acting but i'm doing it... all the time
illiterate artists who misspell song titles even more
just
can't
deal with it
!
find x disgusting
like people.
want something real!
think
they
're
selfish and annoying.
weak.
are
mostly mean.
dirty bastards.
it's nasty don't judge me
don't
even know why?
like cops either.
owe you anything.
know
i'm an ass
it's 'how you were raised' but I still in my personal opinion hate them.
that's not politically correct but that's how I feel. I can't help it.
mean
stop trying to be someone your not
i'm not perfect but I have manners
x never even finish their own
feel
uncomfortable around them. I don't know why. sorry if this offended someone.
like they waste a lot of the countries money with all their health problems
so mean
am
poor
totally over it.
black though
can't
stand the sight of them.
help it, I just don't like it.
wish they
were still slaves
all died.
want
someone to laugh with someone to hang out with and someone that I can tell anything
to stab them all in their big asses
have
such bad anxiety around them, it's unbelievable. get away from me.
a horrible past that made me hate them.
find them so disgusting, I can barely talk I them.
only made an exception to one and its cause his sister is cute.
hope you realize how rude your being...
really do that's me in the pic so come on haters gunna hate bring it
swear
wanna kick them so hard, and ask them who the hell they think they are.
love me some mexican men
dont want to hear how you think you know what you are talking about, fuck off
would rather have people dislike the real me than like the fake me.
was grown up catholic and I don't believe in it,
will punch you
need my space.
they
are
so
evil!
fucking dramatic!!!
hostile
nasty
rude and nasty
ignorant.
gross and smell so nasty. ugh.
pathetic.
giving me nausea
such assholes!
like little plastic toys. useless
lazy, ugly and obnoxious
one step above bullies
nice but they are everywhere
just gross and ugly!
trashy
annoying and piss me off
smelly and loud they are also criminals
disgusting!!!
make me
sick
;
!
uneasy
crazy.
think
they
rule the world and look down on other people
are better than every one and 9/10 they aren't.
their soo great
all
look alike
suck.
should
just kill themselves
be sterilised to save the earth.
annoy
me sooo much
the shit out of me.
always
try to tell me i'm wrong because I don't agree with them. if you're gay, fuck you.
crush me
dont understand anything ._.
just seem trashy
don't even know me.
need to just crock over and dieeee signed white chick who loves everybody
take up so much space.
really piss me off.
gross me out.
irritate my soul.
disgust me.
so bitch made!
piss me off
get upset about the smallest shit!
never listen to me!
know your ready for bed or relaxing and just blow up your phone.
smell
suck
act like they rule the country.

Figure 4: Word tree for context of hate speech starting with the words *I* and *They*.

[9] http://www.bbc.com/news/world-europe-35105003

[10] https://www.nytimes.com/2017/03/21/technology/google-advertising-apologizes-ad.html

Strong and weak online identities: Our findings quantitatively suggest that having a stronger notion of online identify may help individuals to better behave in the online world. This observation has interesting implications for system developers and researchers working in this space. For example, a social media platform could apply different hate speech monitoring strategies in their posts, i.e. heavily monitor the posts of those who do not use strong identities.

Leveraging online hatespeech to detect offline hate: We noted some tension related to racism in US as well as another tension related to sexual orientation in UK. These observations highlight the importance of monitoring hate speech in the online world in order to gain knowledge about hate in the offline world. Although there is no doubt that hate speech should be rapidly removed from social media platforms, the very removed data might provide a unique opportunity to identify the root causes of the offline hate.

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*

ACKNOWLEDGMENTS

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-1, and individual grants from CNPq, CAPES, Fapemig, and Humboldt Foundation.

REFERENCES

[1] Swati Agarwal and Ashish Sureka. 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).*

[2] J. Bartlett, J. Reffin, N. Rumball, and S. Williamson. 2014. *Anti-social media.* DEMOS.

[3] Irfan Chaudhry. 2015. #Hashtagging hate: Using Twitter to track racism online. *First Monday* 20, 2 (2015).

[4] Adrian Chen. 2014. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. https://www.wired.com/2014/10/content-moderation/. (2014). (Accessed on May 2017).

[5] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.*

[6] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *International Conference on Web and Social Media (ICWSM).*

[7] Denzil Correa, Leandro Silva, Mainack Mondal, Fabricio Benevenuto, and Krishna P. Gummadi. 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).*

[8] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Proceedings of 35th European Conference on IR Research.*

[9] Richard Delgado and Jean Stefancic. 2004. *Understanding words that wound.* Westview Press.

[10] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online Hate Speech.* UNESCO.

[11] Liz Gannes. 2013. On Making Our Digital Lives More Real. http://allthingsd.com/20130802/im-so-over-oversharing-on-making-our-digital-lives-more-real/. (August 2013).

[12] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 4 (2015), 215–230.

[13] Edel Greevy and Alan F. Smeaton. 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.*

[14] Erin Griffith. 2013. 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/. (May 2013).

[15] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).

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

[17] Grace Chi En Kwan and Marko M. Skoric. 2013. Facebook Bullying: An Extension of Battles in School. *Computers in Human Behavior* 29, 1 (2013), 16–25.

[18] I. Kwok and Y. Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of The AAAI Conference on Artificial Intelligence (AAAI'13).*

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

[20] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[21] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is It Biased?: Assessing the Representativeness of Twitter's Streaming API. In *Proceedings of the 23rd International Conference on World Wide Web.*

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

[23] Julio Reis, Fabrício Benevenuto, Pedro O.S. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the News: First Impressions Matter on Online News. In *International Conference on Web and Social Media (ICWSM).*

[24] Huascar Sanchez and Shreyas Kumar. 2011. Twitter bullying detection. *ser. NSDI* 12 (2011).

[25] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *International Conference on Web and Social Media (ICWSM).*

[26] M. Stephens. 2013. The Geography of Hate Map. http://users.humboldt.edu/mstephens/hate/hate_map.html. (2013).

[27] John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.

[28] Twitter team. 2017. The Streaming APIs. https://dev.twitter.com/streaming/overview. (2017).

[29] I-Hsien Ting, Hsing-Miao Chi, Jyun-Sing Wu, and Shyue-Liang Wang. 2013. An approach for hate groups detection in facebook. In *Proceedings of The 3rd International Workshop on Intelligent Data Analysis and Management (IADM'13).*

[30] Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, and Ben Y. Zhao. 2014. Whispers in the Dark: Analyzing an Anonymous Social Network. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC'14).*

[31] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media (LSM'12).*

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