Interaction dynamics between hate and counter users on Twitter

Binny Mathew
binnymathew@iitkgp.ac.in
Indian Institute of Technology, Kharagpur

Navish Kumar
navish.iitkgp@gmail.com
Indian Institute of Technology, Kharagpur

Pawan Goyal
pawang.iitk@gmail.com
Indian Institute of Technology, Kharagpur

Animesh Mukherjee
animeshm@gmail.com
Indian Institute of Technology, Kharagpur

ABSTRACT

Social media platforms usually tackle the proliferation of hate speech by blocking/suspending the message or account. One of the major drawback of such measures is the restriction of free speech. In this paper, we investigate the interaction of hatespeech and the responses that counter it (aka *counterspeech*). One of the prime contribution of this work is that we developed and released[1] a dataset where we annotate pairs of hate and counter users.

We perform several lexical, linguistic and psycholinguistic analysis on these annotated accounts and observe that the couterspeakers of the target communities employ different strategies to tackle the hatespeech. The hate users seem to be more popular as we observe that they are more subjective, express more negative sentiment, tweet more and have more followers. While the hate users seem to use words more about *envy, hate, negative emotion, swearing terms, ugliness*, the counter users use more words related to *government, law, leader*. Finally, we build a classifier to detect if a user is a hateful or counter speaker. This identification can help the platform to devise different incentive mechanisms to demote hate and promote counter speakers. Overall, our study unfolds for the first time, the interaction dynamics of the hate and counter users which could pave a more effective way for combating hate content on Twitter than just suspending the hate accounts.

CCS CONCEPTS

• **Social and professional topics** → **Hate speech**; **User characteristics**; **User characteristics**; **Hate speech**; • **Human-centered computing** → *Social media*; *Social network analysis*; • **Applied computing** → *Annotation*.

KEYWORDS

Counterspeech, Hate Speech, Twitter, Dataset

ACM Reference Format:
Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on Twitter. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3371158.3371172

[1]Dataset and Model: https://github.com/binny-mathew/Counterspeech_Twitter

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371172

1 INTRODUCTION

The online proliferation of hatespeech[2] has caused several countries and companies to implement laws against hatespeech to enforce the citizens to restrain from such behavior. Countries like Germany, United States of America, France have laws banning hatespeech. Social media sites such as Twitter, Facebook usually respond to hatespeech with the suspension or deletion of the message or the user account itself.

While these laws may reduce the amount of hatespeech in online media, it does so at the cost of causing harm to the freedom of speech. Another potential alternative to tackle hatespeech is "counterspeech". The main idea behind counterspeech is to add more speech to the conversation and try to change the mindset of the hate speaker.

This requirement led countries and organizations to consider counterspeech as an alternative to blocking [16]. The idea that 'more speech' is a remedy for harmful speech has been familiar in liberal democratic thought at least since the U.S. Supreme Court Justice Louis Brandeis declared it in 1927. There are several initiatives with the aim of using counterspeech to tackle hatespeech. For example, the Council of Europe supports an initiative called 'No Hate Speech Movement'[3] with the aim to reduce the levels of acceptance of hatespeech and develop online youth participation and citizenship, including in Internet governance processes. UNESCO released a study [16] titled 'Countering Online Hate Speech', to help countries deal with this problem. Social platforms like Facebook have started counterspeech programs to tackle hatespeech[4]. Facebook has even publicly stated that it believes counterspeech is not only potentially more effective, but also more likely to succeed in the long run [2]. Combating hatespeech in this way has some advantages: it is faster, more flexible and responsive, capable of dealing with extremism from anywhere and in any language and it does not form a barrier against the principle of free and open public space for debate.

1.1 The present work

In this paper, we perform the first comparative study of the interaction dynamics of the hatespeech and the corresponding counterspeech replies. We choose Twitter as our source of data and curate

[2]https://goo.gl/4rWGiF
[3]No Hate Speech Movement Campaign: http://www.nohatespeechmovement.org/
[4] Counterspeech Campaign by Facebook: https://counterspeech.fb.com/en/

a dataset with 1290 hate tweet and counterspeech reply pairs. After the annotation process, the dataset consists of 558 unique hate tweets from 548 user and 1290 counterspeech replies from 1239 users. We found that 75.39% of these replies are counterspeech that oppose the hatespeech spread by the user.

1.2 Contributions

The main contributions of our paper are as follows:

- We perform the first comparative study that looks into the characteristics of the hateful and counter accounts.
- We provide a dataset[1] of 1290 tweet-reply pair in which the tweets are the hatespeech and their replies are counterspeech.
- We develop a model[1] which predicts if a given Twitter user is a hateful or counterspeech account with an accuracy of 78%.

1.3 Observations

Our study results in several important observations.

- First, we find significant difference in the activity pattern between these users. Hateful accounts tend to express more negative sentiment and profanity in general. If the hateful tweet is from a verified account, it seems to be much more viral as compared to other hateful tweets.
- Another intriguing finding is that hateful users also act as counterspeech users in some situations. In our dataset, such users use hostile language as a counterspeech measure 55% of the times.
- In terms of personality traits, the counterspeakers seem to have a higher quotient of 'agreeableness'. They are more altruistic, modest and sympathetic. The hateful users, on the other hand, seem to have a higher quotient of 'extraversion' indicating that they are more energetic and talkative in nature. They are more cheerful, excitement-seeking, outgoing, and sociable in nature.
- Using the linguistic structure of the general tweets and the account characteristics of the hateful and the counter speakers, it is possible to distinguish them early with an accuracy of 78%.

2 PRELIMINARIES

- **Hatespeech:** We define hatespeech according to the Twitter guidelines. Any tweet that 'promotes violence against other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease' is considered as a hatespeech[5]. A **hate account (HA)** is a Twitter account which posts *one or more* such hateful tweets.
- **Counterspeech:** We follow the definition of counterspeech used in [25]. We call a tweet as a 'counterspeech' if the tweet is a direct reply to a hateful tweet. A **counter account (CA)** is a Twitter account which posts one or more such counterspeech in response to a hateful post.

[5]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

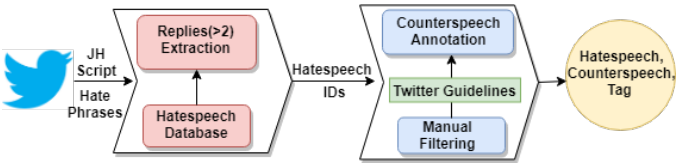


Figure (1) A flowchart showing the process by which we obtained the dataset.

3 DATASET COLLECTION

We use Twitter as the platform for this study. As we are interested in the replies received by a particular hate tweet, we could not use the Twitter API directly as it does not provide any service that helps in directly collecting the replies of a particular tweet. To collect these replies we utilize the PHEME script[6] which allows us to collect the set of tweets replying to a specific tweet, forming a conversation. Other works [39] have used this technique as well. Our overall methodology is a multi-step approach for creating the dataset which can be divided into the following three steps:

- Hateful tweet collection.
- Filtration and annotation of the hateful tweets.
- Extraction and annotation of the reply tweets.

We explain these three steps in detail below.

3.1 Hatespeech collection

We rely on the hatespeech templates defined in Silva et al. [32] to collect the hateful tweets. These templates are of the form:

I <intensity><userintent><hatetarget>

The subject "I" implies that the user is expressing her own personal emotions. The verb, embodied by <userintent> component is used to specify the user's intent which in this case is the word 'hate' or its synonyms. The component <intensity> is optional and acts as a qualifier which some users use to amplify their emotions. Words such as 'really','f**cking' are used to express the <intensity>. In our work, we have used the words that are listed in [26] for each of the component. The component <hatetarget> is used to find the target community on the receiving end of the hatred. Table 2 lists the keywords that we have used to extract hateful tweets for each of these communities. Some examples of the these templates include: "I hate muslims", "I really despise white people" etc.

Next, we utilize the Jefferson-Henrique's web scraping script[7] to collect the hateful posts using the templates defined above. We run the script for all the communities defined in Table 2 to collect around 578K tweets in total. We removed all non-English tweets from the dataset.

3.2 Filtration and annotation of hateful tweets

As an initial filtering step to ensure that we get sufficient replies to a hateful tweet, we filter out all the tweets collected in the previous step that did not have at least two replies. This reduced the number of hatespeech tweets to 13,321. Next, we manually tagged the tweets as containing hatespeech or not. We follow the guidelines defined

[6]https://github.com/azubiaga/pheme-twitter-conversation-collection
[7]https://github.com/Jefferson-Henrique/GetOldTweets-python

Counterspeech Type	Hatespeech	Counterspeech	Total
Presentation of facts	A tragedy that is a DIRECT result of allowing too many f**king Muslims in their country. There should not be any surprises here. #BANMUSLIMS	The guy was born in UK in 1995	135
Pointing our hypocrisy	I hate Fat ppl bruhhhh ughhh	"@user: I hate Fat ppl bruhhhh ughhh" you use to be fat you hypocrite!	177
Warning of consequences	I hate lesbians so much omfg	I report you to the LGBT community, this blatant discrimination will not be tolerated	69
Affiliation	I'm christian so i hate muslims, muslim turkish people and MUSLIM TURKISH BELIEBERS cause YOU ARE THE TERRORIST!Turkish BELIEBERS Need Bieber	@user IM MUSLIM BUT İ DO NOT HATE CRISTIANS	58
Denouncing speech	I hate gays	"@user: I hate gays." Stop being homophobic	132
Images	- - -	- - -	146
Humor	I hate gays	@user so uhhh I guess you hate yourself huh? HAHAHHAHAHAHAHA man I'm funny	135
Positive tone	@user the problem with the world is you f**king muslims! go choke to death on some bacon you child raper pig! f**k you a**hole!	@user, I'm sorry you feel that way. Islam is a beautiful religion, so misunderstood. Not all Muslims are the same.	175
Hostile language	@user f**k the muslims!!!	@user you are truly one stupid backwards thinking mother**cker to believe negativity about Islam	357
Miscellaneous	F**k I hate women! All bit**es to a good guy	@user Hey hey now.. ALL women are NOT the same.	198
	Total		1582

Table (1) Example hatespeech and the corresponding counterspeech received for each type. Please note that the total is more than the reported 1290 counterspeech replies because the users sometimes employ multiple strategies in a single tweet.

Gender	Ethnicity	Physical Trait
Men	Nigger	Fat
Women	Nigga	
Female	White people	
	Black people	

Sexual Orientation	Nationality	Religion
Lesbian	American	Jew
Gay	Indian	Islam
LGBT	Mexican	Muslim
Transgender	Arab	
	Asian	

Table (2) Keywords used during search to select hatespeech for different communities.

by Twitter [5] to identify if a tweet contains hatespeech or not. Each tweet was annotated by two users and in case of a disagreement, we ask one more annotator to resolve it. The annotators were instructed to take the context into consideration when annotating a tweet as hateful. The annotations were performed by a group of two undergraduate students with Major in Computer Science and a PhD student in Social Computing. The final dataset consisted of 558 tweets that were tagged as hateful. Figure 1 illustrates the whole process in a flowchart.

3.3 Extraction and annotation of the reply tweets

Once we selected the hatespeech tweets with at least two replies, the next step involved using the PHEME script[6] to scrape the replies received to these hateful tweets. We observe that the 558 hate tweets received a total of 1711 replies with a mean of 3.067 and a median of 2 replies per tweet. We employ two annotators for the counterspeech tagging. We follow the procedure used in Mathew et al. [25] to annotate the counterspeech data. For each reply text to a hatespeech tweet, the annotators were asked to perform two tasks: (i) decide if the reply is a counterspeech or not, and (ii) if the reply is indeed a counterspeech, then decide the category of the counterspeech. We describe these categories in the subsequent section.

Two independent annotators tagged each reply tweet as counterspeech or not and got an accuracy of 81.35% and κ score of 0.46, which is moderately good. For the second task of deciding on the category of counterspeech, we use the Exact Match Ratio and Accuracy defined in Sorower [33]. The two annotators obtain an Exact Match Ratio of 71.4% and an Accuracy of 77.58%. We employ a third annotator to resolve the conflicting cases and report the final distribution in Table 1.

4 TAXONOMY OF COUNTERSPEECH

A successful counter to a hatespeech will require various types of strategies. In [5], the authors define eight such strategies that are used by counterspeakers. Similar to [25], we divide the *Tone* category into two parts: positive and negative Tone. We aggregate all the counterspeech that would not fit in the above categories into a 'Miscellaneous' category. Note that sometimes the users employ multiple counterspeech strategies in a single tweet. We refer the readers to [5, 25] for a detailed discussion on the taxonomy of counterspeech. Table 1 shows an example for each type of counterspeech.

5 TARGET COMMUNITY ANALYSIS

Our final dataset consisted of 558 hatespeech tweets from 548 Hate Accounts (HAs) and 1290 direct replies that were counterspeech from 1239 Counterspeech Accounts (CAs). Of the total 1711 tweets that were replies to the 558 hate tweets, 75.39% (1290) were counterspeech tweets. This is almost double of what is reported in [25]. We argue that the main reason behind this could be the public nature of Twitter as opposed to the semi-anonymous nature of Youtube.

Hate Target	Gender	Sexuality	Nationality	Religion	Physical Trait	Ethinicity	Total
Presentation of facts	1 (00.35%)	5 (02.49%)	5 (04.07%)	124 (21.87 %)	0 (00.00%)	2 (00.93%)	137 (08.94%)
Pointing out hypocrisy	38 (13.48%)	19 (9.45%)	16 (13.01%)	100 (17.64%)	7 (4.83%)	7 (3.26%)	187 (12.20%)
Warning of consequences	3 (01.06%)	9 (4.48%)	4 (3.25%)	20 (3.53%)	2 (1.38%)	25 (11.63%)	63 (4.11%)
Affiliation	14 (04.97%)	9 (4.48%)	9 (7.32%)	19 (3.35%)	2 (1.38%)	4 (1.86%)	57 (3.72%)
Denouncing speech	15 (05.32%)	20 (9.95%)	12 (9.76%)	50 (8.82%)	3 (2.07%)	34 (15.81%)	134 (8.74%)
Images	25 (08.93%)	15 (7.46%)	15 (12.30%)	10 (1.77%)	2 (1.38%)	17 (7.91%)	84 (5.49%)
Humor	32 (11.35%)	30 (14.93%)	6 (4.88%)	40 (7.06%)	12 (8.28%)	8 (3.72%)	128 (8.35%)
Positive tone	47 (16.67%)	34 (16.92%)	13 (10.57%)	37 (6.53%)	15 (10.34%)	13 (6.05%)	159 (10.37%)
Hostile language	50 (17.73%)	39 (19.40%)	32 (26.02%)	112 (19.75%)	65 (44.83%)	81 (37.67%)	379 (24.72%)
Miscellaneous	55 (19.50%)	21 (10.45%)	10 (8.13%)	54 (9.52%)	37 (25.52%)	24 (11.16%)	201 (13.11%)
Total counter	282	201	123	567	145	215	1533
Total hate	120	110	43	143	91	99	606

Table (3) Counterspeech strategies used by various target communities. The percentage given in the bracket are normalized by the total number of counterspeech in each category. Please note that the total reported is more than the 1290 counterspeech replies because of the presence of multiple target communities in a single tweet.

We also found that 79.07% of the counterspeech employed only one kind of strategy. The p-values reported are calculated using Mann-Whitney U test.

5.1 Strategies used by CAs

The CAs used different strategies to tackle the hatespeech. We can observe from Table 3 that different target communities adopt different measures to respond to the hateful tweets. The largest fraction (23.60%) of the hateful tweets seems to be religious in nature. They also receive the highest number of counterspeech with an average of 3.97 counterspeech for every religious hateful tweet as compared to only 2.7 counterspeech per hateful tweet for the entire dataset. The religious CAs seems to be using two strategies more as compared to other target communities: presentation of facts and pointing out hypocrisy.

In case of the nationality target communities, CAs seem to be relying more on affiliation and using images in their counter speech. The CAs for the ethnicity use warning of the consequences and denounce the hatespeech more as compared to other target communities.

The CAs for the target communities associated with sexuality rely on humor and positive tone to counter the hateful messages. In case of the hatespeech which target the physical traits, the CAs heavily use hostile language to counter them.

Irrespective of the community, the counterspeakers seem to be using hostile language a lot with the lowest being for the gender (17.73%) and highest for the physical traits (44.83%). We can also observe that 24.72% of the counterspeech tweets used hostile language as a strategy. This is less than the 35% reported in [25] on YouTube. One of the main reasons for this could be the more public nature of Twitter.

5.2 Use of images/videos

We observe that only 12 (2.15%) of hate tweets contained images/videos whereas 146 (11.24%) of the counter tweets had images/videos. We also look into the replies of these 12 hate tweets which use images/videos to get a better understanding. Interestingly, we found that 69 (47.26%) out of the 145 counter tweets which use images/videos were in response to hateful tweets which themselves used images/videos. Another interesting observation was that the counterspeech involving images/videos are liked more by the Twitter community as compared to other strategies. The counter images received the highest likes among all the counterspeech strategies with a median of 2.5 likes (average = 10).

6 USER LEVEL ANALYSIS

In this section, we characterize the HAs and CAs based on their activity and tweet history. We first collect for each user their last 3200 tweets using the Twitter API[8]. This would also give us other information about the users such as the number of tweets, number of followers, friends etc.

6.1 User activity

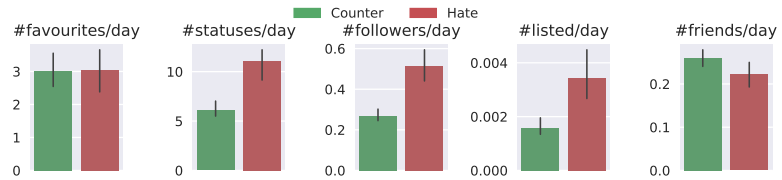
We normalize the user-characteristics by dividing them by the age of the account. Thus each of the user properties are divided by the number of days since the start of the account. In Figure 2a, we can observe several striking differences in the nature of the accounts. We observe that the hate users are more "popular" in that they *tweet more*, *have more followers*, and *are part of more public lists* as compared to the counter users (p-value<0.001).

The counter users have *more friends* per day as compared to hate users (p-value<0.005).

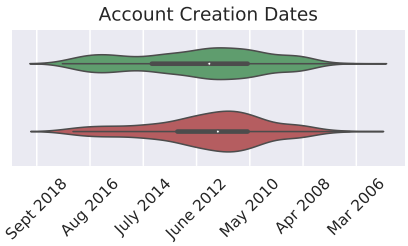
6.2 Creation dates

We analyze the account creation dates of hate and counter users as shown in Figure 2b. Previous works on hate users [13, 30] have reported that the Twitter accounts of hateful users have relatively less age as compared to the normal users. Our results, in contrast, do not seem to support this observation. We find that the *hate accounts are older than the counter accounts* (p-value<0.01). We argue that the main reason for this is due to the way we collected our dataset. Around 80% of the hateful tweets are older than Dec 2017. Around this time Twitter first started enforcing stricter rules on abusive

[8]Twitter API: https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html



(a) Average values for several user profile properties for Hateful and Counter user normalized by the account age. Error bars represent 95% confidence interval.



(b) KDEs of the account creation dates of Hateful and Counter users.

Figure (2) User account characteristics

content, which means that they managed to bypass the platform's strict guidelines. Thus, many of the hate accounts are much older.

6.3 Suspended and deleted accounts

The Twitter API returns an error message when the user account is suspended or the user is not found. According to Twitter[9], the common reasons for account suspension include spam, risk to Twitter's security, or abusive tweet or behavior. Twitter accounts are not found when the user does not exist. This may be because the user deactivated her account or that the account was permanently deleted after thirty days of deactivation. We found that 4.56% and 4.28% of the hate and counter users, respectively, were deleted.

6.4 Verified accounts

Next, we look for the presence of verified accounts in our dataset. We found that 3.28% of the hate users are verified; in contrast, 0.56% of the counter users have verified accounts. We found that the hateful tweets by the HAs generated much more audience as compared to the CAs. On average (median) the verified HAs received 356.42 (57.5), 80.42 (13), 24.08 (16) likes, retweets, and replies, respectively. This is much higher than the verified CAs who received 1.8 (2.0), 0.4 (0.0), 0.4 (0.0) likes, retweets, and replies, respectively.

6.5 Tweet analysis

To understand how the different set of users express their emotions, we make use of the users' tweet history. We apply VADER [21] to find the average negative sentiment expressed by the users. In Figure 3, we observe that the tweets from hateful users express more negative sentiments compared to counter users. We use TextBlob[10] to measure the subjectivity expressed in the tweets. As observed from Figure 3, hate accounts use more subjective tweets as compared to counters accounts (p<0.001). In order to find the profanity expressed in the tweets, we use Shutterstock's "List of Dirty, Naughty, Obscene, and Otherwise Bad Words"[11]. We observe that hate accounts use more profane words as compared to counter

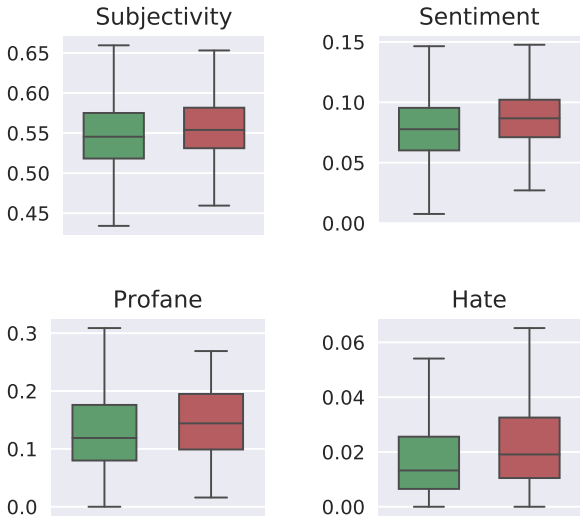


Figure (3) Boxplots for the distribution of Subjectivity, Negative Sentiment, Profanity, and Hatespeech for the Hateful and Counter users

accounts (p-value<0.05). We use the model provided by Davidson et al. [8] to check for hatespeech and abusive language in the tweet history of user. We found that the hate users seem to use more hatespeech and abusive language as compared to counter speakers.

6.6 Lexical analysis

Empath [15] is a tool that can be used to analyze text over 200 pre-built lexical categories. We characterize the users over these categories and observe several interesting results as shown in Figure 4. The hate users seem to be using more words related to categories such as *envy, hate, negative emotion, positive emotion, ridicule, swearing terms, ugliness.* These results are in alignment with the sentiment values obtained for these users. The counter users seem to be using words in the categories such as *government, law, leader, pride, religion, terrorism* indicating that the counter users are more civil in nature. All the values reported have p-value<0.01.

[9]https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts

[10]https://github.com/sloria/textblob

[11]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

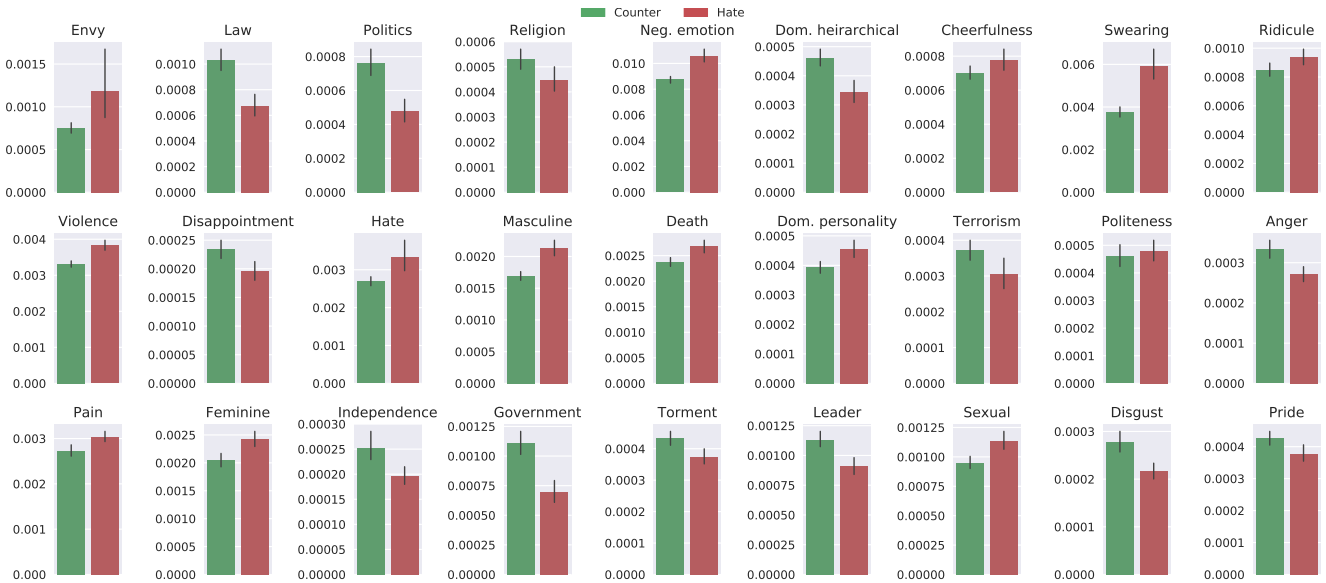


Figure (4) Lexical analysis using Empath for the hateful and counter users. We report the median values for the relative occurrence in several categories of Empath. Error bars represent 95% confidence intervals. That HAs score significantly higher in topics like envy, hate, negative emotion, positive emotion, ridicule, swearing terms, ugliness while CAs score significantly higher in topics like government, law, leader, pride, religion, terrorism

6.7 Personality analysis

In order to get a better understanding of the user characteristics, we perform personality analysis on the user set. We make use of the IBM Watson Personality Insights API [19] [12] to understand the personality of the users. Previous research [20, 23] has used this tool to understand the personality of the users. We provide all the tweets of a particular user as input to the API and the service analyzes the linguistic features to infer the Big-5 personality traits [18]. We observe distinctive difference between the hate and counter users as shown in Figure 5.

CAs score *higher in the personality traits such as agreeableness.* They are *more altruistic, modest, sympathetic.* This is intuitive since the counter speakers in many cases have a simple motive to help the hate speakers. They also seem to be *more self-disciplined.* The counter speakers score higher than the hate speakers in all the sub-traits of the conscientiousness. The counter speakers are *more driven, deliberate, dutiful, persistent*, and *self-assured.*

The HAs seem to score *higher in extraversion* indicating that they are *more energetic* and *talkative* in nature. They are more *cheerful, excitement-seeking, outgoing*, and *sociable* in nature.

7 CLASSIFICATION

In this section, we leverage the unique characteristics of the hate and counter user accounts to develop a predictive model. We pose this as a binary classification problem - distinguishing hate users from counter users. This early automatic distinction can help the platform to develop appropriate incentive mechanisms to demote the hate user accounts and promote the counter user accounts and also urge these users to help in overall purging of the media. Munger [27] showed that counterspeech using automated bots can reduce instances of racist speech if the instigators are sanctioned by a high-follower white male.

7.1 Features

We use the following feature set for our classifier.

- **TF-IDF values**: For each account, we calculate its tf-idf vector using the users tweet history.
- **User profile properties**: We use several user account properties such as #favorites per day, #tweets per day, #followers per day, #friends per day, #listed per day, and whether the account is verified.
- **Lexical properties**: We use the vector of 200 pre-built topics as the feature set for each user.
- **Affect properties**: For each user account we calculate the average sentiment, profanity, and subjectivity expressed in the tweet history.

7.2 Dataset

We first divide the dataset into training and test set in the ratio 90:10 while keeping the test set balanced. In order to find the hyperparameter of our models, we use 10% of the training data as the validation set. We run randomized grid search to find the optimal hyperparameter values.

7.3 Choice of classifier

We choose classifiers such as Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Extra-Tree (ET), XGBoost

[12] IBM Personality Insights API: https://www.ibm.com/watson/services/personality-insights/

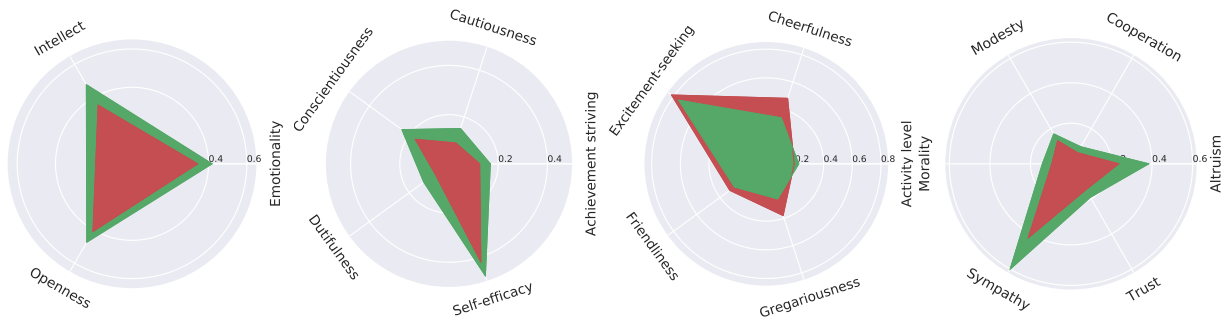


Figure (5) Personality traits of the CAs and HAs. The CAs are more self-disciplined, driven, deliberate, dutiful, persistent, and self-assured while the HAs are more energetic, talkative, cheerful, excitement-seeking, outgoing, and social in nature.

(XGB), and CatBoost (CB) for this task. We use TF-IDF features along with LR as the baseline.

7.4 Results

Table 4 shows the results of the classification task. We can observe that CatBoost (CB) performs the best with an accuracy of 78% followed by XGBoost (XGB) with an accuracy of 74%. Both the classifiers perform much better than the baseline classifier (LR + TF-IDF).

Model	Precision	Recall	F-score	Accuracy
LR + TFIDF	0.68	0.68	0.68	0.68
SVM	0.64	0.63	0.62	0.63
LR	0.66	0.66	0.66	0.66
ET	0.72	0.70	0.69	0.70
RF	0.72	0.72	0.72	0.72
XGB	0.74	0.74	0.74	0.74
CB	0.83	0.78	0.77	0.78

Table (4) Evaluation results for various classifiers - Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Extra-Tree (ET), XGBoost (XGB), and CatBoost (CB) on the task of classification of an account as hateful or counter. The evaluation metrics reported are calculated by taking macro average.

7.5 Feature ablation

We perform feature ablation to understand the importance of the different feature types. Table 5 reports the feature ablation study carried out by CatBoost (CB) classifier. From the table, we can observe that TF-IDF feature is the most useful followed by lexical features.

Feature excluded	Precision	Recall	F-score	Accuracy
TF-IDF	0.59	0.53	0.43	0.53
User profile	0.84	0.79	0.78	0.79
Lexical	0.65	0.56	0.49	0.56
Affect	0.83	0.77	0.76	0.77

Table (5) Feature ablation study for the CatBoost (CB) classifier. The evaluation metrics reported are calculated by taking macro average.

8 RELATED WORK

8.1 Hateful or harmful speech

Hatespeech lies in a complex nexus with freedom of expression, group rights, as well as concepts of dignity, liberty, and equality [16]. Owing to this, there can be several issues in defining what constitutes as hatespeech [3]. The authors usually adopt a definition that fits a general description of hatespeech. There is substantial literature on the analysis of hatespeech. Silva et al. [32] studies the targets of hatespeech on Twitter and Whisper and observes that 'Blacks' are among the most frequent targets of hatespeech. Mondal et al. [26] studies the effects of anonymity on hatespeech. There are works where the authors study the perceptions and experience of people regarding the hatespeech [17, 22]. In Van Spanje and De Vreese [35], the authors study the impact of hatespeech prosecution of a politician on electoral support for his party and find an immediate increase in the support for the party.

Another line of research corresponds to detection of hatespeech in various social media platforms like Twitter [1, 6, 8, 37], Facebook [10], Yahoo! Finance and News [11, 28, 36], and Whisper [26]. In another online effort, a Canadian NGO, the Sentinel Project[13], launched a site in 2013 called HateBase[14], which invites Internet users to add to a list of slurs and insulting words in many languages.

There are some works which have tried to characterize the hateful users. In Ribeiro et al. [29, 30], the authors study the user characteristics of hateful accounts on twitter and found that the hateful user accounts differ significantly from normal user accounts on the basis of activity, network centrality, and the type of content they produce. In ElSherief et al. [13], the authors perform a comparative study of the hate speech instigators and target users on twitter. They found that the hate instigators target more popular and high profile twitter users, which leads to greater online visibility. In ElSherief et al. [12], the authors focus on studying the target of the hatespeech - directed and generalized. They look into linguistic and psycholinguistic properties of these two types of hatespeech

[13] https://thesentinelproject.org/
[14] https://www.hatebase.org/

and found that while directed hate speech is more personal and directed, informal and express anger, the generalized hate is more of religious type and uses lethal words such as 'murder', 'exterminate', and 'kill'.

8.2 Countering hatespeech

The current solution adopted by several organization and companies to tackle online hate speech involves blocking/suspending the account or the particular hateful post. While these options are very powerful, they tend to violate the free speech doctrine. Counterspeech is considered to be a promising solution in this direction as it can help in controlling the hate speech problem and at the same time, it supports free speech. Institutions can help in educating the public about hatespeech and its implications, consequences and how to respond. Programmes such as 'No Hate Speech' movement [3] and Facebook's counterspeech program [4] help in raising awareness, providing support and seeking creative solutions [7].

Online social media sites such as Twitter, YouTube, and Facebook have been used to study counterspeech. Wright et al. [38] study the conversations on Twitter, and find that some arguments between strangers lead to favorable change in discourse and even in attitudes. Ernst et al. [14] study the comments in YouTube counterspeech videos related to Islam and find that they are dominated by messages that deal with devaluating prejudices and stereotypes corresponding to Muslims and/or Islam. In Schieb and Preuss [31], the authors study counterspeech on Facebook and through simulation, find that the defining factors for the success of counter speech are the proportion of the hatespeech and the type of influence the counter speakers can exert on the undecided. Stroud and Cox [34] perform case studies on feminist counterspeech.

Benesch et al. [4] describes strategies that have favorable impact or are counterproductive on users who tweet hateful or inflammatory content. In Mathew et al. [25], the authors curate a counterspeech dataset from YouTube and build a general classifier which achieves an F1-score of 0.73 .

This work is the first to study the interaction dynamics of the hate and counter speakers in a coupled fashion. We have released the dataset used in our experiment as well as the model for classifying an account as hateful or counter [1].

9 DISCUSSION

9.1 Success of counterspeech

We found cases in which the counterspeech provided was successful in changing the mindset of the user who had posted the hateful tweet. In one case, the user who had posted the hateful message on Twitter, later apologized to everyone saying that she was really sorry for what she did. Although, we rarely receive such direct evidence of a counterspeech being successful, these cases prove that counterspeech can actually help in changing the attitude of the hate users without resorting to aggressive measures such as blocking/suspension of the account.

9.2 New venues for hatespeech

The harsh nature of Twitter on hatespeech has led to several user account suspension and deletion. This is true for other social media sites as well. Due to this, several new social media sites have sprung up in the recent years like Gab[15], Wrongthink[16] which support free speech on their site and allow users to posts contents that would not be allowed on other sites. This has resulted in an increase in the spread of hate speech [24] and gab becoming an echo chamber for right-leaning dissemination [9].

9.3 Common users

We found some users which were common in HAs and CAs. On examining these accounts we observe that these accounts take the role of hate or counter user depending on the situation. For example, we observe that a user who had performed a hate speech targeting the 'LGBTQ' community also performed a counterspeech in response to a hatespeech targeting 'Fat people'. We found in cases where the HAs act as counter speaker, 55% of the times they use the hostile language strategy.

9.4 Hate supporters

While annotating the dataset, we also found several tweets that were in support of the hate tweet. Specifically, we annotated 223 direct replies to the hate tweets that were in support. We found several interesting properties of these hate support accounts (**HSAs**). To begin with, we found that HSAs had several similarities with the HAs. They use more profane and subjective words, their accounts are relatively newer as compared to HAs and CAs, and hashtags such as 'whitegenocide' were also used by them. As per lexical analysis, the HSAs use more words in the categories such as *anger, independence, suffering*. The HSAs seem to be more open in nature as compared to CAs and HAs. They are *more imaginative, philosophical* and *authority-challenging*.

10 CONCLUSION

In this paper, we perform the first characteristic study comparing the hateful and counterspeech accounts in Twitter. We provide a dataset of 1290 tweet-reply pairs of hatespeech and the corresponding counterspeech tweets. We observe that the counter speakers of different communities adopt different techniques to counter the hateful tweets. We perform several interesting analysis on these accounts and find that hateful accounts express more negative sentiments and are more profane. The hateful users in our dataset seem to be more popular as they tweet more and have more followers. We also find that the hate tweets by verified accounts have much more virality as compared to a tweet by a non-verified account. While the hate users seem to use words more about envy, hate, negative emotion, swearing terms, ugliness, the counter users use more words related to government, law, leader. We also build a supervised model for classifying the hateful and counterspeech accounts on Twitter and obtain an F-score of 0.77. Our work should be useful in appropriately designing incentive mechanisms to make Twitter-like platforms free of hate content.

ACKNOWLEDGMENTS

The authors are grateful to Ravina, for her help in the initial data collection. Similarly, the author would like to thank the editor and anonymous referees for their helpful comments.

[15] https://gab.com
[16] https://wrongthink.net/

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets *(WWW)*. 759–760.

[2] Jamie Bartlett and Alex Krasodomski-Jones. 2015. Counter-speech examining content that challenges extremism online. *Demos. Available at: http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf* (2015).

[3] Susan Benesch. 2014. Defining and diminishing hate speech. *Freedom from hate: State of the world's minorities and indigenous peoples 2014* (2014), 18–25.

[4] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for Successful Counterspeech. *Dangerous Speech Project. Available at: https://dangerousspeech.org/considerations-for-successful-counterspeech/* (2016).

[5] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A Field Study. *Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/* (2016).

[6] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying. In *WWW '17 Companion*. 1285–1290.

[7] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.* 91 (2011), 1435.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. (2017).

[9] Lucas Riguete Pereira de Lima, Julio C. S. Reis, Philipe F. Melo, Fabricio Murai, Leandro Araújo Silva, Pantelis Vikatos, and Fabrício Benevenuto. 2018. Inside the Right-Leaning Echo Chambers: Characterizing Gab, an Unmoderated Social System. (2018).

[10] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. (2017).

[11] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *WWW '15 Companion*. 29–30.

[12] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Y. Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media *(ICWSM '18)*.

[13] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets.

[14] Julian Ernst, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer, Gary Bente, and Hans-Joachim Roth. 2017. Hate Beneath the Counter Speech? A Qualitative Content Analysis of User Comments on YouTube Related to Counter Speech Videos. *Journal for Deradicalization* (2017), 1–49.

[15] Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4647–4657.

[16] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing.

[17] Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities* 22, 3 (2016), 324–341.

[18] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[19] Liang Gou, Michelle X. Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: Understanding Automatically Discovered Personality Traits from Social Media and User Sharing Preferences *(CHI '14)*. ACM, 955–964. https://doi.org/10.1145/2556288.2557398

[20] Tianran Hu, Haoyuan Xiao, Jiebo Luo, and Thuy-vy Thi Nguyen. 2016. What the Language You Tweet Says About Your Occupation. In *Tenth International AAAI Conference on Web and Social Media*.

[21] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

[22] Laura Leets. 2002. Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of social issues* 58, 2 (2002), 341–361.

[23] Zhe Liu, Anbang Xu, Yi Wang, Jerald Schoudt, Jalal Mahmud, and Rama Akkiraju. 2017. Does Personality Matter?: A Study of Personality and Situational Effects on Consumer Behavior *(HT '17)*. ACM, 185–193. https://doi.org/10.1145/3078714.3078733

[24] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 173–182.

[25] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 369–380.

[26] Mainack Mondal, Leandro Araujo Silva, and Fabricio Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *HT*.

[27] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* 39, 3 (2017), 629–649.

[28] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *WWW '16*. 145–153.

[29] Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter.

[30] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2017. " Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. In *WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. 8.

[31] Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ICA Annual Conference, At Fukuoka, Japan*. 1–23.

[32] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media.. In *ICWSM*. 687–690.

[33] Mohammad S Sorower. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis* 18 (2010).

[34] Scott R Stroud and William Cox. 2018. The Varieties of Feminist Counterspeech in the Misogynistic Online World. In *Mediating Misogyny*. Springer, 293–310.

[35] Joost Van Spanje and Claes De Vreese. 2015. The good, the bad and the voter: The impact of hate speech prosecution of a politician on electoral support for his party. *Party Politics* 21, 1 (2015), 115–130.

[36] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media (LSM '12)*. 19–26.

[37] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[38] Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for Counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*. 57–62.

[39] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS ONE* 11 (03 2016), 1–29. https://doi.org/10.1371/journal.pone.0150989