Thou shalt not hate: Countering Online Hate Speech

Preprint · August 2018

7 authors, including:

Binny Mathew
Indian Institute of Technology Kharagpur
11 PUBLICATIONS **2** CITATIONS

Subham Rajgaria
Indian Institute of Technology Kharagpur
2 PUBLICATIONS **0** CITATIONS

Prajwal Singhania
Indian Institute of Technology Kharagpur
2 PUBLICATIONS **0** CITATIONS

Suman Kalyan Maity
Indian Institute of Technology Kharagpur
36 PUBLICATIONS **101** CITATIONS

Thou shalt not hate: Countering Online Hate Speech

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania,
Suman Kalyan Maity*, **Pawan Goyal and Animesh Mukherjee**
IIT Kharagpur, India
*Kellogg School of Management, Northwestern University
Email: binnymathew@iitkgp.ac.in, hardik.tharad@gmail.com, subham.rajgaria@gmail.com,
prajwal1210@gmail.com, suman.maity@kellogg.northwestern.edu,
pawang@cse.iitkgp.ernet.in, animeshm@cse.iitkgp.ac.in

Abstract

Hate content in social media is ever increasing. While Facebook, Twitter, Google have attempted to take several steps to tackle this hate content, they most often risk the violation of freedom of speech. Counterspeech, on the other hand, provides an effective way of tackling the online hate without the loss of freedom of speech. Thus, an alternative strategy for these platforms could be to promote counterspeech as a defense against hate content. However, in order to have a successful promotion of such counterspeech, one has to have a deep understanding of its dynamics in the online world. Lack of carefully curated data largely inhibits such understanding. In this paper, we create and release the first ever dataset for counterspeech using comments from YouTube. The data contains 9438 manually annotated comments where the labels indicate whether a comment is a counterspeech or not. This data allows us to perform a rigorous measurement study characterizing the linguistic structure of counterspeech for the first time. This analysis results in various interesting insights such as: the counterspeech comments receive double the likes received by the non-counterspeech comments, for certain communities majority of the non-counterspeech comments tend to be hate speech, the different types of counterspeech are not all equally effective and the language choice of users posting counterspeech is largely different from those posting non-counterspeech as revealed by a detailed psycholinguistic analysis. Finally, we build a set of machine learning models that are able to automatically detect counterspeech in YouTube videos with an F1-score of 0.73.

1 Introduction

> "If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence." – Louis Brandeis

The advent of social media has brought several changes to our society. It allowed people to share their knowledge and opinions to a huge mass in a very short amount of time. While the social media sites have been very helpful, they have some unintended negative consequences as well. One such major issue is the proliferation of hate speech (Massaro, 1990). To tackle this problem, several countries have created laws against hate speech[1]. Organizations such as Facebook, Twitter, and YouTube have come together and agreed to fight hate speech as well[2].

1.1 Current protocols to combat hatespeech and their limitations

One of the main tools that these organizations use to combat online hate speech is blocking or suspending the message or the user account itself. Although, several social media sites have taken strict actions to prohibit hate speech on websites they own and operate, they have not been very effective in this enterprise[3]. At the same time, one should not block/suspend free speech because selective free speech is a dangerous precedent.

While blocking of hateful speech may reduce its impact on the society, one always has the risk of violation of free speech. Therefore, the preferred remedy to hate speech would be to add more speech (Richards and Calvert, 2000).

[1] https://goo.gl/tALXsH
[2] https://goo.gl/sH87W2
[3] https://goo.gl/G7hNtS, https://goo.gl/zEu4aX, https://goo.gl/CFmsqM

1.2 Can countering hate speech be an effective solution?

This requirement led countries and organizations to consider countering of hate speech as an alternative to blocking (Gagliardone et al., 2015). 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 hate speech. For example, the Council of Europe supports an initiative called 'No hate speech Movement'[4] with the aim to reduce the levels of acceptance of hate speech and develop online youth participation and citizenship, including in Internet governance processes. UNESCO released a study (Gagliardone et al., 2015) titled 'Countering Online Hate Speech', to help countries deal with this problem. Social platforms like Facebook have started counterspeech programs to tackle hate speech[5]. 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 (Bartlett and Krasodomski-Jones, 2015). Combating hate speech 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.

1.3 Working definition of counterspeech

In this paper, we define counterspeech as a direct response/comment (not reply to a comment) that *counters* the hateful or harmful speech. Taking the YouTube videos that contain hateful content toward three target communities: *Jews*, *African-American (Blacks)* and *LGBT*, we collect user comments to create a dataset which contains counterspeech. To annotate this dataset, we use the different classes of counterspeeh described in Benesch et. al. (2016b) with a slight modification to the 'Tone' Category. While the paper includes all kinds of tones in this category, we split this class further into two categories: 'Positive Tone' and 'Hostile Language'.

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

[5]https://counterspeech.fb.com/en/

1.4 Our contributions and observations

We annotate and release the first ever dataset[6] on counterspeech. The dataset is based on counterspeech targeted to three different communities: *Jews*, *Black*, and *LGBT*. It consists of 4111 comments annotated as counterspeech and an additional 5327 comments tagged as non-counterspeech. The counterspeech comments are further labeled into one or more of the categories listed in Table 1.

While developing the dataset, we had several interesting observations. We find that overall counterspeech comments receive more (twice) likes than non-counterspeech comments. Psycholinguistic analysis reveals striking differences between the language used by the users posting counter and non-counterspeech. We also observe that the different communities attract different proportions of counterspeech. 'Humor' as a counterspeech seems to be more prevalent when *LGBT* is the target community, while in case of the *Jews* community, 'Positive tone' of speech seems to be more widely used.

As an additional contribution, we define three classification tasks for the dataset and develop machine learning models: (a) counterspeech vs non-counterspeech classification, in which Logistic Regression performs the best with an F1-score of 0.73, (b) multi-label classification of the types of counterspeech present in a give counterspeech text, in which multi-layer perceptron performs the best, (c) cross-community classification in which SVM performs the best with an F1-score of 0.61.

2 Related work

In this section, we review some of the related literature. "Counter-speech is a common, crowd-sourced response to extremism or hateful content. Extreme posts are often met with disagreement, derision, and counter campaigns" (Bartlett and Krasodomski-Jones, 2015). Citron and Norton (2011) categorizes four ways in which one can respond to hateful messages – (i) **Inaction:** By not responding to the hate speech, we might be actually causing more harm. It sends a message that people do not care about the target community. (ii) **Deletion/Suspension:** The removal of hate speech is the most powerful option available in response to hate speech. Removal of the hateful content is sometimes accompanied by the removal

[6]https://goo.gl/uGXSEk

or suspension of the user account as well. This strategy is used by most of the social networks such as Facebook, Twitter, Quora, etc. (iii) **Education:** Institutions can help in educating the public about hate speech and its implications, consequences and how to respond. Programmes such as 'NO HATE SPEECH' movement[4] and Facebooks Counterspeech program[5] help in raising awareness, providing support and seeking creative solutions. (iv) **Counterspeech:** Counterspeech is considered as the preferred remedy to hate speech as it does not violate the normative of free speech. While government or organizations rarely take part in counterspeech, a large proportion of the counterspeech is actually generated by the online users.

Silence in response to digital hate carries significant expressive costs as well. When powerful intermediaries rebut demeaning stereotypes (like the Michelle Obama image) and invidious falsehoods (such as holocaust denial), they send a powerful message to readers. Because intermediaries often enjoy respect and a sense of legitimacy, using counterspeech, they can demonstrate what it means to treat others with respect and dignity (Citron and Norton, 2011).

While blocking might work as a counter at the individual scale, it might actually be detrimental for the community as a whole. Deletion of comments that seem hateful might affect a person's freedom of speech. Also, with blocking, it is not possible to recover from the damage that the message has already caused. Counterspeech can therefore be regarded as the most important remedy which is constitutionally preferred (Benesch, 2014).

Counterspeech has been studied on social media sites like Twitter (Wright et al., 2017; Benesch et al., 2016b), YouTube (Ernst et al., 2017) and Facebook (Schieb and Preuss, 2016). Wright et al. (2017) 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. (2017) 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. Schieb and Preuss (2016) study counterspeech on Facebook and through simulation, find that the defining factors for the success of counter speech are the proportion of the hate speech and the type of influence the counter speakers can exert on the undecided. Stroud and Cox (2018) perform case studies on feminist counterspeech. Another line of research considers ascertaining the success of the counterspeech. Benesch et al. (2016a) describes strategies that have favorable impact or are counterproductive on users who tweet hateful or inflammatory content.

Most of these studies are not conducted at large scale. Computational approaches are required in order to study and engage counterspeech efforts at scale, automatic detection being the key component (Wright et al., 2017). This paper is the first to release a considerably large annotated counterspeech data and propose an automatic counterspeech detection model.

3 Dataset

YouTube is one of the key online platforms on the Internet with 1.5 billion logged-in users visiting the site every month[7]. Many of these videos contain hate speech targeted toward various communities. In this paper, we focus on such hateful videos and scrape their comment section.

3.1 Data collection from YouTube

In order to gather a diverse dataset, we focus on three target communities: *Jews*, *Blacks*, and *LGBT*. First, we manually select videos[6] that contain some act of hate against one of these communities. Next, we use the YouTube comment scraper[8] to collect all the comments from the selected videos. Each comment had fields such as the comment text, username, date, number of likes, etc.

3.2 Dataset annotation

There are different types of counterspeech that have different effects on the user. In order to understand the differences between them, we annotate the dataset at two levels.

First level annotation: In the first level, we select comments from the hate speech video and ask the annotators to annotate each of these comments as a counter/non-counter to the hate message/action in the video. We define a comment as counterspeech if it opposes the hatred expressed in the video. We only consider those comments which

[7] http://goo.gl/eEqWAt

[8] http://ytcomments.klostermann.ca/

are direct response to the video and ignore all the replies to these comments as we observe that they usually tend to drift off-topic and the discussion becomes more personal and noisy. Each comment has been annotated by two users and the conflicting cases have been resolved by a third annotator. We achieve 95.9% agreement between the two annotators with a Cohen's κ of 0.917. As a result of this step, we arrive at 4111 counterspeech comments and 5327 non-counterspeech comments. To our surprise, we find that 43.5% of the direct responses to the selected hate videos are counterspeech.

Second level annotation: In order to obtain a deeper understanding of the types of counterspeech, we perform a second level annotation. We give the annotators a counterspeech text and ask them to label all the types of counterspeech that are present in it. We use the taxonomy of counterspeech described in Benesch et al. (2016b) for this purpose. For ease of readability we describe these categories in the subsequent section.

Two independent annotators tagged each comment annotated as counterspeech in the first level into appropriate types. We obtain a loose κ score of 0.956 and a strict κ score of 0.872 for this task (Ravenscroft et al., 2016). We employ a third annotator for deciding on the conflicting cases. The final distribution of the different types of counterspeech is noted in Table 1.

3.3 Types of counterspeech

There are numerous strategies that could be used to counter the hateful messages in the online social media. Benesch et al. (2016b) distinguishes eight such strategies that are used by counterspeakers. We decided on using these eight types of counterspeech with a slight modification to the category 'Tone'. While the authors define multiple sub-categories, we only use 'Positive Tone' and 'Hostile' as categories for the dataset. A single comment can consist of multiple types of counterspeech as shown in Figure 1. Below, we discuss these various categories.

Presenting facts to correct misstatements or mis-perceptions: In this strategy, the counterspeaker tries to persuade by correcting misstatements. An example of this type of counterspeech toward the *LGBT* community from our dataset is as follows: "*Actually homosexuality is natural. Nearly all known species of animal have their gay communities. Whether it be a lion or a whale, they have or had (if they are endangered) a gay community. Also marriage is an unnatural act. Although there are some species that do have longer relationships with a partner most known do not.*"

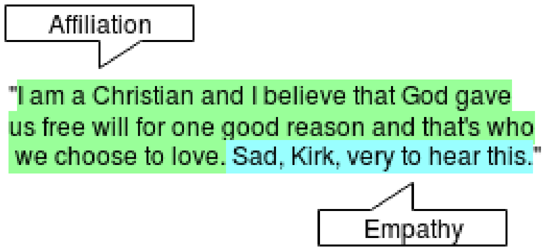


Figure 1: An example comment containing two types of counterspeech: *affiliation* and *empathy*.

Pointing out hypocrisy or contradictions: In this strategy, the counterspeaker points out the hypocrisy or contradiction in the user's (hate) statements. In order to discredit the accusation, the individual may explain and rationalize their previous behavior, or if they are persuadable, resolve to avoid the dissonant behavior in the future (Beauvois et al., 1993). An example of this type of counterspeech toward *LGBT* community from our dataset is as follows: "*You don't love homosexuals. You hate them. To say you want homosexuals to go to hell is pure hate. Whether they are worthy of death or not doesn't even matter because once you say you "hate" someone you automatically contradict Jesus Christ Himself.* "

Warning of offline or online consequences: In this strategy, counterspeaker warns the user of possible consequences of his/her actions. This can sometimes cause the original speaker of the hatespeech to retract from his/her original opinion. An example of this type of counterspeech toward the *African-American* community from our dataset is as follows: "*I strongly suggest a local lawyer take this case pro bono and get the cops fired or at least disciplined and the man receive a public apology from the mayor, or, the city be sued too.* "

Affiliation: Affiliation is "...establishing, maintaining, or restoring a positive affective relationship with another person or group of persons" (Byrne, 1961). People are more likely to credit the counterspeech of those with whom they affiliate, since they tend to "evaluate ingroup members as more trustworthy, honest, loyal, cooperative, and valuable to the group than outgroup members" (Kane et al., 2005). In our dataset, couterspeakers who use *Affiliation* receive the highest number of likes for their comments among all the counterspeech types. An example of this type of counterspeech toward the *African-American* community from our dataset is as follows: "*As a black man I feel for this guy because I think he truly believed he would get fairly by the police if he spoke in an eloquent nonthreatening manner. My brother this is America shit don't go that way.* "

Denouncing hateful or dangerous speech: In this strategy, the counterspeakers denounce the message as being hateful. This strategy can help the counterspeakers in reducing the impact of the hate message. An example of this type of counterspeech toward *LGBT* community from our dataset is as follows: "*This is not Christianity. This is a man preaching hate, bullying, and violence, and saying it's what God wants. Not my God.* "

Humor and sarcasm: Humor is one of the most powerful tools used by the counterspeakers to combat hate speech. It can de-escalate conflicts and can be used to garner more attention toward the topic. Humor in online settings also eases hostility, offers support to other online speakers, and encourages social cohesion (Marone, 2015). Often, the humor is sarcastic, like the following counterspeech comment subscribing the *LGBT* community from our dataset: "*Yeah because using force to change someone's sexuality or thought process is going to work just fine and they will start to love you. Can't see this plan going wrong at all. What a genius!*"

Positive Tone: The counterspeaker uses a wide variety of tones to respond to hate speech. In this strategy, we consider different forms of speech such as empathic, kind, polite, or civil. Increasing empathy with members of opposite groups counteracts incitement (Benesch, 2014). We would like to point out that the original authors actually defined Tone to contain hostile counterspeech as well. Instead, we decide to make 'Hostile Language' as a separate type of counterspeech. An example of this type of counterspeech toward *African-American* community from our dataset is as follows: "*it was actually sad to watch i cried because the man wanted to get his child. breaks my heart honestly and when he said help me i was just devastated i would sue these cops*"

Hostile language: In this strategy, the counterspeaker uses abusive, hostile, or obscene comments in response to the original hate message. Such a response can persuade an original speaker to delete his message or even a whole account, but is unlikely to either de-escalate the conversation or persuade the original speaker to recant or apologize. An example of this type of counterspeech toward *Jews* from our dataset is as follows: "*do you muslims want an award from jews Christians and white people for not having a problem with jews? its called basic human decency you f*****g filthy c***s how about you get out of europe and stop ruining germany*"

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	202	34	255	491
Pointing out hypocrisy or contradictions	147	100	245	492
Warning of offline or online consequences	77	50	99	226
Affiliation	166	66	110	342
Denouncing hateful or dangerous speech	253	212	221	686
Humor	141	115	355	611
Positive Tone	287	121	114	522
Hostile	531	510	635	1676
Total	1804	1208	2034	5046

Table 1: Statistics of the counterspeech dataset. Numbers corresponding to each of the target community, grouped as per the type of counterspeech are shown.

4 Detailed analysis

In this section, we perform a detailed analysis over the dataset. We observe that 78.2% of the counterspeech comments belong to exactly one counterspeech category. Thus, majority of the counterspeakers rely on a single strategy for counterspeech. As noted in Table 1, different communities attract different types of counterspeech. We observe that 'Hostile Language' is the major category for all the classes. Other than that, the counterspeakers for the *Jews* community seem to be using 'Positive Tone' strategy in their counterspeech more often, while the counterspeakers of the *LGBT* community more often use 'Humor' to tackle the hatespeech.

4.1 Likes and comments

We first analyze the comments as per the likes and comments received. We consider two groups - counterspeech comments and non-counterspeech comments. For our analysis, we also perform MannWhitney U test (Mann and Whitney, 1947) to compare the two distributions.

On average, we find counterspeech comments in our data receiving 3.26 likes, in contrast to non-counterspeech comments receiving 1.68 likes, which is almost half of counterspeech ($p \sim$

0.0). Similarly, we investigate into the number of replies received and find that counterspeech comments received more replies (average: 2.58) than non-counterspeech comments (average: 1.65). However, the differences were not as significant ($p > 0.1$).

We further look into the comments and likes received by each of the community for each type of counterspeech. Figures 2a and 2d show the average number of likes and replies received by the *Blacks* community, respectively. Here, we observe that the comments that talk about online/offline consequences get more likes. Also, strategies such as 'Consequences' and 'Humor' seem to garner more replies than other strategies. For the *Jews* community, Figures 2b and 2e plots the average number of likes and replies received, respectively. 'Affiliation' here receives more likes than other types. 'Affiliation' and 'Contradiction' receive the majority of replies. Figures 2c and 2f plot the average number of likes and replies received by the *LGBT* community, respectively. 'Humor' and 'Contradictions' receives more likes than other types. 'Contradiction', 'Humor', and 'Positive Tone' receives more replies than other types.

4.2 Psycholinguistic analysis

The language that online users choose, provide important psychological cues to their thought processes, emotional states, intentions, and motivations (Tausczik and Pennebaker, 2010). The LIWC tool[9] helps in understanding several psycholinguistic properties using the text. In order to understand the psycholinguistic differences, we apply LIWC (i.e., the fraction of words in different linguistic and cognitive dimensions identified by the LIWC tool) on both counter and non-counter comments. Finally, we look for statistically significant differences between these two groups with respect to the above analysis. We run MannWhitney U test (Mann and Whitney, 1947) and report the significantly different categories in Table 2.

We observe several LIWC categories that show significant differences between counter and non-counter comments. The 'spoken' category of LIWC ('assent' and 'non-fluencies') is more pronounced in non-counterspeech, whereas 'affective processes' ('anxiety', 'anger', 'sadness', 'negative emotion' and 'affect') are more strong in counterspeech. 'Personal concern' ('religion' and 'leisure') is more pronounced in non-counter comments. The 'biological processes' ('body', 'health', 'sexual'), on the other hand, seems to be more dominant in the language of the counterspeakers.

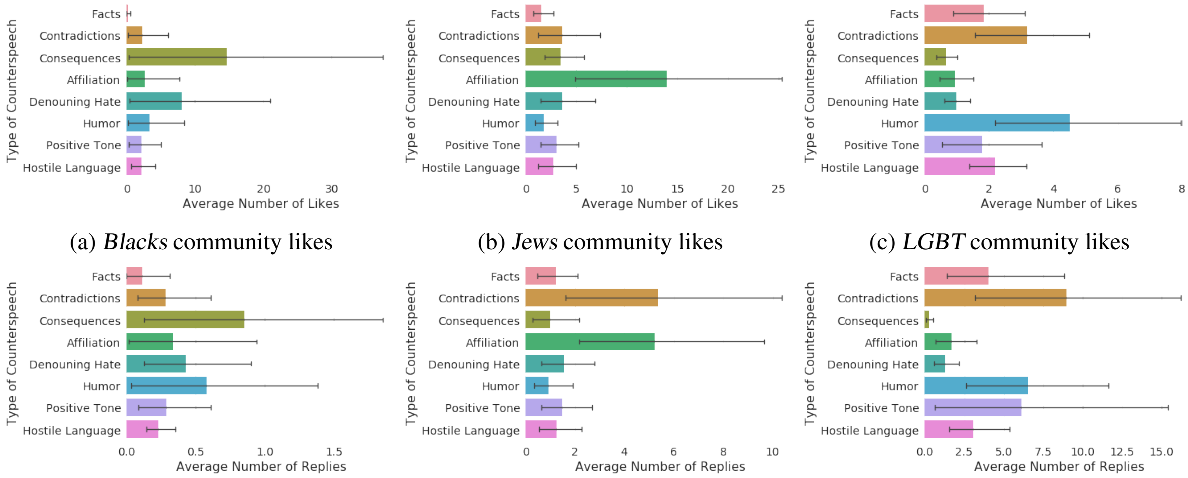
[9] http://liwc.wpengine.com/

Dimension	Category	Counter (mean)	Non-counter (mean)
Personal concerns	Leisure***	0.184	**0.298**
	Relig***	1.341	**1.615**
Spoken categories	Assent***	0.197	**0.275**
	Nonflu***	0.034	**0.051**
Biological processes	Body***	**0.314**	0.248
	Health***	**0.19**	0.154
	Sexual***	**0.525**	0.439
Perceptual processes	Hear***	0.243	**0.305**
Cognitive processes	Insight***	0.706	**0.867**
	Discrep**	**0.713**	0.699
	Certain***	0.569	**0.674**
Affective processes	Anx***	**0.141**	0.106
	Negemo***	**1.546**	1.31
	Posemo***	1.164	**1.306**
	Affect***	**2.699**	2.596
	Anger***	**0.929**	0.752
	Sad**	**0.191**	0.178
Social processes	Humans***	**0.758**	0.648
Linguistic processes	Funct*	17.136	**18.479**
	Swear***	**0.3**	0.184
	I***	**0.645**	0.545
	Article*	2.077	**2.475**
	Ipron***	**2.089**	2.083
	Negate*	0.811	**0.881**
	Past***	0.745	**1.011**
	Present***	**3.415**	3.346
	Pronoun**	4.24	**4.303**
	They***	0.43	**0.59**
	Verbs*	4.884	**5.092**
	You*	0.498	**0.541**
	SheHe***	**0.62**	0.594

Table 2: LIWC analysis of the counter and non-counter comments. Only those LIWC categories are shown which are statistically significant: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). Note that each LIWC category is either dense in green cells (red cells) for the counter (non-counter) comments or for the non-counter (counter) comments.

5 Classification model

We consider three classification tasks that naturally manifest in this problem context. The first task is a binary classification problem in which we present the system with a comment and the task is to predict whether the comment is a counterspeech or non-counterspeech. The second one is a multi-label classification task in which we present the system with a known counterspeech comment and the task is to predict all the types of counterspeech present in the comment. The third task is similar to first, except that it is cross-community, i.e., while the training data is drawn from two of the three communities, the test data is drawn from the remaining community.



(a) *Blacks* community likes (b) *Jews* community likes (c) *LGBT* community likes

(d) *Blacks* community replies (e) *Jews* community replies (f) *LGBT* community replies

Figure 2: Plots showing average number of likes and replies received by different types of counterspeech in the three communities.

Preprocessing: Before the classification, we preprocess all the data by eliminating URLs, numerals and punctuations. The text is then lower cased, tokenized and used as input for the classification pipeline. We observe that the stop words provide important cues to the classifier and therefore retain them through the preprocessing stage.

Training, validation and test split: The data is split into a training, validation and test set, where the validation and test set are randomly sampled while respecting their overall type distribution in the dataset. As there are more non-counter comments in the training data, we randomly select a subset of these to make the dataset balanced.

Features: For the task of classification we use *tf-idf* vectors and *bag of words* vectors (BoWV). The BoWV approach uses the average of the GloVe (Pennington et al.) word embeddings to represent a sentence. We set the size of the vector embeddings to 200.

Choice of classifiers: We experiment with multiple classifiers such as Random Forest (RF), Logistic Regression (LR), SVMs, XGBoost (XGB), Decision Tree (DT), and neural models such as Multi-layer Perceptron(MLP), fastText, LSTM, CNN.

5.1 Counterspeech classification

In this task, a binary classifier is built to predict if the given input text is a counterspeech or non-counterspeech. We consider balanced test data and report results in Table 3. As the table shows, logistic regression with TF-IDF performs the best. Classifier such as fastText with random embeddings also perform well.

Model	Precision	Recall	F1-Score
Random Embedding + LSTM	0.64	0.64	0.64
BoWV + SVM	0.65	0.65	0.65
Random Embedding + CNN	0.66	0.66	0.66
TF-IDF + XGBoost	0.67	0.67	0.67
BoWV + RF	0.67	0.67	0.67
Random Embedding + Fasttext	0.70	0.70	0.70
TF-IDF + LR	0.73	0.73	0.73

Table 3: Classification scores for the task of predicting if the given comment is counterspeech or non-counterspeech.

5.2 Counterspeech type classification

Here, we build models for a multi-label classification task in which the input to the classifier is a counter comment and the output are the types of counterspeech present in the comment. Table 4 shows the results of this task. For evaluation purpose, we report the strict metric *exact match ratio* (Sorower, 2010). In addition, to account for partial correctness, we report the variants of accuracy, precision, recall and F1-score proposed by Godbole and Sarawagi (2004). As the table shows, MLP gives the best performance in all the metrics.

Model	Exact match ratio	Accuracy	Precision	Recall	F1-score
Random Forest	0.17	0.19	0.21	0.19	0.20
Decision Tree	0.21	0.30	0.33	0.36	0.33
MLP	0.32	0.39	0.42	0.42	0.41

Table 4: Classification scores for the task of multi-label classification of the types of counterspeech.

5.3 Cross-community classification

In this section, we build models that draw the training data points from two communities to predict the labels for the test data drawn from the third community. Note that this application is motivated by the fact that in the context of the current problem there might exist communities for which in-community training instances are scarce and therefore the only way to perform the classification is to resort to the training instances available for other communities (see (Rudra et al., 2015) for a similar approach). For evaluation, we report weighted precision, recall and F1-Score. Table 5 shows the results of this task. The models are able to produce comparable results even while they are trained using instances from a different community. This is an extremely desirable feature to avoid requirement of fresh annotations every time the model is used for a new (and so far unseen) community.

Train instances	Test instances	Model	Precision	Recall	F1-Score
Blacks + Jews	LGBT	XGB	0.65	0.58	0.58
Jews + LGBT	Blacks	SVM	0.62	0.62	0.61
LGBT + Blacks	Jews	LR	0.62	0.58	0.58

Table 5: Classification scores for the task of predicting if the given comment is counterspeech or non-counterspeech in one community using the training instances from the other two communities.

6 Discussion

We observe that the non-counterspeech consist mainly of comments that agree with the main content in the video or hatespeech toward the target community itself. These vary depending on the community involved. In case of *Jews*, we find that majority of the comments claimed that the Jews are controlling the economy and are responsible for the destruction of their society. Many of the non-counterspeech also included holocaust denial (Gerstenfeld et al., 2003). In case of *Blacks*, we find that the majority of non-counterspeech were hatespeech in the form of racist remarks such as ni**ers, slavery etc. In case of *LGBT*, we observe that the majority of non-counterspeech are linked to religious groups claiming that it is unnatural and forbidden in their religion.

Not all types of counterspeech are equally effective (Benesch et al., 2016a). To understand the nature of replies received by each type of counterspeech comments, we randomly select some comments corresponding to each type and analyze the nature of the responses received. This would tell us how the community views these statements provided by the counterspeakers. In specific, we check if the reply is agreeing with the statement provided by the counterspeaker. We observe that strategies like ‘Warning of Consequences’, ‘Denouncing Hate’ and ‘Humor’ receive more acceptance in the replies than other strategies. Strategies like ‘Pointing out Hypocrisy’ and ‘Hostile Language’ receive least acceptance from the community. Although, using ‘Hostile Language’ seems to be very prevalent (see Table 1), we observe that this strategy is not welcomed by even the target community in whose favor these are posted. In many instances, the target community users tend to oppose this form of counterspeech and request the counterspeakers to refrain from using such language of hate.

7 Conclusion and future works

The proliferation of hateful content in online social media is a growing concern. Currently used methods such as blocking or suspension of messages/accounts cause problems to the freedom of speech. Counterspeech is emerging as a very promising option backed by several organizations and NGOs. With no dataset and model available for counterspeech detection, no large scale study can be conducted. In this paper, we took the first step toward creating a dataset of counterspeech against hateful videos in YouTube. We found that counter comments receive more likes than non-counter comments. Further, the psycholinguistic analysis of the comments reveal striking differences between the language choice of counter and non-counter speakers. We found that different communities seem to have different preferences for the selection of counterspeech type. Our models and dataset are placed in the public domain.

There are several directions, which can be taken up as future research. One immediate step is to develop automatic counterspeech detection models for other social media sites like Facebook and Twitter. Another direction could be to study the effectiveness of different types of counterspeech for different communities. A connected research objective could be to investigate how effective the counterspeakers are in changing the mindset of the hate users.

References

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.

Jean-Leon Beauvois, Robert-Vincent Joule, and Fabien Brunetti. 1993. Cognitive rationalization and act rationalization in an escalation of commitment. *Basic and Applied Social Psychology*, 14(1):1–17.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: United States Holocaust Memorial Museum*.

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

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

D Byrne. 1961. Anxiety and the experimental arousal of affiliation need. *The Journal of Abnormal and Social Psychology*, 63(3):660.

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

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*, pages 1–49.

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

Phyllis B Gerstenfeld, Diana R Grant, and Chau-Pu Chiang. 2003. Hate online: A content analysis of extremist internet sites. *Analyses of social issues and public policy*, 3(1):29–44.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *PAKDD*, pages 22–30.

Aimée A Kane, Linda Argote, and John M Levine. 2005. Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational behavior and human decision processes*, 96(1):56–71.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Vittorio Marone. 2015. Online humour as a community-building cushioning glue. *The European Journal of Humour Research*, 3(1):61–83.

Toni M Massaro. 1990. Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.*, 32:211.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

James Ravenscroft, Anika Oellrich, Shyamasree Saha, and Maria Liakata. 2016. Multi-label annotation in scientific articles-the multi-label cancer risk assessment corpus. In *LREC*.

Robert D Richards and Clay Calvert. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, page 553.

Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *CIKM*, pages 583–592.

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

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

Scott R Stroud and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, pages 293–310. Springer.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

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*, pages 57–62.