Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Thou Shalt Not Hate: Countering Online Hate Speech

Binny Mathew, Punyajoy Saha,[1] Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity,[2] Pawan Goyal, Animesh Mukherjee
Indian Institute of Technology(IIT), Kharagpur
[1]Indian Institute of Engineering Science and Technology (IIEST), Shibpur
[2]Kellogg School of Management, Northwestern University
binnymathew@iitkgp.ac.in, {punyajoysaha1998, hardik.tharad, subham.rajgaria, prajwal1210}@gmail.com,
suman.maity@kellogg.northwestern.edu, {pawang, 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 the hateful content, they have mostly been unsuccessful. Counterspeech is seen as an effective way of tackling the online hate without any harm to the 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 13,924 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 much more likes as compared to 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 0.71. We also build multilabel models that can detect different types of counterspeech in a comment with an F1-score of 0.60.

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].

Current protocols to combat hate speech 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, some have argued that 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).

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

[1]Hate speech Laws: https://goo.gl/tALXsH

[2]https://goo.gl/sH87W2

[3]https://goo.gl/G7hNtS, https://goo.gl/zEu4aX, https://goo.gl/CFmsqM

[4]No hate speech Movement Campaign: http://www.nohatespeechmovement.org/

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.

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 counterspeech 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'. Note that when we say that a comment is a counterspeech to a video, we are focusing on the person about whom the video is about. The video may contain other people as well (such as an interviewer).

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*, *Blacks*, and *LGBT*. It consists of 6,898 comments annotated as counterspeech and an additional 7,026 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 much more 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 XGBoost performs the best with an F1-score of 0.71, (b) multi-label classification of the types of counterspeech present in a given counterspeech text, in which XGBoost performs the best, (c) cross-community classification with an F1-score in the range 0.62 - 0.65. With these tasks and analysis, we hope that our research can help in reducing the spread of hate speech online.

[5] Counterspeech Campaign by Facebook: https://counterspeech.fb.com/en/

[6] The dataset and models are available here: https://github.com/binny-mathew/Countering_Hate_Speech_ICWSM2019

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/conflicts, derision, counter campaigns" (Bartlett and Krasodomski-Jones 2015; Maity et al. 2018). 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 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 counterspeech 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. Munger (2017) conducted an experiment to examine the impact of group norm promotion and social sanctioning on racist online harassment and found that subjects who were sanctioned by a high-follower in-group male significantly reduced their use of a racist slur.

Our work is different from the existing literature in several ways. As noted in (Benesch et al. 2016b), the literature for counterspeech is pretty less. The existing literature on counterspeech have been qualitative and anecdotal in nature, while ours is the first work which tries to study counterspeech empirically. (Wright et al. 2017) noted that "Computational approaches are required in order to study and engage counterspeech efforts at scale and there is no work which perform automatic detection of counterspeech". In this work we attempt for the first time various learning algorithms to detect counterspeech. Further, one of our main contributions is that we release the first ever dataset for counterspeech identification. Our paper not only does the classification for the counterspeech task; we take it a step forward and do multi-label classification for various types of counterspeech.

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.

Data collection from YouTube

In order to gather a diverse dataset, we focus on three target communities: *Jews*, *Blacks*, and *LGBT*. The first step in our work involved finding videos that contained hateful content. In order to find such videos, we searched YouTube for phrases such as 'I hate Jews', 'I hate blacks' etc. We then 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.

Dataset annotation

Annotations were performed by a group of two PhD students working in the area of social computing and three undergraduate students in computer science with ages ranging between 21-30.

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

[8]YouTube Comment Scrapper: http://ytcomments.klostermann.ca/

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 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 90.23% agreement between the two annotators with a Cohen's κ of 0.804. As a result of this step, we arrive at 6,898 counterspeech comments and 7,026 non-counterspeech comments. To our surprise, we find that 49.5% 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 0.868 and a strict κ score of 0.743 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.

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'. The authors have considered the whole spectrum of Tone as a single category. While one end of the spectrum ('Hostile') can cause the original speaker to delete his post/account and thus is unlikely to de-escalate the conversation, the other end of the spectrum ('Positive tone') could help in generating a positive attitude and thus de-escalate the conversation. So, we decided to divide the 'Tone' category into 'Positive tone' and 'Hostile language' categories for our work. Note that a single comment can consist of multiple types of counterspeech as shown in Figure 1. Also, the types of countespeech strategies discussed here are not comprehensive; there could be other types as well. In this paper, we focus on just these eight types of counterspeech. Below, we discuss these various categories.

Presenting facts to correct misstatements or misperceptions: In this strategy, the counterspeaker tries to persuade by correcting misstatements. An example of this type

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

Table 1: Statistics of the counterspeech dataset. Numbers corresponding to each of the target community, grouped as per the type of counterspeech are shown. Note that if a comment utilizes multiple strategies, we would include that particular comment in all the corresponding counterspeech types. Thus, we have a total of 9,119 counterspeech from 6,898 comments.

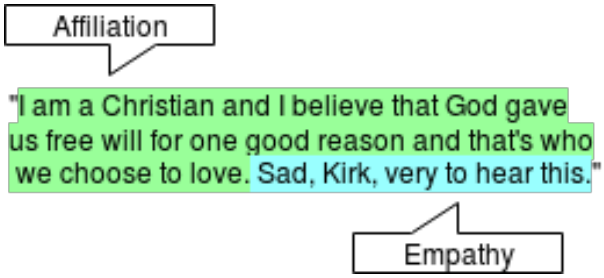


Figure 1: An example comment containing two types of counterspeech: *affiliation* and *empathy*. This comment was in response to a interview video in which the interviewee (Kirk) says that homosexuality is unnatural, detrimental and destructive to the society.

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*". This comment was in response to a interview video in which the interviewee says that homosexuality is unnatural, detrimental and destructive to the society.

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, Joule, and Brunetti 1993). An example of this type of counterspeech toward *LGBT* community from our dataset is as follows: "*The 'US Pastor' can't accept gays because the Bible says not to be gay. But...he ignores:The thing about eating shrimp or pork, The thing about touching the skin of a dead pig (Football), The thing about mixing fabrics, The thing about torn clothes, The thing about tattoos, The thing about planting two crops in one field, The thing about working on the Holy Day (Saturday or Sunday depending)...for any and all of those sins one should burn for an eternity, yet is ignored. But when it comes to loving the wrong person (gays) this will not do! Christians only follow the parts of the bible that supports their bigotry. YOUR A HYPOCRITE.* ". This comment was in response to a interview video in which the interviewee encourages hate against homosexual people.

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 hate speech to retract from his/her original opinion. An example of this type of counterspeech toward the *LGBT* community from our dataset is as follows: "*I'm not gay but nevertheless, whether You are beating up someone gay or straight, it is still an assault and by all means, this preacher should be arrested for sexual harassment and instigating!!!* ". This comment was in response to a video in which a preacher advised people to beat the kids if they are gay.

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, Argote, and Levine 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 *LGBT* community from our dataset is as follows: "*Hey I'm Christian and I'm gay and this guy is so wrong. Stop the justification and start the accepting. I know who my heart and soul belong to and that's with God: creator of heaven and earth. We all live in his plane of consciousness so it's time we started accepting one another. That's all* ". This comment was in response to a interview video in which the interviewee encourages hate against homosexual people.

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 *Jews* community from our dataset is as follows: "*please take this down YouTube. this is hate speech.* ". This comment was in response to a video in which a preacher is advocating hatred and killing of Jewish people.

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: "*HAHAHAHAHAHAHAH...oh you were serious. That's even funnier :)*". This comment was in response to a video in which a preacher advocates hate and killing of homosexual people.

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 *Jews* community from our dataset is as follows: "*I am a Christian, and I believe we're to love everyone!! No matter age, race, religion, sex, size, disorder...whatever!! I LOVE PEOPLE!! We are not going to go anywhere as a country if we don't put God first in our lives, and treat EVERYONE with respect*". This comment was in response to a video in which a preacher is advocating hatred and killing of Jewish people.

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 *African-American* from our dataset is as follows: "*This is ridiculous!!!!!! I hate racist people!!!! Those police are a**holes!!!*". This comment was in response to a video in which the police are performing a hate crime against.

Detailed analysis

In this section, we perform a detailed analysis over the dataset. We observe that 71.24% 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 among all the classes and is present in around 39.74% of the counterspeech. 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' and 'Pointing out hypocrisy or contradiction' to tackle the hate speech.

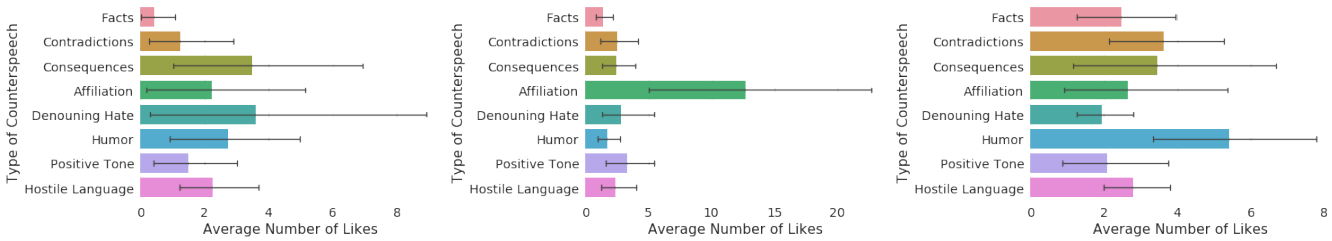
Likes and comments

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 Mann–Whitney U test (Mann and Whitney 1947) to compare the two distributions.

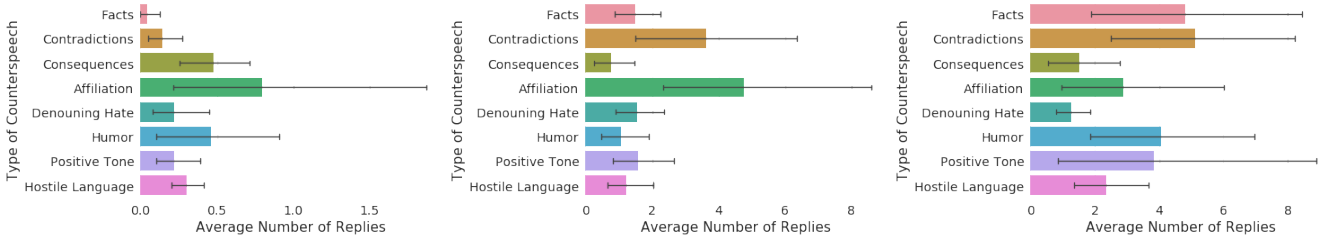
On average, we find counterspeech comments in our data receiving 3.0 likes, in contrast to non-counterspeech comments receiving 1.73 likes, which is very less as compared to the counterspeech ($p \sim 0.0$). Similarly, we investigate into the number of replies received and find that counterspeech comments receive more replies (average: 1.94) than non-counterspeech comments (average: 1.50). However, the differences were not as significant ($p > 0.1$).

We would also like to point out that the average likes and replies in YouTube videos are generally less. In (Thelwall, Sud, and Vis 2012), the authors analyze large samples of text comments in YouTube videos. They found that 23.4% of YouTube comments in complete comment sets are replies (called reply-density). The authors also analyzed the types of videos receiving the least and the most replies and found that videos pertaining to news, politics, and religion are at the top. This point is exemplified by our dataset where we found the reply density to be 45.17%, which is almost double of a normal YouTube video. If we consider just the counterspeech comments, the reply density is even higher at 72.07%. Due to the controversial nature of the content, it is expected to generate such discussion. We can thus state that our dataset is representative enough. In another work by (Siersdorfer et al. 2014), the authors analyzed over 6 million comments from 67,290 videos and found the average rating to be 0.61. This is much less than the average 3.0 likes for our counterspeech comments.

We view the likes received by the counterspeech comments as an endorsement by the community. In this sense, we can observe from Figure 2 that different communities seem to like different types of counterspeech. In case of the *African-American* community, the average likes received by the counterspeech category which 'Warn of offline/online consequences' and 'Denouncing of hateful/dangerous speech' seem to be more as compared to the other types. In these counterspeech comments, the counterspeakers call out for racism and talk about how the person in focus could be sued for his actions. In case of the *Jews* community, 'Affiliation' seems to be the most endorsed form of counterspeech. In the comments, we observe that the people affiliate with both the target and the source community ('Muslims', 'Christians') to counter the hate message. In some of the comments, the counterspeakers identify themselves as belonging to the same community as that of the hate speaker and claim that the hate message is unacceptable. Previous works have shown that these kinds of counterspeech are successful in changing the attitude of the hate speaker (Berger and Strathearn 2013; Munger 2017). In case of the *LGBT* community, we can observe that the community endorses several types of counterspeech with the 'Humor' and 'Pointing out hypocrisy or contradiction' receiving more average likes than others. In these comments, the counterspeakers make use of sarcasm



(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. The whiskers represent 95% confidence interval.

(a) Presenting facts (b) Pointing out hypocrisy (c) Warning of consequences (d) Affiliation

(e) Denouncing speech (f) Humor and sarcasm (g) Positive tone (h) Hostile language

Figure 3: Word clouds for the different types of counterspeech used by the counterspeaker for hate speech against **Blacks**.

and provide several points which contradict the statements expressed by the hate speaker.

Lexical analysis

To understand the language used in each type of counterspeech, we look into the words that are employed by the counterspeakers. We observe that some of the words such as *Islam, Jew, People, Muslim, Black* were present in all the categories with high frequency. In order to filter out such words that are common in all categories, we use the tf-idf method. First, we generate a tf-idf matrix using the whole corpus. Then, for each type of counterspeech, we use the matrix to rank the words according to their tf-idf values. We display the top 200 words present in each type of counterspeech. The size of a word is proportional to its tf-idf value. We observe from the figures (summarized below) that the words present are representative of the categories that they belong to.

In Figure 3, we plot the word cloud for the different types of counterspeech employed for the African-American community. Observe the presence of words such as 'Illogical' in Figure 3b, 'Sue', 'Lawyer' in Figure 3c, 'Im', 'Friend' in Figure 3d, 'Race', 'Racist' in Figure 3e, 'Oh', 'Funny' in Figure 3f, 'Sorry', 'Feel', 'Heart' in Figure 3g.

In Figure 4, we plot the word cloud for the different types of counterspeech employed for the LGBT community. Observe the presence of words such as 'Hypocrite' in Figure 4b, 'Arrest', 'Die', 'Meet' in Figure 4c, 'Im', 'Christian', 'Support' in Figure 4d, 'Hatred', 'Bigot' in Figure 4e, 'Lmao', 'Funny' in Figure 4f, 'Happiness', 'Love' in Figure 4g.

In Figure 5, we plot the word cloud for the different types of counterspeech employed for the Jews community. Observe the presence of words such as 'Paradoxical' in Figure 5b, 'Deporting', 'Enforcement', 'Jail', and 'Fbi' in Figure 5c, 'Im', 'Ur', 'Love' in Figure 5d, 'Hatred', 'Racist' in Figure 5e, 'Lol', 'Funny' in Figure 5f, 'Bless', 'Love' in Figure 5g.

Psycholinguistic analysis

The language that online users choose, provides 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 used in 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 Mann–Whitney 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', 'achievement', 'work', 'leisure', and 'money') 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.

Dimension	Category	Counter (mean)	Non-counter (mean)	Significance Level
Personal concerns	Achiev	0.334	**0.383**	*
	Work	0.316	**0.397**	**
	Leisure	0.179	**0.251**	*
	Home	**0.057**	0.046	***
	Money	0.123	**0.152**	**
	Relig	1.148	**1.362**	***
Spoken categories	Assent	0.080	**0.095**	**
	Nonflu	0.021	**0.031**	***
	Filler	**0.162**	0.136	***
Biological processes	Body	**0.263**	0.175	***
	Health	**0.131**	0.108	***
	Sexual	**0.461**	0.357	***
Perceptual processes	See	0.382	**0.391**	*
	Hear	0.259	**0.306**	***
	Feel	**0.084**	0.078	***
Cognitive processes	Insight	**0.660**	0.586	***
	Discrep	**0.626**	0.586	***
	Certain	0.551	**0.655**	***
	Incl	1.400	**1.417**	*
	Excl	0.976	**1.067**	**
Affective processes	Anx	**0.121**	0.086	***
	Negemo	**1.429**	1.089	***
	Posemo	1.066	**1.149**	***
	Affect	**2.488**	2.217	***
	Anger	**0.937**	0.654	***
	Sad	**0.093**	0.079	**
Social processes	Humans	**0.759**	0.621	***
	Family	**0.113**	0.105	***
	Friends	**0.042**	0.033	***
Linguistic processes	Funct	17.013	**17.811**	***
	Swear	**0.353**	0.164	***
	I	**0.658**	0.543	***
	Ipron	**2.006**	1.951	***
	Negate	0.779	**0.859**	***
	Past	0.746	**0.941**	***
	Present	**3.389**	3.301	***
	Pronoun	4.281	**4.161**	***
	They	0.441	**0.566**	***
	Verbs	4.888	**4.957**	***
	You	0.517	**0.541**	*
	SheHe	**0.683**	0.578	***

Table 2: LIWC analysis of the counter and non-counter comments. Only those LIWC categories are shown which are statistically significant: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

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.

Preprocessing: Before the classification, we preprocess all the data by eliminating URLs, numerals, stopwords and punctuations[10]. The text is then lower cased, tokenized and used as input for the classification pipeline.

Features: For the task of classification we use *tf-idf* vectors (TF-IDF), *bag of words* vectors (BoWV) and *sentence* vectors (SV). The BoWV approach uses the average of the GloVe (Pennington, Socher, and Manning 2014) word embeddings to represent a sentence. We set the size of the vector embeddings to 300. The sentence vector is generated using a Universal Sentence Encoder (Cer et al. 2018) which outputs a 512 dimensional vector representation of the text. Recent works (Conneau et al. 2017) have shown better performance using pre-trained sentence level embeddings as compared to word level embeddings.

Choice of classifiers: We experiment with multiple classifiers such as Gaussian Naive Bayes (GNB), Random Forest (RF), Logistic Regression (LR), SVMs, XGBoost (XGB), CatBoost (CB) (Dorogush, Ershov, and Gulin 2017), Decision Tree (DT), and neural models such as Multi-layer Perceptron (MLP), LSTM.

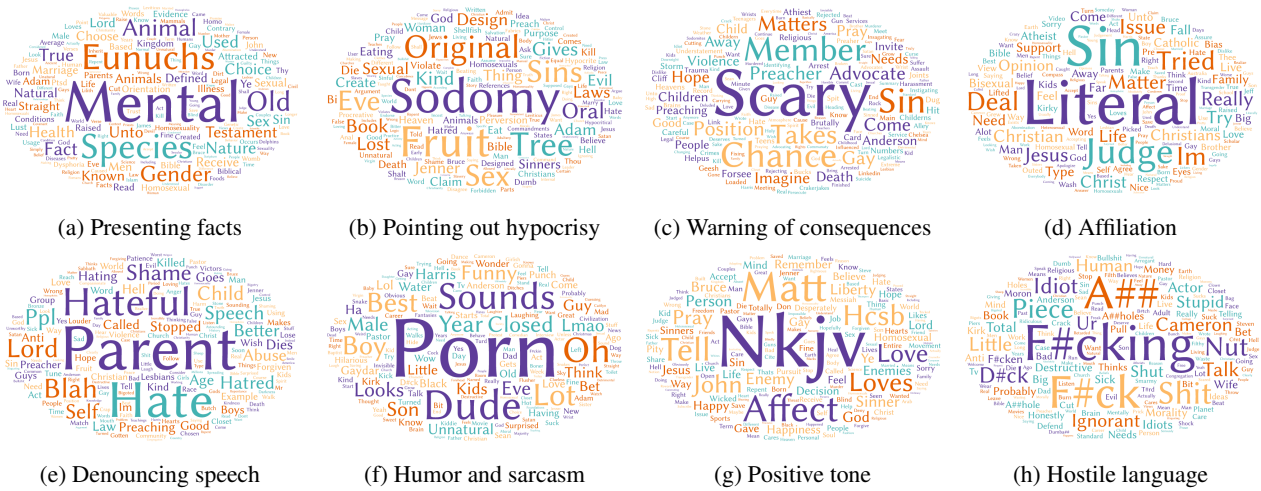
Counterspeech classification

In this task, a binary classifier is built to predict if the given input text is a counterspeech or non-counterspeech. We perform stratified 10-fold cross validation on the dataset. The whole training set is partitioned into 10 folds, one is set apart for testing, and the remaining nine are used to train the model and evaluate it on the test fold.

The process is repeated 10 times until each fold is used for testing exactly once. We use a held-out validation set to fine tune the parameters of the classifier. The results are

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

[10]We did not observe any significant change in the scores by including the stopwords and punctuations.



(a) Presenting facts (b) Pointing out hypocrisy (c) Warning of consequences (d) Affiliation

(e) Denouncing speech (f) Humor and sarcasm (g) Positive tone (h) Hostile language

Figure 4: Word clouds for the different types of counterspeech used by the counterspeaker for hate speech against **LGBT**.

(a) Presenting facts (b) Pointing out hypocrisy (c) Warning of consequences (d) Affiliation

(e) Denouncing speech (f) Humor and sarcasm (g) Positive tone (h) Hostile language

Figure 5: Word clouds for the different types of counterspeech used by the counterspeaker for hate speech against **Jews**.

computed from the 10 tests and the means and standard deviations of different evaluation measures are reported in Table 3. We use accuracy along with weighted precision, recall, and F1-score as the evaluation measure.

Among the different features, we observe that sentence vectors seem to be performing much better than BoWV and TF-IDF in most of the cases. We got the best performance using XGBoost along with SV+TF-IDF+BoWV as features. Classifiers such as MLP and CatBoost also performed comparably well. Our best performing model achieves an accuracy of 71.6%. In the table we show some of the best results obtained using different classifier choices. The results from all the different classifiers and the different feature types are listed in our repository[6].

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. As a baseline, we use $General_B$ (Metz et al. 2012) which predicts the top most frequent labels of the dataset based on the cardinality[11] of the dataset. For our dataset, only the most frequent label 'Hostile language' was predicted to be relevant.

The performance of a multi-label classifier should be always assessed by means of several evaluation metrics (Madjarov et al. 2012). In this paper, the multi-label classifiers are evaluated using five measures: accuracy, precision, recall, F1-score and hamming loss (Godbole and Sarawagi 2004; Schapire and Singer 2000).

Accuracy is defined as the proportion of predicted *correct* labels to the *total* number of labels, averaged over all

[11]Cardinality represents the average size of the multi-labels in the dataset, which is 1.32 in our case. We follow the same procedure as the authors and take the closest integer value of the cardinality as the cardinality of the dataset (which will be 1).

Method	Precision	Recall	F1-Score	Accuracy
XGB+SV+TF-IDF+BOWV	0.716(+/-0.038)	0.715(+/-0.039)	0.715(+/-0.04)	0.716(+/-0.038)
MLP+SV+TF-IDF	0.714(+/-0.031)	0.713(+/-0.033)	0.713(+/-0.033)	0.714(+/-0.032)
CB+SV+TF-IDF+BOWV	0.708(+/-0.04)	0.706(+/-0.043)	0.705(+/-0.043)	0.707(+/-0.042)
RF+SV+TF-IDF+BOWV	0.697(+/-0.043)	0.693(+/-0.045)	0.692(+/-0.046)	0.695(+/-0.044)
SVC+SV+TF-IDF+BOWV	0.693(+/-0.029)	0.691(+/-0.03)	0.691(+/-0.03)	0.692(+/-0.029)

Table 3: Classification scores for the task of predicting if the given comment is counterspeech or non-counterspeech. The values reported are the means and standard deviations over 10-folds for each of the evaluation metric.

instances.

$$Accuracy = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Y_i \cup Z_i \mid} \quad (1)$$

Precision is defined as the proportion of predicted *correct* labels to the total number of *actual* labels, averaged over all instances

$$Precision = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Z_i \mid} \quad (2)$$

Recall is defined as the proportion of predicted *correct* labels to the total number of *predicted* labels, averaged over all instances

$$Recall = \frac{1}{\mid D \mid} \sum_{i=1}^{|D|} \frac{\mid Y_i \cap Z_i \mid}{\mid Y_i \mid} \quad (3)$$

F1-score is defined simply as the harmonic mean of precision and recall.

$$\textit{F1-score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Hamming loss is equal to 1 over $|D|$ (number of multi-label samples), multiplied by the sum of the symmetric differences between the predictions (Z_i) and the true labels (Y_i), divided by the number of labels (L), giving

$$Hamming\ Loss = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|}. \quad (5)$$

All these performance measures have values in the interval $[0...1]$. For Hamming loss, the smaller the value, the better the multi-label classifier performance is, while for the other measures, the greater values indicate better performance.

For the multi-label classification, we perform multi-label stratified[12] 10-fold cross validation (Sechidis, Tsoumakas, and Vlahavas 2011) on the dataset. The whole training set is partitioned into 10 folds, one is set apart for testing, and the remaining nine are used to train the model and evaluate it on the test fold. The process is repeated 10 times until each fold is used for testing exactly once. We use a held-out validation set to fine tune the parameters of the classifier. The results are computed from the 10 tests and the means and standard deviations of different evaluation measures are reported in Table 4. We obtain the best performance on using XGBoost with SV+TF-IDF+BoWV as the feature set.The results from all the different classifiers and the different feature types are listed in our repository[6]

To get a better understanding of how the classifier is performing for each counterspeech type, we look into the label-wise performance as illustrated in Table 5. These label-wise results are obtained using the best-classifier (XGBoost with SV+TF-IDF+BoWV) we obtained in the previous step. We observe that the classifier is able to perform good for certain classes such as 'Hostile' and 'Affiliation', while it performs poorly for other types such as 'Warning of offline/online consequences' and 'Pointing out Hypocrisy/contradiction'.

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. In this task, a binary classifier is built to predict if the given input text is a counterspeech or non-counterspeech. Since this task is the same as first task (counterspeech classification), we use the best model (XGBoost+SV+TF-IDF+BoWV) that we obtained from the first task. The training consists of data points from two communities and the test set will be the third community. The process is repeated until each community is used for testing exactly once. 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 accuracy, weighted precision, recall and F1-score. Table 6 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.

Discussion

We observe that non-counterspeech consists mainly of comments that agree with the main content in the video or hate speech 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

[12]https://github.com/trent-b/iterative-stratification

Method	Accuracy	Precision	Recall	F1-Score	Hamming Loss
$General_B$	0.322	0.397	0.322	0.356	0.191
XGB+SV+TF-IDF+BOWV	0.472(+/-0.012)	0.509(+/-0.012)	0.733(+/-0.011)	0.601(+/-0.011)	0.212(+/-0.015)
MLP+SV+TF-IDF	0.44(+/-0.014)	0.504(+/-0.013)	0.527(+/-0.021)	0.515(+/-0.015)	0.295(+/-0.018)
LR+SV+TF-IDF	0.469(+/-0.014)	0.5(+/-0.014)	0.734(+/-0.02)	0.595(+/-0.015)	0.212(+/-0.015)
GNB+SV	0.339(+/-0.014)	0.357(+/-0.016)	0.71(+/-0.02)	0.475(+/-0.018)	0.072(+/-0.012)
DT+SV+TF-IDF	0.301(+/-0.015)	0.356(+/-0.02)	0.361(+/-0.02)	0.358(+/-0.019)	0.193(+/-0.012)

Table 4: Classification scores for the task of multi-label classification of the types of counterspeech. The values reported are the means and standard deviations over 10-folds for each of the evaluation metric.

Counterspeech Type	Precision	Recall	F1-Score
Presenting facts	0.443(+/-0.036)	0.648(+/-0.049)	0.526(+/-0.040)
Pointing out hypocrisy	0.353(+/-0.026)	0.630(+/-0.043)	0.452(+/-0.031)
Warning of consequences	0.434(+/-0.057)	0.470(+/-0.087)	0.450(+/-0.066)
Affiliation	0.567(+/-0.052)	0.605(+/-0.084)	0.582(+/-0.052)
Denouncing hateful speech	0.413(+/-0.023)	0.694(+/-0.041)	0.518(+/-0.025)
Humor	0.462(+/-0.036)	0.661(+/-0.042)	0.543(+/-0.036)
Positive tone	0.430(+/-0.050)	0.598(+/-0.046)	0.500(+/-0.047)
Hostile	0.547(+/-0.016)	0.919(+/-0.017)	0.686(+/-0.015)

Table 5: The performance scores of each type of counterspeech as given on using the XGB+SV+TF-IDF+BoWV. The values reported are the means and standard deviations over 10-folds for each of the evaluation metric.

Train	Test	Precision	Recall	F1-Score	Accuracy
Blacks+Jews	LGBT	0.652	0.655	0.653	0.644
Jews+LGBT	Blacks	0.617	0.616	0.617	0.616
LGBT+Blacks	Jews	0.621	0.628	0.624	0.620

Table 6: 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.

	Jews		LGBT		Blacks	
Type of Counterspeech	#R	%A	#R	%A	#R	%A
Presenting facts	112	28.57	49	46.94	22	36.36
Pointing out hypocrisy or contradictions	114	42.98	68	41.18	33	39.39
Warning of offline or online consequences	45	71.11	30	26.67	55	34.54
Affiliation	121	51.24	76	40.79	86	1.16
Denouncing hateful or dangerous speech	124	44.35	85	43.53	120	19.17
Humor	62	46.77	101	65.35	49	28.57
Positive tone	156	42.31	54	38.89	165	6.06
Hostile	87	35.63	105	42.86	66	48.48

Table 7: Percentage of replies that agree with the opinion of the counter speaker.'#R' represents the number of replies that were tagged and '%A' represents the percentage of replies that agree with the counterspeakers comment.

are controlling the economy and are responsible for the destruction of their society. Many of the non-counterspeech also included holocaust denial (Gerstenfeld, Grant, and Chiang 2003). In case of *Blacks*, we find that the majority of non-counterspeech were hate speech 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 investigated into the people's reaction to the counterspeech. This would tell us how the community views these statements provided by the counterspeakers. For each target community and for each type of counterspeech, we randomly select 10 counterspeech comments that have at least two replies. Next, we ask the annotators to check if the response to the counterspeech comment is in agreement with the counterspeaker's opinion or against it. We report the results in Table 7. As observed from the table, the community acceptance (as observed by % agreement) of the type of counterspeech varies. In Jew community, counterspeech strategies involving 'Warning offline or online consequences' and 'Affiliation' seem to be more favoured by the community. In case of the LGBT community, 'Humor' seems to be the most acceptable form of counterspeech by the community while 'Warning offline or online consequences' seems to be the least favored tactics. In case of the Blacks community, we observe that counterspeech strategies such as 'Affiliation' and 'Positive Tone' receives very less community acceptance. We found several cases, in which the replies to such

counterspeech were stating that such 'Positive Tone' and 'Affiliation' will not change the stance of the hate speakers. Although, using 'Hostile language' seems to be very prevalent (see Table 1), we found 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.

The counterspeech classifiers can be used in several scenarios. One such promising area, is studying the effectiveness of the types of counterspeech on a larger scale. One could also use such classifiers to generate datasets that could potentially be used to build systems that automatically counter hate messages in online social media. Such a system would be very effective as they could characterize the hate speaker and provide an effective counterspeech based on his/her profile.

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

Bartlett, J., and Krasodomski-Jones, A. 2015. Counterspeech examining content that challenges extremism online. *Demos. Available at: http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf*.

Beauvois, J.-L.; Joule, R.-V.; and Brunetti, F. 1993. Cognitive rationalization and act rationalization in an escalation of commitment. *Basic and Applied Social Psychology* 14(1):1–17.

Benesch, S.; Ruths, D.; Dillon, K. P.; Saleem, H. M.; and Wright, L. 2016a. Considerations for successful counterspeech. *Dangerous Speech Project. Available at: https://dangerousspeech.org/considerations-for-successful-counterspeech/*.

Benesch, S.; Ruths, D.; Dillon, K. P.; Saleem, H. M.; and Wright, L. 2016b. Counterspeech on twitter: A field study. *Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/*.

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

Berger, J., and Strathearn, B. 2013. Who matters online: Measuring influence, evaluating content and countering violent extremism in online social networks. *International Centre for the Study of Radicalisation and Political Violence*.

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

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Citron, D. K., and Norton, H. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.* 91:1435.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670–680.

Dorogush, A. V.; Ershov, V.; and Gulin, A. 2017. Catboost: gradient boosting with categorical features support. In *Workshop on ML Systems at NIPS*.

Ernst, J.; Schmitt, J. B.; Rieger, D.; Beier, A. K.; Vorderer, P.; Bente, G.; and Roth, H.-J. 2017. Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization* 1–49.

Gagliardone, I.; Gal, D.; Alves, T.; and Martinez, G. 2015. *Countering online hate speech*. UNESCO Publishing.

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

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

Kane, A. A.; Argote, L.; and Levine, J. M. 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.

Madjarov, G.; Kocev, D.; Gjorgjevikj, D.; and Džeroski, S. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition* 45(9):3084–3104.

Maity, S. K.; Chakraborty, A.; Goyal, P.; and Mukherjee, A. 2018. Opinion conflicts: An effective route to detect incivility in twitter. *Proc. ACM Hum.-Comput. Interact.* 2(CSCW):117:1–117:27.

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

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

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

Metz, J.; de Abreu, L. F.; Cherman, E. A.; and Monard, M. C. 2012. On the estimation of predictive evaluation measure baselines for multi-label learning. In *Ibero-American Conference on Artificial Intelligence*, 189–198. Springer.

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

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Ravenscroft, J.; Oellrich, A.; Saha, S.; and Liakata, M. 2016. Multi-label annotation in scientific articles-the multi-label cancer risk assessment corpus. In *LREC*.

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

Rudra, K.; Ghosh, S.; Ganguly, N.; Goyal, P.; and Ghosh, S. 2015. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *CIKM*, 583–592.

Schapire, R. E., and Singer, Y. 2000. Boostexter: A boosting-based system for text categorization. *Machine learning* 39(2-3):135–168.

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

Sechidis, K.; Tsoumakas, G.; and Vlahavas, I. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 145–158. Springer.

Siersdorfer, S.; Chelaru, S.; Pedro, J. S.; Altingovde, I. S.; and Nejdl, W. 2014. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)* 8(3):17.

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

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

Thelwall, M.; Sud, P.; and Vis, F. 2012. Commenting on youtube videos: From guatemalan rock to el big bang. *Journal of the American Society for Information Science and Technology* 63(3):616–629.

Wright, L.; Ruths, D.; Dillon, K. P.; Saleem, H. M.; and Benesch, S. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, 57–62.