Countering hate on social media: Large-scale classification of hate and counter speech

Joshua Garland*
Santa Fe Institute
Santa Fe, NM 87501 USA
joshua@santafe.edu

Keyan Ghazi-Zahedi*
Max Planck Institute for Mathematics in the Sciences
Inselstrasse 22, 04103
Leipzig, Germany

Jean-Gabriel Young
Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI 48109, USA

Laurent Hébert-Dufresne
Vermont Complex Systems Center
University of Vermont,
Burlington, VT 05405 USA

Mirta Galesic
Santa Fe Institute
Santa Fe, NM 87501 USA

Abstract

Hateful rhetoric is plaguing online discourse, fostering extreme societal movements and possibly giving rise to real-world violence. A potential solution to this growing global problem is citizen-generated counter speech where citizens actively engage with hate speech to restore civil non-polarized discourse. However, its actual effectiveness in curbing the spread of hatred is unknown and hard to quantify. One major obstacle to researching this question is a lack of large labeled data sets for training automated classifiers to identify counter speech. Here we use a unique situation in Germany where self-labeling groups engaged in organized online hate and counter speech. We use an ensemble learning algorithm which pairs a variety of paragraph embeddings with regularized logistic regression functions to classify both hate and counter speech in a corpus of millions of relevant tweets from these two groups. Our pipeline achieves macro F1 scores on out of sample balanced test sets ranging from 0.76 to 0.97—accuracy in line and even exceeding the state of the art. We then use the classifier to discover hate and counter speech in more than 135,000 fully-resolved Twitter conversations occurring from 2013 to 2018 and study their frequency and interaction. Altogether, our results highlight the potential of automated methods to evaluate the impact of coordinated counter speech in stabilizing conversations on social media.

*Denotes equal contribution.

1 Introduction

Hate speech is a growing problem in many countries [Bakalis, 2015, Hawdon et al., 2017], it can have serious psychological consequences [Oksanen et al., 2018], and is related to, and perhaps even contributing to, real-world violence [Müller and Schwarz, 2019]. While censorship can help curb hate speech [Álvarez-Benjumea and Winter, 2018], it can also impinge on civil liberties and might merely disperse rather than reduce hate [Chandrasekharan et al., 2017]. A promising alternative approach to reduce toxic discourse without recourse to censorship is so-called *counter speech*, which broadly refers to citizens' response to hateful speech in order to stop it, reduce its consequences, and discourage it [Benesch et al., 2016, Rieger et al., 2018].

It is unknown, however, whether counter speech is actually effective due to the lack of systematic large-scale studies on its impact [Gaffney et al., 2019, Gagliardone et al., 2015]. A major reason has been the difficulty of designing automated algorithms for discovering counter speech in large online corpora, stemming mostly from the lack of labeled training sets including both hate and counter speech. Past studies that provided insightful analyses of the effectiveness of counter speech mostly used hand-coded examples and were thus limited to small samples of discourse [Mathew et al., 2018, 2019, Wright et al., 2017, Ziegele et al., 2018, Ziems et al., 2020].

We perform the first large-scale classification

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 102–112

https://doi.org/10.18653/v1/P17

study of hate and counter speech, using a unique situation in Germany, where self-labeling hate and counter speech groups engaged in discussions around current societal topics such as immigration and elections. One is "Reconquista Germanica" (RG), a highly-organized hate group which aimed to disrupt political discussions and promote the right-wing populist, nationalist party Alternative für Deutschland (AfD). At their peak time, RG had between 1,500 and 3,000 active members. The counter group "Reconquista Internet" (RI) formed in late April 2018 with the aim of countering RG's hateful messaging through counter speech and to re-balance the public discourse. Within the first week, approximately 45,000 users joined the discord server where RI was being organized. At their peak, RI had an estimated 62,000 registered and verified members, of which over 4,000 were active on their discord server for the first few months. However, RI quickly lost a significant amount of active members, splintering into independent though cooperating smaller groups. We collected millions of tweets from members of these two groups and built labeled training set orders of magnitude larger than past studies of counter speech. By building an ensemble learning system with this large corpus we trained highly accurate classifiers which matched human judgment. We also used this system to study more than 135,000 conversations on German Twitter to understand the interactions between counter and hate groups online—an important first step in studying the impacts of counter speech on a large scale.

2 Background and Past Research

2.1 Hate and counter speech

There are many definitions of online hate speech and its meaning is developing over time. According to more narrow definitions, it refers to insults, discrimination, or intimidation of individuals or groups on the Internet, on the grounds of their supposed race, ethnic origin, gender, religion, or political beliefs [Blaya, 2019, Weber, 2009]. However, the term can also be extended to speech that aims to spread fearful, negative, and harmful stereotypes, call for exclusion or segregation, incite hatred, and encourage violence against a particular group [Gagliardone et al., 2015, you, 2019, twi, 2019, fac, 2019], be it using words, symbols, images, or other media.

Counter speech entails a citizen-generated response to online hate in order to stop and prevent the spread of hate speech, and if possible discourage it by changing perpetrators' attitudes and prevailing social norms. Counter speech intervention programs focus on empowering Internet users to speak up against online hate [Gagliardone et al., 2015]. For instance, programs such as *seriously* [ser, 2019] and the *Social Media Helpline* [smh, 2019] help users to recognize different kinds of online hate and prepare appropriate responses. Counter speech is seen as a feasible way of countering online hate, with a potential to increase civility and deliberation quality of online discussions [Ziegele et al., 2018, Habermas, 2015].

2.2 Classification of Hate and Counter Speech

There has been a lot of work on developing classifiers to detect hate speech online (e.g., [Brassard-Gourdeau and Khoury, 2018, Basile et al., 2019, Burnap et al., 2015, Burnap and Williams, 2016, Ribeiro et al., 2018, Zhang and Luo, 2019, Bosco et al., 2018, de Gibert et al., 2018, Kshirsagar et al., 2018, MacAvaney et al., 2019, Malmasi and Zampieri, 2018, Pitsilis et al., 2018, Al-Hassan and Al-Dossari, 2019, Vidgen and Yasseri, 2020, Zimmerman et al., 2018]). Many different learning algorithms have been used to perform this classification, ranging from support vector machines and random forests to convolutional and recurrent neural networks [Zhang and Luo, 2019, Burnap and Williams, 2016, Bosco et al., 2018, de Gibert et al., 2018, Kshirsagar et al., 2018, Malmasi and Zampieri, 2018, Pitsilis et al., 2018, Al-Hassan and Al-Dossari, 2019, Vidgen and Yasseri, 2020, Zimmerman et al., 2018]. These algorithms use a variety of feature extraction methods, for example, frequency scores of different n-grams, word and document embeddings [Le and Mikolov, 2014, Pennington et al., 2014], sentiment scores [Brassard-Gourdeau and Khoury, 2018, Burnap et al., 2015], part-of-speech scores such as the frequency of adjectives versus nouns used to describe target groups, 'othering' language (e.g., 'we' vs. 'them' [Burnap and Williams, 2016]), and meta-information about the text authors (e.g., keywords from user bios, usage patterns, their connections based on replies, retweets, and following patterns [Ribeiro et al., 2018]). Zhang and Luo [2019] compare several state-of-the art methods for automatic detection of hate speech, including

SVM and different implementations of convolutional neural networks (CNN) using word embeddings, on seven different Twitter data sets. The best performing methodology, based on a combination of CNN and gated recurrent networks (GRU), yields macro F1 scores ranging from 0.64 to 0.83. Other promising approaches include an ensemble of different CNNs with different weight initializations proposed by Zimmerman et al. [Zimmerman et al., 2018], BERT [Devlin et al., 2019], and a multi-view stacked SVM approach proposed by McAvaney et al. [MacAvaney et al., 2019]. The best results reported for these approaches are achieved by MacAvaney et al. [2019] using the `hatebase.org` database (a set of 24,802 tweets provided by Davidson et al. [2017], receiving F1 scores of 0.91 with a neural ensemble and 0.89 using BERT.

Compared to the number of studies investigating automatic detection of online hate, there have been far fewer studies that aim to automatically detect counter speech. One reason for this is the difficulty and subjectivity of automated identification of counter speech [Kennedy et al., 2017]. As a result, most past studies use hand-coded examples for this task. For instance, Mathew et al. [2019] analyzed more than 9,000 hand-coded counter speech and neutral comments posted in response to hateful YouTube videos. They found that for discriminating counter speech vs. non-counter speech, the combination of tf-idf vectors as features and logistic regression as the classifier performed best, achieving an F1 score of 0.73. In another study, Mathew et al. [2018] analyzed 1,290 pairs of Twitter messages containing hand-coded hate and counter speech. In this data set, a boosting algorithm based mostly on tf-idf values and lexical properties of tweets performed best, achieving F1 score of 0.77. Wright et al. [2017] provide a qualitative analysis of individual examples of counter speech. Ziegele et al. [2018] employed 52 undergraduate students to hand-code 9,763 Facebook messages. A study concurrent to ours [Ziems et al., 2020] investigated hate and counter speech in the context of racist sentiment surrounding COVID-19. They hand-coded 2,319 tweets, of which they labeled 678 as hateful, 359 as counter speech, and 961 as neutral. They were able to achieve F1 scores on unbalanced sets of 0.49 for counter and 0.68 for hate.

While extremely useful as a first step in analyzing counter speech, these studies are intrinsically limited because manual coding of counter speech is costly and hard to scale to the size needed to train sophisticated classifiers.

3 Data and Methods

3.1 Data Collection Strategy

We built our corpus of hate speech by collecting the timelines of 2,120 publicly known members of RG using the Twitter API. We used the list of hate accounts known as the “Böhmermann Liste," which was promoted as a list of accounts that spread hateful rhetoric, promote alt-right propaganda, or engage in directly hateful speech. As a secondary check of the list, we further verified these accounts by ensuring that the names and/or bios of these accounts contained known RG badges and no known badges of RI (see Table S1 in the Supplementary Materials for a list of these features). Finally, we had an expert hand-verify a large random sample of these accounts to ensure they were indeed actively taking part in hate speech. This resulted in more than 4.6 million tweets which with high likelihood contained some hateful rhetoric. While we cannot guarantee that every single one of these tweets was hate speech, given the purpose of these accounts, we can be reasonably confident in labeling the tweets sent from these accounts as hate speech.

Building our corpus of counter speech was a bit more challenging. We began our search with a hand-curated list of 103 accounts comprised of the core RI Twitter team, each of which were highly focused to their primary objective of engaging in various forms of counter speech. We collected the timelines of each of these known members of RI using the Twitter API. While we were highly confident in this sample of tweets being counter speech, it did not provide enough examples to build balanced training sets. Therefore, to expand our counter speech corpus we also collected the follower-followee network of these 103 core RI members using the Twitter API. This resulted in a list of 70,537 potential counter accounts.

We narrowed down this list of potential counter accounts by only including users that appeared in at least 5 of the follower-followee networks of core RI members. We then further required that each user self-identify as an RI member by using language features typical of RI members in their bios (see Table S1 in the Supplementary Materials for a list of these features) and also eliminated all users from

this subset who used any RG features in their bios, to remove troll accounts that used both classes of features in their bios. Finally, we also enrolled an expert to check many of these accounts to ensure they were actively taking part in counter speech. This process resulted in a total of 1,472 profiles which we labeled as counter accounts.

To build our corpus of counter speech we collected the timelines of each of these additional accounts as well as the timelines of the core RI Twitter team. This resulted in a total of 4,323,881 tweets which had a high probability of containing counter speech. For training, we labeled all of these tweets as counter speech. It is likely that not all of these tweets in fact contained counter speech, especially those written by users that were identified through our network search. However, one can think of our data gathering process as trading-off some accuracy for a significant increase in scale. To check that accuracy was not strongly impacted, we verified that our classification of counter speech aligned with human judgment after classification (see Section 3.2) and added a post-hoc criterion to eliminate tweets that are not confidently labeled as counter speech (or hate speech) by the trained classifier.

In addition to these labeled tweets, we collected 204,544 fully-resolved conversations (reply trees) that grew in response to tweets of prominent accounts engaged in political speech on German Twitter from 2013 to 2018. These included accounts of large news organizations (e.g., faznet, tagesschau, tagesthemen, derspiegel and spiegelonline, diezeit, and zdfheute), well-known journalists and bloggers (e.g., annewilltalk, dunjahayali, janboehm, jkasek, maischberger, nicolediekmann), and politicians (e.g. cem_oezdemir, c_lindner, goeringeckardt, heikomaas, olafscholz, renatekuenast), all of which were known to be targets of hate speech. Indeed, the majority of these conversations involve instances of both hate and counter speech. We focused on 137,725 trees which originated from 11 accounts that contributed trees in at least 69 of 72 possible months throughout the examined period: derspiegel, goeringeckardt, jkasek, olafscholz, regsprecher, zdfheute, c_lindner, faznet, janboehm, nicolediekmann, and tagesschau. The tweets in these trees were used to study the dynamics of hate and counter speech over time and were not used to train or evaluate the accuracy of the classifiers. Figure 1 shows a few example trees labeled using the pipeline described in Section 3.2.

3.2 Classification Pipeline

As is common in the literature [Schmidt and Wiegand, 2017] we split our classification pipeline into two stages: extraction of features from text, and classification based on those features. Before tweets were used in this pipeline they went through a minor preprocessing stage. All of the text was made lower case, and hashtags, usernames e.g., @username, punctuation and "RT:" were all stripped out of the tweet's text. Finally, depending on the model being trained, we removed stop words using one of two lists ("heavy" and "light"), or we did not remove any stop words. The "heavy" stop word list eliminated 231 German words based on nltk's German stop word list. The "light" stop word list was based on the heavy list without all words which have been shown to be relevant identifiers in an "us vs. them" discourse [Burnap and Williams, 2016], e.g., wir, uns, sie (we, us, them). This list eliminated 48 words (see Supplementary Materials).

To extract features from each processed timeline tweet, we constructed paragraph embeddings, also known as doc2vec models [Le and Mikolov, 2014], using the standard gensim implementation [Řehůřek and Sojka, 2010]. We will refer to a generic doc2vec model as $\mathcal{M}_{d2v}$. We performed a parameter sweep following standard practice and the guidelines of [Lau and Baldwin, 2016]. This sweep includes the analysis of several doc2vec parameters e.g, maximum distance between current and predicted words, "distributed-memory" vs "distributed bag of words" frameworks, and five different document label types, as well as three levels of stop word removal.

The five different document label types used to train $\mathcal{M}_{d2v}$ were as follows. 1) Each tweet was treated as a single document and labeled with a *unique* label, *viz.,* the unique tweet-id assigned by Twitter. 2) All tweets by a single *author* used the same document label, *viz.,* the user-id assigned by Twitter. This effectively made every tweet by a particular user a single document. These are the more traditional choices for document labeling. We also used three other labels which incorporate the classification stage into the feature development: 3) Each tweet was assigned a *group* label, with all tweets from RG accounts labeled "hate" and all

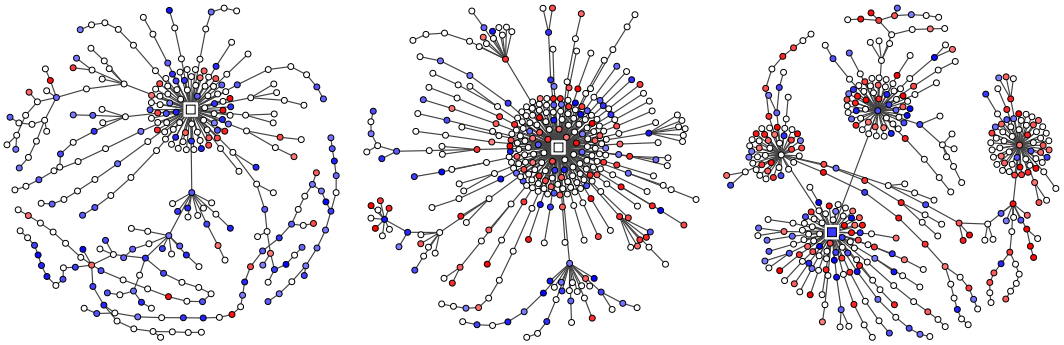


Figure 1: Examples of Twitter conversations (reply trees) with labeled hate (red), counter (blue), and neutral speech (white). The root node is shown as a large square. We used a confidence threshold of $\gamma = 0.75$ and a panel of 25 experts to classify these tweets, as described later in Section 3.2.

tweets from RI labeled "counter". This treats all RG tweets as one document and all RI tweets as another document, incorporating the the label we care about into the feature development stage. However, it conflates all the tweets into two documents. To avoid this we also trained $\mathcal{M}_{d2v}$ using multi-label setups. In particular, we trained models where we 4) labeled each tweet using both the author's identifier as well as the group identifier and separate models which 5) labeled each tweet with a unique identifier as well as the group identifier.

Every $\mathcal{M}_{d2v}$ was trained on five different but partially overlapping training sets (approximately 27% overlap). Each training set included 500,000 randomly selected tweets originating from RG accounts and another 500,000 coming from RI accounts. This produced a balanced training set with 50% hate speech and 50% counter speech. This is important in interpreting our classification results correctly, and avoiding accuracy inflation due to unbalanced sets, an apparent frequent problem with much of the current literature where hate speech is highly under sampled [Zhang and Luo, 2019, MacAvaney et al., 2019]. We refer to these training sets as $\mathcal{T}_{in,i}$, to denote the i^{th} in-sample training set.

Let $\{\mathcal{M}_{d2v}, \mathcal{T}_{in,i}\}$ be a trained doc2vec model and the corresponding training set it was trained on. For each tweet $t_j \in \mathcal{T}_{in,i}$ we use $\mathcal{M}_{d2v}$ to infer a corresponding feature vector $x_j \in \mathbb{R}^{300}$, as $x_j = \mathcal{M}_{d2v}(t_j)$. With each tweet mapped to a feature vector we constructed a decision boundary between tweets from RG members and tweets from RI members using regularized logistic regression. In other words, we wrote the likelihood that tweet j is labeled as coming from an RG/RI account as:

$$h_\theta(x_j) = g(\theta^T x_j), \qquad g(z) = \frac{1}{1+e^{-z}}. \quad (1)$$

where $\theta \in \mathbb{R}^{300}$ is the vector of feature weights. Given a set of labels $\mathcal{L} = \{H, C\}$ for all tweets, we then learned the vector θ that best separates the data by minimizing the loss $-\sum_j \log h_\theta(x_j)$ under an ℓ_2 regularization constraint $\frac{1}{\lambda}||\theta||_2$, where λ is a fixed regularization parameter. We finally solved for θ using the the LBFGS algorithm as implemented in scikit-learn [Pedregosa et al., 2011].

To evaluate the accuracy of the resulting hypothesis function h_θ we evaluated its predictive accuracy on an out of sample test set denoted $\mathcal{T}_{out,i}$. Each out of sample test set $\mathcal{T}_{out,i}$ consisted of 50,000 tweets from both groups, chosen at random while ensuring that $\mathcal{T}_{out,i} \cap \mathcal{T}_{in,i} = \emptyset$. For each, $\mathcal{M}_{d2v}$, h_θ, $\mathcal{T}_{out,i}$ combination we determined the probability of each class label $l \in \mathcal{L}$ for each $t \in \mathcal{T}_{out,i}$. In particular, for each tweet $t \in \mathcal{T}_{out,i}$ and each label $l \in \{H, C\}$ we computed $h_\theta(\mathcal{M}_{d2v}(t)) = p(l|\mathcal{M}_{d2v}(t); \theta)$, where $p(l|\mathcal{M}_{d2v}(t); \theta)$ denotes the probability that a tweet t has label l when classified with the feature vector calculated with model $\mathcal{M}_{d2v}$. The accuracy of this prediction was then assessed against the known labels.

In addition to logistic regressions, we also used word bias and n-gram based classifiers like those used in [Jaki and De Smedt, 2019], as well as xgboost [Chen and Guestrin, 2016] with a variety of parameters. However, in both cases the accuracy was worse (only slightly so for xgboost) than the logistic regression experiments reported in Section 4, so we omit these details for brevity.

Instead of looking for the single optimal

$(\mathcal{M}_{d2v}, h_\theta)$ parameterization we used an ensemble learning approach to classification by constructing a "panel of experts." The panel is comprised of N experts which are defined to be the combination of a feature extraction method $\mathcal{M}_{d2v}$ as well as a classification or hypothesis function h_θ. An ensemble learning approach combines multiple hypothesis functions to form a more robust hypothesis jointly which can lead to greater generalizability and increased out-of-sample accuracy.

In this ensemble classification method, each expert is given a tweet in a balanced out-of-sample test set $\mathcal{T}_{out,i}$ and asked to assign to it a probability that it belongs to each class $l \in \mathcal{L}$. For each tweet $t \in \mathcal{T}_{out,i}$ we computed a hate and counter score, S_h and S_c respectively, in the following way:

$$S_h(t) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_i(t; H), \tag{2}$$

where $\mathcal{E}_i(t; l)$ is the probability that expert i assigns to label $l \in \{H, C\}$ for tweet t. Note that $S_c(t) = 1 - S_h(t)$. Note that $\mathcal{T}_{out,i} \cap \mathcal{T}_{in,j}$ need not be empty when $i \neq j$. As such, if a tweet appeared in the training set of an expert, we withheld its vote to avoid leaking training data.

For final classification we then defined a "confidence threshold" $\gamma \in [1/2, 1]$, and used a confidence voting system with thresholding to assign labels to tweets. If $S_h(t) > \gamma$ then t is labeled H, and if $S_c(t) > \gamma$ then t is labeled C. If $S_c(t)$ and $S_h(t)$ are both less than the given threshold the tweet is marked as neutral speech and the panel effectively abstains from voting. This results in some tweets which the panel of experts is not confident labeling as hate or counter speech—a crucial feature of the classifier. Indeed, the primary goal of our classifier is to identify hate and counter speech in online political discourse. Since not all online political discourse is hate or counter speech, the classifier must be able to flag neutral speech, too. The confidence threshold allows us to identify neutral speech by contrasting it with counter and hate speech.

An alternative solution would be to build a ternary classifier that can distinguish between hate, counter and neutral speech. A big challenge would then be to obtain a corpus of neutral speech relevant to political discourse, yet free of hate or counter speech. One could use tweets from politicians or news outlets to build a neutral corpus but then one would have to be careful to have a balanced representation across the political spectrum so that the neutral class is not biased toward the speech patterns of a particular party or news outlet. As we were confident in our hate and counter speech labeling and not confident in labeling an unbiased neutral corpus we chose to use our ranked classification method instead. However, this is certainly a potentially fruitful area for future research.

3.3 Crowdsourcing

To test whether the automated classifier corresponded to human judgment, we conducted a crowdsourcing study in which human judges evaluated some of the same tweets evaluated by the classifier. Since our corpus mostly contained German tweets, judges were recruited among members of Mechanical Turk who indicated that they can speak German. To qualify, they had to complete a relatively difficult German test item taken from a Goethe Institut's test for B1 German level, which asked them to interpret comments of three individuals about violence in video games. They also evaluated a test sample of a few dozen tweets from across the score spectrum. We also checked their answers to ensure a basic level of conscientiousness. Of the initial 55 candidate raters, 28 raters both solved the German test correctly and gave ratings to the test sample of tweets which indicated that they paid attention to their content. These 28 raters were asked to evaluate 5000 randomly selected tweets evenly spread across the whole range of scores $S_h(t)$. Raters ranked tweets on a scale of 1 to 5, from "very likely counter speech" to "very likely hate speech," with 3 corresponding to neutral content. Each tweet was evaluated by at least 2 different raters. Standard inter-rater reliability measures are not possible because most tweets were evaluated by a different pair of raters. However, median difference in ratings of each tweet was 0, and absolute mean was 0.57. In other words, different raters evaluated the tweets well within one point on our 5-point scale, suggesting reasonable correspondence of evaluations by different raters.

4 Results

Classification Results All combinations of feature extraction models $\mathcal{M}_{d2v}$ and classification functions h_θ produced a total of $N = 289$ possible experts. We found that the top 10 highest performing parameter sets across all five balanced training sets were the same for all $\mathcal{M}_{d2v}$. In particular, a max-

imum distance between the current and predicted word within a sentence of 5, ignoring all words that occurred less than 10 times, and an initial learning rate of 0.025 with a minimum learning rate of 0.00025, resulted in the highest accuracy. Each of these top performing experts were trained for 20 epochs, with a distributed bag-of-words framework. The optimal preprocessing parameters were also the same across the top 10. Each of these used light stop word removal. The optimal document labeling was also the same across these 10 experts, namely the author-group labeling. Recall that this labeling scheme tagged each tweet with the authors' unique identifier as well as the known group identification (RG vs RI). While the top models had the same training parameters aside from λ, the models were trained across 5 different training sets, which led to experts that could differ significantly.

These top 10 experts had individual F1 scores of 0.755 ± 0.0012 on their individual test sets, when forced to make a classification (confidence $\gamma = 1/2$). Taking into account that each $\mathcal{T}_{out,i}$ was balanced, containing 50,000 hate tweets and 50,000 counter tweets, these F1 scores do not suffer from accuracy inflation that would occur with an unbalanced test set [Zhang and Luo, 2019, MacAvaney et al., 2019]. This result compares well to previous studies that used smaller unbalanced data sets and achieved F1 scores ranging from 0.49 to 0.77 [Mathew et al., 2019, 2018, Ziems et al., 2020].

As mentioned in Section 3, we did not only use experts in isolation, but also in an ensemble learning approach where the experts could vote on the class label for each tweet in a given test set. Due to variations in the training sets and parameters, each expert had a slightly different view of the language, suggesting that combining their knowledge might be beneficial. Using the top 10 experts as a panel, instead of individually as just discussed, we obtained an improved average F1 score across all 5 out-of-sample test sets of 0.7616 ± 0.00083. Increasing the size of the panel to include the top 25 experts resulted in an average F1 score across the 5 test sets of 0.7618 ± 0.0007, see Table 1. We used this large panel for all of our subsequent results.

We also obtained improved results when we varied confidence threshold γ and allowed the experts to withhold their vote on contentious tweets. Increasing the confidence threshold naturally decreased the number of tweets classified as hate or counter speech. As expected, we found that this led to an increased overall precision, recall, and F1 score, since the labeled tweets were those for which the panel was more certain (see Table 1).

The scores for $\gamma > 1/2$ should be viewed with cautious optimism. The thresholding procedure causes many examples—correctly and incorrectly labeled—to be ignored from these calculations, which obviously may bias these scores in unpredictable ways. Even so, these scores provide a rough approximation of how accurate we can expect the classifier to be when applied to the reply tree dataset at a given confidence threshold. As discussed earlier, while this makes these scores challenging to interpret, the threshold is necessary to avoid mislabeling neutral speech in these conversations.

Comparison to Human Judgment Our crowdsourcing results, shown in Figure 2, suggest that our automated classifier aligns well with human judgment. Overall correlation between classifier scores and human judgments was $r = 0.94$. The correlation was somewhat lower for tweets classified as counter speech ($r = 0.75$) than for those classified as hate ($r = 0.96$). This could indicate that to humans counter speech looks more like 'neutral' discourse than hate speech does, or this could be a reflection of the slightly weaker counter speech labeling scheme described in Section 3.1 and suggest that counter speech is more challenging for the classifier to identify, or a combination of both. As expected, classifier scores around 0.5 received intermediate hate scores from human judges as well. The labels assigned by human judges were not used during the classification training or annotation process and were only used as a sanity check on the alignment of the classifier with human judgment.

Tree Coloring and Analysis We next used our classifier to label 137,725 fully-resolved conversations (reply trees) related to current societal and political issues on German Twitter between 2013 and 2018. Due to limited space, here we focus our analysis on two primary questions. For the interested reader, please see [Garland et al., 2020] for a much deeper analysis of this rich dataset.

First, how do hate and counter speech develop over time? To study this, we calculated the proportion of hate and counter speech of all speech occurring in each month (using $\gamma = 0.75$), as well as the average hate and counter score for all tweets

Results for Top Classifier Overall				
γ	Precis.	Recall	F1	Labeled
0.50	0.757	0.757	0.757	100.0%
0.65	0.837	0.837	0.837	70.29%
0.75	0.883	0.882	0.882	53.99%
0.85	0.924	0.924	0.924	37.93%
0.95	0.970	0.970	0.970	18.49%

Results for Panel of Experts				
γ	Precis.	Recall	F1	Labeled
0.50	0.763	0.762	0.762	100.0%
0.65	0.854	0.854	0.854	66.44%
0.75	0.897	0.897	0.897	49.43%
0.85	0.939	0.939	0.939	33.45%
0.95	0.977	0.977	0.977	15.38%

Table 1: Classification scores for the top classifier (left) and a panel of experts using the top 25 experts (right). γ is the confidence threshold, and "Labeled" is the percentage of examples in the test set that are labeled as either hate or counter at a confidence level of γ. The top row represents traditional accuracy measures, which compare favorably to previous studies that used smaller unbalanced data sets and achieved F1 scores ranging from 0.49 to 0.77. The remaining rows are more nuanced and need to be viewed with cautious optimism, see text for a discussion of these issues.

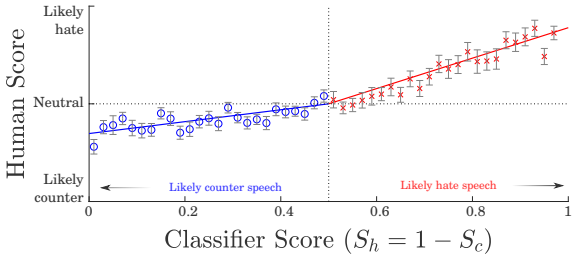


Figure 2: Human judgment of hate and counter speech corresponds to automated classification (panel of 25 experts). Average human judgments of tweets classified as counter speech by our method are shown in blue (left-half), and judgments for tweets classified as hate are shown in red (right-half). Individual human judgments are averaged across bins of width 0.02 of classifier scores for the original tweet. Error bars represent $\pm$ one standard error.

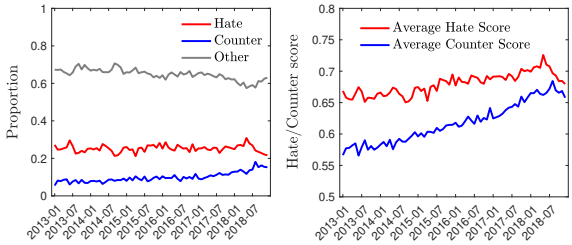
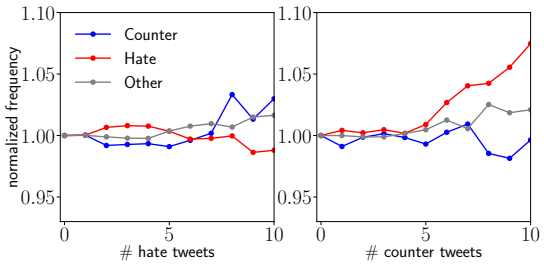


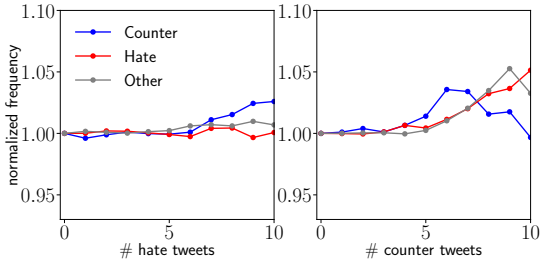
Figure 3: Proportion of hate, counter, and other speech in reply trees from 2013-2018, using a $\gamma = 0.75$ threshold (left panel), and average hate and counter score of tweets exceeding the $\gamma = 1/2$ threshold (right panel). After the establishment of RI in April 2018, the proportion of counter speech increases, and the ongoing increase in polarization is slowed down as indicated by a decrease in average hate and counter scores.

exceeding $\gamma = 1/2$. As Figure 3 shows, the proportion of hate speech was rather stable throughout the examined period, slightly increasing towards the end (red line in the left panel). However, its average score was consistently increasing over time (red line in the right panel). The proportion of counter speech was increasing somewhat throughout this period (blue line in the left panel), but its score increased quite strongly towards more extreme speech (blue line in the right panel). A notable change occurred in May 2018, when RI became active: the proportion of counter and other speech increased, and the proportion as well as extremity of hate speech decreased in the following months. This result suggests that organized counter speech might have helped in balancing polarized and hateful discourse, although causality is difficult to establish given the complex web of online and offline events and process in the broader society throughout that time.

Second, we conducted an initial analysis of how hate and counter speech interact in reply trees. We asked, how do tweets identified as hate or counter speech change the expected frequency of future hate and counter speech in a reply tree? For this analysis, we used reply trees that have at least 10 tweets identified as hate and at least 10 identified as counter speech, using a 70% threshold on scores assigned by a panel of the top 25 experts. We measured the overall frequency of assigned labels in every individual tree, and tracked how this frequency increases or decreases in time as more tweets identified as hate or counter are posted. We compared 6-month periods before and after the establishment of RI. Results are shown in Figure 4. Before RI was founded (Figure 4a), a low amount of hate tweets



(a) November 2017 - April 2018



(b) May 2018 - October 2018

Figure 4: Frequency of hate, counter, and other tweets following a hate (counter) tweet, normalized by the overall frequency of these types of tweets in a tree. Panel (a) shows the 6-months period before the establishment of RI, and panel (b) shows the 6-months period after RI was formed. By comparing the right panels of both (a) and (b), tweets from organized counter speech tend to attract more counter and other speech and attract less hate speech than tweets from non-organized counter speech.

(first panel) somewhat attracted additional hate and suppressed counter speech. However, once many hate tweets were posted, counter speech increased and hate decreased. Similarly, counter tweets (second panel) did not have much effect on hate at first but once there were many counter tweets in a tree they attracted much more hate speech. Importantly, counter speech attracted less hate and stimulated additional counter speech more effectively after RI was formed in April 2018 (Figure 4b). In all time periods, we also found that counter speech tweets were more likely than hate speech to stimulate neutral or unclassified speech; suggesting that counter speech contributed to depolarizing individual discussions.

Taken together, these results suggest that organized counter speech was associated with a more balanced discourse, reflected in an increased proportion of counter speech in discussions and reduced extremity of hate (Figure 3) and counter speech having a strong influence in attracting more counter and neutral speech while not attracting more hate (Figure 4).

5 Conclusion

Online hate speech is a problem shared by every social media platform, and yet there are still no clear solutions to this growing problem. A potential solution aimed at returning online discourse to civility is citizen-generated counter speech. Until now, studying counter speech and its effectiveness has been limited to small-scale hand-labeled studies. In this paper, we leveraged a unique situation in Germany to perform the first large-scale automated classification of counter speech. Our methods provided F1 scores on a balanced set of 100,000 out-of-sample tweets ranging from 0.76 to 0.97 depending on the confidence threshold being used. Beyond accuracy measures, we used crowdsourcing to verify that the conclusions reached by our classifier were in-line with human judgment.

We were able to use this classification algorithm to identify hate and counter speech in over 135,000 fully resolved Twitter conversations from 2013-2018. Our results suggest that counter speech might have contributed to depolarization of discussions and that organized counter speech by RI might have stimulated further counter speech and attracted less hateful responses. While causality cannot be established due to the many other ongoing societal processes at the time, our results suggest that organized counter speech may be a powerful solution to combating the spread of hate online. We hope that the framework developed in this paper will be a starting point to understand the dynamics between hate and counter speech and help develop actionable strategies.

Acknowledgments

The authors would like to thank Will Tracy and Santa Fe Institute's Applied Complexity team for support and resources throughout this project. J.G. was partially supported by an Omidyar and an Applied Complexity Fellowship at the Santa Fe Institute. J.-G.Y. was supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award. J.-G.Y and L.H.-D. were supported by Google Open Source under the Open-Source Complex Ecosystems And Networks (OCEAN) project. M.G. was partially supported by NSF-DRMS 1757211. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funders.

References

Chara Bakalis. *Cyberhate: An issue of continued concern for the council of Europe's anti-racism commission*. Council of Europe, 2015.

James Hawdon, Atte Oksanen, and Pekka Räsänen. Exposure to online hate in four nations: A cross-national consideration. *Deviant Behav.*, 38(3):254–266, 2017.

Atte Oksanen, Markus Kaakinen, Jaana Minkkinen, Pekka Räsänen, Bernard Enjolras, and Kari Steen-Johnsen. Perceived societal fear and cyberhate after the November 2015 paris terrorist attacks. *Terror. Political Violence*, pages 1–20, 2018.

Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *SSRN:3082972*, 2019.

Amalia Álvarez-Benjumea and Fabian Winter. Normative change and culture of hate: An experiment in online environments. *Eur. Sociol. Rev.*, 34(3):223–237, 2018.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *Proceedings of the ACM on Human-Computer Interaction*, volume 1, pages 1–22, 2017.

S Benesch, D Ruths, KP Dillon, H M Saleem, and L Wright. Considerations for successful counterspeech, 2016. URL https://https://dangerousspeech.org/considerations-for-successful-counterspeech/.

Diana Rieger, Josephine B Schmitt, and Lena Frischlich. Hate and counter-voices in the internet: Introduction to the special issue. *SCM Stud. Commun. Media*, 7(4):459–472, 2018.

Hannah Gaffney, David P Farrington, Dorothy L Espelage, and Maria M Ttofi. Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. *Aggress. Violent Behav.*, 45:134–153, 2019.

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

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. Analyzing the hate and counter speech accounts on Twitter. *arXiv:1812.02712*, 2018.

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

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. Vectors for counterspeech on Twitter. In *Proceedings of the first workshop on abusive language online*, pages 57–62, 2017.

Marc Ziegele, Pablo Jost, Marike Bormann, and Dominique Heinbach. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Stud. Commun. Media*, 7(4):525–554, 2018.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the COVID-19 crisis. *arXiv:2005.12423*, 2020.

Catherine Blaya. Cyberhate: A review and content analysis of intervention strategies. *Aggress. Violent Behav.*, 45: 163–172, 2019.

Anne Weber. *Manual on hate speech*. Council Of Europe, 2009.

Youtube: Hate speech policy, 2019. URL https://support.google.com/youtube/answer/2801939.

Twitter: Hateful conduct policy, 2019. URL https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Facebook: Hate speech, 2019. URL https://www.facebook.com/communitystandards/hate_speech.

Seriously, 2019. URL http://www.seriously.ong.

Social Media Helpline, 2019. URL https://socialmediahelpline.com/counterspeech-dos-and-donts-for-students/.

Jürgen Habermas. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons, 2015.

Éloi Brassard-Gourdeau and Richard Khoury. Impact of sentiment detection to recognize toxic and subversive online comments. *arXiv:1812.01704*, 2018.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online communities with computational Twitter analysis. *Technol. Forecast. Soc. Change*, 95:96–108, 2015.

Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11, 2016.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on Twitter. In *Twelfth international AAAI conference on web and social media*, 2018.

Ziqi Zhang and Lei Luo. Hate speech detection: A solved problem? the challenging case of long tail on Twitter. *Semantic Web*, 10(5):925–945, 2019.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. Overview of the evalita EVALITA hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263. CEUR, 2018.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20. Association for Computational Linguistics, 2018.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32. Association for Computational Linguistics, 2018.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8), 2019.

Shervin Malmasi and Marcos Zampieri. Challenges in discriminating profanity from hate speech. *J. Exp. Theor. Artif. Intell.*, 30(2):187–202, 2018.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.*, 48(12):4730–4742, 2018.

Areej Al-Hassan and Hmood Al-Dossari. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*, 2019.

Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *J. Inf. Technol. Politics*, 17(1):66–78, 2020.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171—-4186, 2019.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on Web and social media*, 2017.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77, 2017.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86. Association for Computational Linguistics, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12: 2825–2830, 2011.

Sylvia Jaki and Tom De Smedt. Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv:1910.07518*, 2019.

Tianqi Chen and Carlos Guestrin. xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *arXiv preprint arXiv:2009.08392*, 2020.