A Benchmark Dataset for Learning to Intervene in Online Hate Speech

Abstract

Countering online hate speech is a critical yet challenging task, but one which can be aided by the use of Natural Language Processing (NLP) techniques. Previous research has primarily focused on the development of NLP methods to automatically and effectively detect online hate speech while disregarding further action needed to calm and discourage individuals from using hate speech in the future. In addition, most existing hate speech datasets treat each post as an isolated instance, ignoring the conversational context. In this paper, we propose a novel task of generative hate speech intervention, where the goal is to automatically generate responses to intervene during online conversations that contain hate speech. As a part of this work, we introduce two fully-labeled large-scale hate speech intervention datasets collected from Gab and Reddit. These datasets provide conversation segments, hate speech labels, as well as intervention responses written by Mechanical Turk Workers. In this paper, we also analyze the datasets to understand the common intervention strategies and explore the performance of common automatic response generation methods on these new datasets to provide a benchmark for future research.

Introduction

The growing popularity of online interactions through social media has been shown to have both positive and negative impacts. While social media improves information sharing, it also facilitates the propagation of online harassment, including hate speech. These negative experiences can have a measurable negative impact on users. Recently, the Pew Research Center BIBREF0 reported that “roughly four-in-ten Americans have personally experienced online harassment, and 63% consider it a major problem.”

To address the growing problem of online hate, an extensive body of work has focused on developing automatic hate speech detection models and datasets BIBREF1, BIBREF2, BIBREF3, BIBREF4, BIBREF5, BIBREF6, BIBREF7, BIBREF8. However, simply detecting and blocking hate speech or suspicious users often has limited ability to prevent these users from simply turning to other social media platforms to continue to engage in hate speech as can be seen in the large move of individuals blocked from Twitter to Gab BIBREF9. What's more, such a strategy is often at odds with the concept of free speech. As reported by the Pew Research Center BIBREF0, “Despite this broad concern over online harassment, 45% of Americans say it is more important to let people speak their minds freely online; a slightly larger share (53%) feels that it is more important for people to feel welcome and safe online.” The special rapporteurs representing the Office of the United Nations High Commissioner for Human Rights (OHCHR) have recommended that “The strategic response to hate speech is more speech.” BIBREF10 They encourage to change what people think instead of merely changing what they do, so they advocate more speech that educates about cultural differences, diversity, and minorities as a better strategy to counter hate speech.

Therefore, in order to encourage strategies of countering online hate speech, we propose a novel task of generative hate speech intervention and introduce two new datasets for this task. Figure FIGREF5 illustrates the task. Our datasets consist of 5K conversations retrieved from Reddit and 12k conversations retrieved from Gab. Distinct from existing hate speech datasets, our datasets retain their conversational context and introduce human-written intervention responses. The conversational context and intervention responses are critical in order to build generative models to automatically mitigate the spread of these types of conversations.

To summarize, our contributions are three-fold:

We introduce the generative hate speech intervention task and provide two fully-labeled hate speech

datasets with human-written intervention responses.

Our data is collected in the form of conversations, providing better context.

The two data sources, Gab and Reddit, are not well studied for hate speech. Our datasets fill this gap.

Due to our data collecting strategy, all the posts in our datasets are manually labeled as hate or non-hate speech by Mechanical Turk workers, so they can also be used for the hate speech detection task. The performance of commonly-used classifiers on our datasets is shown in Section SECREF6.

Related Work

In recent years, a few datasets for hate speech detection have been built and released by researchers. Most are collected from Twitter and are labeled using a combination of expert and non-expert hand labeling, or through machine learning assistance using a list of common negative words. It is widely accepted that labels can vary in their accuracy overall, though this can be mitigated by relying on a consensus rule to rectify disagreements in labels. A synopsis of these datasets can be found in Table TABREF10.

BIBREF2 collect 17k tweets based on hate-related slurs and users. The tweets are manually annotated with three categories: sexist (20.0%), racist (11.7%), and normal (68.3%). Because the authors identified a number of prolific users during the initial manual search, the resulting dataset has a small number of users (1,236 users) involved, causing a potential selection bias. This problem is most prevalent on the 1,972 racist tweets, which are sent by only 9 Twitter users. To avoid this problem, we did not identify suspicious user accounts or utilize user information when collecting our data.

BIBREF3 use a similar strategy, which combines the utilization of hate keywords and suspicious user accounts to build a dataset from Twitter. But different from BIBREF2, this dataset consists of 25k tweets randomly sampled from the 85.4 million posts of a large number of users (33,458 users). This dataset is proposed mainly to distinguish hateful and offensive language, which tend to be conflated by many studies.

BIBREF11 focus on online harassment on Twitter and propose a fine-grained labeled dataset with 6 categories. BIBREF14 introduce a large Twitter dataset with 100k tweets. Despite the large size of this dataset, the ratio of the hateful tweets are relatively low (5%). Thus the size of the hateful tweets is around 5k in this dataset, which is not significantly larger than that of the previous datasets.

The dataset introduced by BIBREF12 is different from the other datasets as it investigates the behavior of hate-related users on Twitter, instead of evaluating hate-related tweets. The large majority of the 1.5k users are labeled as spammers (31.8%) or normal (60.3%). Only a small fraction of the users are labeled as bullies (4.5%) or aggressors (3.4%). While most datasets are from single sources, BIBREF13 introduce a dataset with a combination of Twitter (58.9%), Reddit, and The Guardian. In total 20,432 unique comments were obtained with 4,136 labeled as harassment (20.2%) and 16,296 as non-harassment (79.8%).

Since most of the publicly available hate speech datasets are collected from Twitter, previous research of hate speech mainly focus on Twitter posts or users BIBREF2, BIBREF17, BIBREF18, BIBREF19, BIBREF3. While there are several studies on the other sources, such as Instagram BIBREF20, Yahoo! BIBREF1, BIBREF15, and Ask.fm BIBREF16, the hate speech on Reddit and Gab is not widely studied. What's more, all the previous hate speech datasets are built for the classification or detection of hate speech from a single post or user on social media, ignoring the context of the post and intervention methods needed to effectively calm down the users and diffuse negative online conversations.

Dataset Collection ::: Ethics

Our study got approval from our Internal Review Board. Workers were warned about the offensive content before they read the data and they were informed by our instructions to feel free to quit the task at any time if they are uncomfortable with the content. Additionally, all personally identifiable information such as user names is masked in the datasets.

Dataset Collection ::: Data Filtering

Reddit: To retrieve high-quality conversational data that would likely include hate speech, we referenced the list of the whiniest most low-key toxic subreddits. Skipping the three subreddits that have been removed, we collect data from ten subreddits: r/DankMemes, r/Imgoingtohellforthis, r/KotakuInAction, r/MensRights, r/MetaCanada, r/MGTOW, r/PussyPass, r/PussyPassDenied, r/The_Donald, and r/TumblrInAction. For each of these subreddits, we retrieve the top 200 hottest submissions using Reddit's API. To further focus on conversations with hate speech in each submission, we use hate keywords BIBREF6 to identify potentially hateful comments and then reconstructed the conversational context of each comment. This context consists of all comments preceding and following a potentially hateful comment. Thus for each potentially hateful comment, we rebuild the conversation where the comment appears. Figure FIGREF14 shows an example of the collected conversation, where the second comment contains a hate keyword and is considered as potentially hateful. Because a conversation may contain more than one comments with hate keywords, we removed any duplicated conversations.

Gab: We collect data from all the Gab posts in October 2018. Similar to Reddit, we use hate keywords BIBREF6 to identify potentially hateful posts, rebuild the conversation context and clean duplicate conversations.

Dataset Collection ::: Crowd-Sourcing

After we collected the conversations from both Reddit and Gab, we presented this data to Mechanical Turk workers to label and create intervention suggestions. In order not to over-burden the workers, we filtered out conversations consisting of more than 20 comments. Each assignment consists of 5 conversations. For Reddit, we also present the title and content of the corresponding submission in order to give workers more information about the topic and context. For each conversation, a worker is asked to answer two questions:

Q1: Which posts or comments in this conversation are hate speech?

Q2: If there exists hate speech in the conversation, how would you respond to intervene? Write down a response that can probably hold it back (word limit: 140 characters).

If the worker thinks no hate speech exists in the conversation, then the answers to both questions are “n/a”. To provide context, the definition of hate speech from Facebook: “We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.” is presented to the workers. Also, to prevent workers from using hate speech in the response or writing responses that are too general, such as “Please do not say that”, we provide additional instructions and rejected examples.

Dataset Collection ::: Data Quality

Each conversation is assigned to three different workers. To ensure data quality, we restrict the workers to be in an English speaking country including Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States, with a HIT approval rate higher than 95%. Excluding the rejected answers, the collected data involves 926 different workers. The final hate speech labels (answers to Q1) are aggregated according to the majority of the workers' answers. A comment is considered hate speech only when at least two out of the three workers label it as hate speech. The responses (answers to Q2) are aggregated according to the aggregated result of Q1. If the worker's label to Q1 agrees with the aggregated result, then their answer to Q2 is included as a candidate response to the corresponding conversation but is otherwise disregarded. See Figure FIGREF14 for an example of the aggregated data.

Dataset Analysis ::: Statistics

From Reddit, we collected 5,020 conversations, including 22,324 comments. On average, each conversation consists of 4.45 comments and the length of each comment is 58.0 tokens. 5,257 of the comments are labeled as hate speech and 17,067 are labeled as non-hate speech. A majority of the conversations, 3,847 (76.6%), contain hate speech. Each conversation with hate speech has 2.66 responses on average, for a total of 10,243 intervention responses. The average length of the intervention responses is 17.96 tokens.

From Gab, we collected 11,825 conversations, consisting of 33,776 posts. On average, each conversation consists of 2.86 posts and the average length of each post is 35.6 tokens. 14,614 of the posts are labeled as hate speech and 19,162 are labeled as non-hate speech. Nearly all the conversations, 11,169 (94.5%), contain hate speech. 31,487 intervention responses were originally collected for conversations with hate speech, or 2.82 responses per conversation on average. The average length of the intervention responses is 17.27 tokens.

Compared with the Gab dataset, there are fewer conversations and comments in the Reddit dataset, comments and conversations are longer, and the distribution of hate and non-hate speech labels is more imbalanced. Figure FIGREF20 illustrates the distributions of the top 10 keywords in the hate speech collected from Reddit and Gab separately. The Gab dataset and the Reddit dataset have similar popular hate keywords, but the distributions are very different. All the statistics shown above indicate that the characteristics of the data collected from these two sources are very different, thus the challenges of doing detection or generative intervention tasks on the dataset from these sources will also be different.

Dataset Analysis ::: Intervention Strategies

Removing duplicates, there are 21,747 unique intervention responses in the aggregated Gab dataset and 7,641 in the aggregated Reddit dataset. Despite the large diversity of the collected responses for intervention, we find workers tend to have certain strategies for intervention.

Identify Hate Keywords: One of the most common strategies is to identify the inappropriate terms in the post and then urge the user to stop using that work. For example, “The C word and language attacking gender is unacceptable. Please refrain from future use.” This strategy is often used when the hatred in the post is mainly conveyed by specific hate keywords.

Categorize Hate Speech: This is another common strategy used by the workers. The workers classify hate speech into different categories, such as racist, sexist, homophobic, etc. This strategy is often combined with identifying hate keywords or targets of hatred. For example, “The term ""fa**ot"" comprises

homophobic hate, and as such is not permitted here.”

Positive Tone Followed by Transitions: This is a strategy where the response consists of two parts combined with a transitional word, such as “but” and “even though”. The first part starts with affirmative terms, such as “I understand”, “You have the right to”, and “You are free to express”, showing kindness and understanding, while the second part is to alert the users that their post is inappropriate. For example, “I understand your frustration, but the term you have used is offensive towards the disabled community. Please be more aware of your words.”. Intuitively, compared with the response that directly warns, this strategy is likely more acceptable for the users and be more likely to clam down a quarrel full of hate speech.

Suggest Proper Actions: Besides warning and discouraging the users from continuing hate speech, workers also suggest the actions that the user should take. This strategy can either be combined with other strategies mentioned above or be used alone. In the latter case, a negative tone can be greatly alleviated. For example, “I think that you should do more research on how resources are allocated in this country.”

Generative Intervention

Our datasets can be used for various hate speech tasks. In this paper, we focus on generative hate speech intervention.

The goal of this task is to generate a response to hate speech that can mitigate its use during a conversation. The objective can be formulated as the following equation:

where c is the conversation, r is the corresponding intervention response, and D is the dataset. This task is closely related to the response generation and dialog generation, though several differences exist including dialog length, language cadence, and word imbalances. As a baseline, we chose the most common methods of these two tasks, such as Seq2Seq and VAE, to determine the initial feasibility of automatically generate intervention responses. More recent Reinforcement Learning method for dialog generation BIBREF21 can also be applied to this task with slight modification. Future work will explore more complex, and unique models.

Similar to BIBREF21, a generative model is considered as an agent. However, different from dialog generation, generative intervention does not have multiple turns of utterance, so the action of the agent is to select a token in the response. The state of the agent is given by the input posts and the previously generated tokens. Another result due to this difference is that the rewards with regard to ease of answering or information flow do not apply to this case, but the reward for semantic coherence does. Therefore, the reward of the agent is:

where $rw(c,r)$ is the reward with regard to the conversation c and its reference response r in the dataset. $p(r|c)$ denotes the probability of generating response r given the conversation c, and $p_{back}(c|r)$ denotes the backward probability of generating the conversation based on the response, which is parameterized by another generation network. The reward is a weighted combination of these two parts, which are observed after the agent finishing generating the response. We refer the readers to BIBREF21 for details.

Experiments

We evaluate the commonly-used detection and generation methods with our dataset. Due to the different characteristics of the data collected from the two sources (Section SECREF4), we treat them as two independent datasets.

Experiments ::: Experimental Settings

For binary hate speech detection, we experimented the following four different methods.

Logistic Regression (LR): We evaluate the Logistic Regression model with L2 regularization. The penalty parameter C is set to 1. The input features are the Term Frequency Inverse Document Frequency (TF-IDF) values of up to 2-grams.

Support Vector Machine (SVM): We evaluate the SVM model with linear kernels. We use L2 regularization and the coefficient is 1. The features are the same as in LR.

Convolutional Neural Network (CNN): We use the CNN model for sentence classification proposed by BIBREF22 with default hyperparameters. The word embeddings are randomly initialized (CNN in Table TABREF27) or initialized with pretrained Word2Vec BIBREF23 embeddings on Google News (CNN$^\ast $ in Table TABREF27).

Recurrent Neural Network (RNN): The model we evaluated consists of 2-layer bidirectional Gated Recurrent Unit (GRU) BIBREF24 followed by a linear layer. Same as for CNN, we report the performance of RNN with two different settings of the word embeddings.

The methods are evaluated on testing data randomly selected from the dataset with the ratio of 20%. The input data is not manipulated to manually balance the classes for any of the above methods. Therefore, the training and testing data retain the same distribution as the collected results (Section SECREF4). The methods are evaluated using F-1 score, Precision-Recall (PR) AUC, and Receiver-Operating-Characteristic (ROC) AUC.

For generative hate speech intervention, we evaluated the following three methods.

Seq2Seq BIBREF25, BIBREF24: The encoder consists of 2 bidirectional GRU layers. The decoder consists of 2 GRU layers followed by a 3-layer MLP (Multi-Layer Perceptron).

Variational Auto-Encoder (VAE) BIBREF26: The structure of the VAE model is similar to that of the Seq2Seq model, except that it has two independent linear layers followed by the encoder to calculate the mean and variance of the distribution of the latent variable separately. We assume the latent variable follows a multivariate Gaussian Distribution. KL annealing BIBREF27 is applied during training.

Reinforcement Learning (RL): We also implement the Reinforcement Learning method described in Section SECREF5. The backbone of this model is the Seq2Seq model, which follows the same Seq2Seq network structure described above. This network is used to parameterize the probability of a response given the conversation. Besides this backbone Seq2Seq model, another Seq2Seq model is used to generate the backward probability. This network is trained in a similar way as the backbone Seq2Seq model, but with a response as input and the corresponding conversation as the target. In our implementation, the function of the first part of the reward ($\log p(r|c)$) is conveyed by the MLE loss. A curriculum learning strategy is adopted for the reward of $\log p_{back}(c|r)$ as in BIBREF28. Same as in BIBREF21 and BIBREF28, a baseline strategy is employed to estimate the average reward. We parameterize it as a 3-layer MLP.

The Seq2Seq model and VAE model are evaluated under two different settings. In one setting, the input for the generative model is the complete conversation, while in the other setting, the input is the filtered conversation, which only includes the posts labeled as hate speech. The filtered conversation was necessary to test the Reinforcement Learning model, as it is too challenging for the backward model to reconstruct the complete conversation based only on the intervention response.

In our experiments on the generative hate speech intervention task, we do not consider conversations without hate speech. The testing dataset is then randomly selected from the resulting dataset with the ratio of 20%. Since each conversation can have multiple reference responses, we dis-aggregate the responses and construct a pair (conversation, reference response) for each of the corresponding references during training. Teacher forcing is used for each of the three methods. The automatic evaluation metrics include BLEU BIBREF29, ROUGE-L BIBREF30, and METEOR BIBREF31.

In order to validate and compare the quality of the generated results from each model, we also conducted human evaluations as previous research has shown that automatic evaluation metrics often do not correlate with human preference BIBREF32. We randomly sampled 450 conversations from the testing dataset. We then generated responses using each of the above models trained with the filtered conversation setting. In each assignment, a Mechanical Turk worker is presented 10 conversations, along with corresponding responses generated by the three models. For each conversation, the worker is asked to evaluate the effectiveness of the generated intervention by selecting a response that can best mitigate hate speech. 9 of the 10 questions are filled with the sampled testing data and the generated results, while the other is artificially constructed to monitor response quality. After selecting the 10 best mitigation measures, the worker is asked to select which of the three methods has the best diversity of responses over all the 10 conversations. Ties are permitted for answers. Assignments failed on the quality check are rejected.

Experiments ::: Experimental Results and Discussion

The experimental results of the detection task and the generative intervention task are shown in Table TABREF27 and Table TABREF29 separately. The results of the human evaluation are shown in Table TABREF30. Figure FIGREF25 shows examples of the generated responses.

As shown in Table TABREF27 and TABREF29, all the classification and generative models perform better on the Gab dataset than on the Reddit dataset. We think this stems from the datasets' characteristics. First, the Gab dataset is larger and has a more balanced category distribution than the Reddit dataset. Therefore, it is inherently more challenging to train a classifier on the Reddit dataset. Further, the average lengths of the Reddit posts and conversations are much larger than those of Gab, potentially making the Reddit input nosier than the Gab input for both tasks. On both the Gab and Reddit datasets, the SVM classifier and the LR classifier achieved better performance than the CNN and RNN

model with randomly initialized word embeddings. A possible reason is that without pretrained word embeddings, the neural network models tend to overfit on the dataset.

For the generative intervention task, three models perform similarly on all three automatic evaluation metrics. As expected, the Seq2Seq model achieves higher scores with filtered conversation as input. However, this is not the case for the VAE model. This indicates that the two models may have different capabilities to capture important information in conversations.

As shown in Table TABREF29, applying Reinforcement Learning does not lead to higher scores on the three automatic metrics. However, human evaluation (Table TABREF30) shows that the RL model creates responses that are potentially better at mitigating hate speech and are more diverse, which is consistent with BIBREF21. There is a larger performance difference with the Gab dataset, while the effectiveness and the diversity of the responses generated by the Seq2Seq model and the RL model are quite similar on the Reddit dataset. One possible reason is that the size of the training data from Reddit (around 8k) is only 30% the size of the training data from Gab. The inconsistency between the human evaluation results and the automatic ones indicates the automatic evaluation metrics listed in Table TABREF29 can hardly reflect the quality of the generated responses. As mentioned in Section SECREF4, annotators tend to have strategies for intervention. Therefore, generating the common parts of the most popular strategies for all the testing input can lead to high scores of these automatic evaluation metrics. For example, generating “Please do not use derogatory language.” for all the testing Gab data can achieve 4.2 on BLEU, 20.4 on ROUGE, and 18.2 on METEOR. However, this response is not considered as high-quality because it is almost a universal response to all the hate speech, regardless of the context and topic.

Surprisingly, the responses generated by the VAE model have much worse diversity than the other two methods according to human evaluation. As indicated in Figure FIGREF25, the responses generated by

VAE tend to repeat the responses related to some popular hate keyword. For example, “Use of the r-word is unacceptable in our discourse as it demeans and insults people with mental disabilities.” and “Please do not use derogatory language for intellectual disabilities.” are the generated responses for a large part of the Gab testing data. According to Figure FIGREF20, insults towards disabilities are the largest portion in the dataset, so we suspect that the performance of the VAE model is affected by the imbalanced keyword distribution.

The sampled results in Figure FIGREF25 show that the Seq2Seq and the RL model can generate reasonable responses for intervention. However, as is to be expected with machine-generated text, in the other human evaluation we conducted, where Mechanical Turk workers were also presented with sampled human-written responses alongside the machine generated responses, the human-written responses were chosen as the most effective and diverse option a majority of the time (70% or more) for both datasets. This indicates that there is significant room for improvement while generating automated intervention responses.

In our experiments, we only utilized the text of the posts, but more information is available and can be utilized, such as the user information and the title of a Reddit submission.

Conclusion

Towards the end goal of mitigating the problem of online hate speech, we propose the task of generative hate speech intervention and introduce two fully-labeled datasets collected from Reddit and Gab, with crowd-sourced intervention responses. The performance of the three generative models: Seq2Seq, VAE, and RL, suggests ample opportunity for improvement. We intend to make our dataset freely available to facilitate further exploration of hate speech intervention and better models for generative intervention.

Acknowledgments

This research was supported by the Intel AI Faculty Research Grant. The authors are solely responsible for the contents of the paper and the opinions expressed in this publication do not reflect those of the funding agencies.