

# A Benchmark Dataset for Learning to Intervene in Online Hate Speech

Jing Qian<sup>†</sup>, Anna Bethke\*, Yinyin Liu\*, Elizabeth Belding<sup>†</sup>, William Yang Wang<sup>†</sup>

<sup>†</sup> University of California, Santa Barbara

\* Intel AI

{jing.qian, ebelding, william}@cs.ucsb.edu

{anna.bethke, yinyin.liu}@intel.com

## 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<sup>1</sup> collected from Gab<sup>2</sup> and Reddit<sup>3</sup>. These datasets provide conversation segments, hate speech labels, as well as intervention responses written by Mechanical Turk<sup>4</sup> 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.

## 1 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

<sup>1</sup><https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech>

<sup>2</sup><https://gab.ai>

<sup>3</sup><https://www.reddit.com>

<sup>4</sup><https://www.mturk.com>

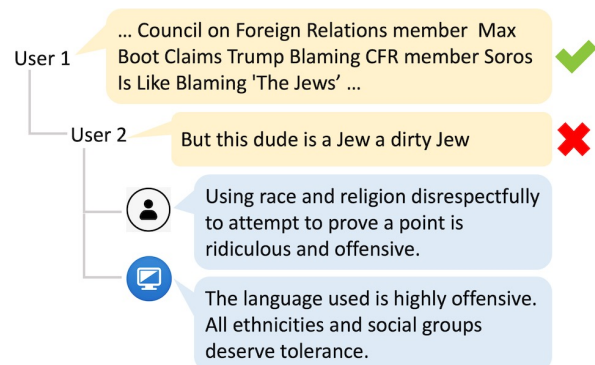


Figure 1: An illustration of hate speech conversation between User 1 and User 2 and the interventions collected for our datasets. The check and the cross icons on the right indicate a normal post and a hateful post. The utterance following the human icon is a human-written intervention, while the utterance following the computer icon is machine-generated.

the propagation of online harassment, including hate speech. These negative experiences can have a measurable negative impact on users. Recently, the Pew Research Center (Center, 2017) 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 (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Davidson et al., 2017; Schmidt and Wiegand, 2017; ElSherief et al., 2018a,b; Qian et al., 2018a,b). 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<sup>5</sup> to Gab (Ohlheiser, 2016). What’s more, such a

<sup>5</sup><https://www.twitter.com>

strategy is often at odds with the concept of free speech. As reported by the Pew Research Center (Center, 2017), “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.” (Bielefeldt et al., 2011) 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 1 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 6.

## 2 Related Work

In recent years, a few datasets for hate speech detection have been built and released by re-

searchers. 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 1.

Waseem and Hovy (2016) 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.

Davidson et al. (2017) use a similar strategy, which combines the utilization of hate keywords and suspicious user accounts to build a dataset from Twitter. But different from Waseem and Hovy (2016), 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.

Golbeck et al. (2017) focus on online harassment on Twitter and propose a fine-grained labeled dataset with 6 categories. Founta et al. (2018) 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 Chatzakou et al. (2017) 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, Kennedy III et al. (2017) introduce a dataset with

	Source	#Posts	Conv.	Categories	Interv.
Waseem and Hovy (2016)	Twitter	17k	No	racist, sexist, normal	No
Davidson et al. (2017)	Twitter	25k	No	hateful, offensive, neither	No
Golbeck et al. (2017)	Twitter	35k	No	the worst, threats, hate speech, direct harassment, potentially offensive, non-harassment	No
Chatzakou et al. (2017)	Twitter	9k	No	aggressive, bullying, spam, normal	No
Kennedy III et al. (2017)	Twitter, Reddit, The Guardian	20k	No	harassment, non-harassment	No
Founta et al. (2018)	Twitter	100k	No	abusive, hateful, normal, spam	No
Warner and Hirschberg (2012)	Yahoo!	9k	No	anti-semitic, anti-black, anti-asian, anti-woman, anti-muslim, anti-immigrant, other-hate	No
Nobata et al. (2016)	Yahoo!	2k	No	clean, hate, derogatory, profanity	No
Van Hee et al. (2015)	Ask.fm	85k	No	threat/blackmail, insult, defamation, sexual talk, curse/exclusion, defense, encouragement to the harasser	No
Ours	Reddit	22k	Yes	hate, non-hate	Yes
Ours	Gab	34k	Yes	hate, non-hate	Yes

Table 1: Comparison of our datasets with previous hate speech datasets. Conv.: Conversation. Interv.: Intervention.

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 (Waseem and Hovy, 2016; Gao et al., 2017; Burnap and Williams, 2016; Badjatiya et al., 2017; Davidson et al., 2017). While there are several studies on the other sources, such as Instagram (Zhong et al., 2016), Yahoo! (Warner and Hirschberg, 2012; Nobata et al., 2016), and Ask.fm (Van Hee et al., 2015), 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.

### 3 Dataset Collection

#### 3.1 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.

#### 3.2 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<sup>6</sup>. 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/TheDonald*, 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 (ElSherief et al., 2018b) 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 2 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 (ElSherief et al., 2018b) to identify potentially hateful posts, rebuild the conversation context and clean duplicate conversations.

<sup>6</sup>[https://www.vice.com/en\\_us/article/8xxymb/here-are-reddits-whiniest-most-low-key-toxic-subreddits](https://www.vice.com/en_us/article/8xxymb/here-are-reddits-whiniest-most-low-key-toxic-subreddits)

Conversation	Hate Speech	Human-Written Intervention Responses
1. User 1: United Kingdom: 'Schoolboy, 15, given detention for backing UKIP during classroom debate' 2. User 2: The education system is full of re***ds! Yes, most school teachers are ret***ed lefties! Teach your children to laugh at these ret***ed lefties! 3. User 3: Asking a teacher to not be a leftist is like asking a medieval monk to question the Pope. 4. User 4: The Jews are like Sjws, they infest everything.	2, 4	➤ Use of this language is not tolerated and it is uncalled for. ➤ Use of the slurs and insults here is unacceptable in our discourse as it demeans and insults and alienates others. ➤ I recommend that you research the holocaust, you might change your opinion.

Figure 2: An example of the aggregated data. The first column is the conversation text. Indexes are added to each post. Indentations before each post indicate the structure of replies. The second column is the indexes of the human-labeled hateful post. Each bullet point in the third column is a human-written response.

### 3.3 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<sup>7</sup>: “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.

### 3.4 Data Quality

Each conversation is assigned to three different workers. To ensure data quality, we restrict the

<sup>7</sup>[https://m.facebook.com/communitystandards/hate\\_speech/](https://m.facebook.com/communitystandards/hate_speech/)

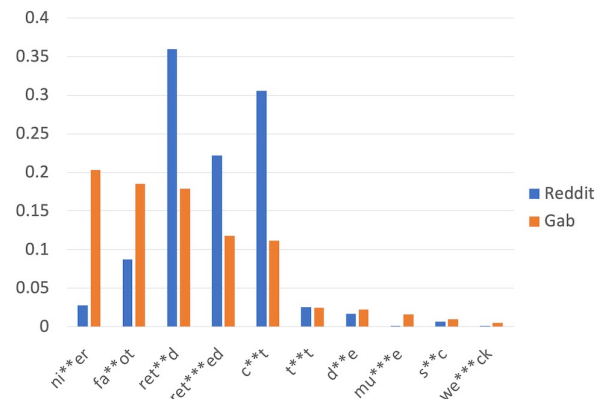


Figure 3: The distributions of the top 10 keywords in the hate speech collected from Reddit and Gab. Hate keywords are masked.

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 2 for an example of the aggregated data.

## 4 Dataset Analysis

### 4.1 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 3 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.

## 4.2 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 word. 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.*”

## 5 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:

$$Obj = \max \sum_{(c,r) \in D} \log p(r|c) \quad (1)$$

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 (Li et al., 2016) can also be applied to this task with slight modification. Future work will explore more complex, and unique models.

Similar to (Li et al., 2016), 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:

$$rw(c, r) = \lambda_1 \log p(r|c) + \lambda_2 \log p_{back}(c|r) \quad (2)$$

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 Li et al. (2016) for details.

## 6 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 4), we treat them as two independent datasets.

### 6.1 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 Kim (2014) with default hyperparameters. The word embeddings are randomly initialized (CNN in Table 2) or initialized with pre-trained Word2Vec (Mikolov et al., 2013) embeddings on Google News (CNN\* in Table 2).

**Recurrent Neural Network (RNN):** The model we evaluated consists of 2-layer bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014) 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 4). 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 (Sutskever et al., 2014; Cho et al., 2014):** 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) (Kingma and Welling, 2013):** 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 (Bowman et al., 2016) is applied during training.

**Reinforcement Learning (RL):** We also implement the Reinforcement Learning method described in Section 5. 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

Conversation	Seq2Seq	VAE	RL
User 1: Ok. Keep White Knighting for a sand ni**er who f**ks tons of White women. You're entitled to your opinion.	This is is not allowed and un time to treat people by their skin color	Use of the r-word is unacceptable in our discourse as it demeans and insults people with mental disabilities	I do not use derogatory racial slurs
User 1: Feminists are f**king ret***ed, why are they so pissed about manspreading shit makes no difference in space. I'm gonna start throwing things in womens cleavage as they walk by.	Using talking about someones character to describe a woman is insulting to hateful	I don't think using words that are sexist in nature contribute to a productive conversation	Your post uses speech that denigrates sex and gender, which is in violation of our Community Guidelines and TOS Agreement
User 1: Only a nig****ick could reach that c**t User 2: I bet a midget football team could hide under her overhang during a rainstorm and not get wet. User 3: Ni**er probably uses her as a waterbed	Use of the c-word is unacceptable in our discourse as it demeans and insults women	Please do not use derogatory language for intellectual disabilities	If you don't agree with you, there 's no need to resort to name calling

Figure 4: Examples of the generated intervention responses. The hateful terms in the conversation are masked.

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 Ranzato et al. (2016). Same as in Li et al. (2016) and Ranzato et al. (2016), 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 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

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 (Paulus et al., 2018). We randomly sampled 450 conversations from the testing dataset. We then gen-

Dataset	Gab			Reddit		
Metric	F1	PR	ROC	F1	PR	ROC
LR	88.2	94.5	95.4	64.7	80.4	91.4
SVM	88.6	94.7	95.6	75.7	<b>81.1</b>	<b>92.0</b>
CNN	87.5	92.8	92.6	74.8	76.8	87.5
RNN	87.6	93.9	94.2	71.7	76.1	88.6
CNN*	<b>89.6</b>	<b>95.2</b>	<b>95.8</b>	76.9	80.1	90.9
RNN*	89.3	94.8	95.5	<b>77.5</b>	79.4	90.6

Table 2: Experimental results for the detection task. PR is Precision-Recall AUC and ROC is ROC AUC. The models marked with \* use pretrained Word2Vec embeddings. The best results are in bold.

erated 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.

## 6.2 Experimental Results and Discussion

The experimental results of the detection task and the generative intervention task are shown in Table 2 and Table 3 separately. The results of the human evaluation are shown in Table 4. Figure 4 shows examples of the generated responses.

As shown in Table 2 and 3, all the classification

Dataset	Gab						Reddit					
Inp. Set.	Complete			Filtered			Complete			Filtered		
Metric	B	R	M	B	R	M	B	R	M	B	R	M
Seq2Seq	<b>13.2</b>	<b>33.8</b>	23.0	<b>15.0</b>	<b>34.2</b>	23.6	5.5	<b>29.5</b>	19.5	5.9	28.2	20.0
VAE	12.2	32.5	<b>23.4</b>	12.4	32.8	21.8	<b>6.8</b>	29.0	<b>20.2</b>	<b>7.0</b>	<b>29.1</b>	<b>20.1</b>
RL	-	-	-	14.5	33.1	<b>23.9</b>	-	-	-	4.4	<b>29.1</b>	18.7

Table 3: Experimental results for generative intervention task. Inp. Set.: Input Setting (Section 6.1). B: BLEU. R: ROUGE-L. M: METEOR. Best results are in bold.

Dataset	Gab		Reddit	
Metric	Eff.	Div.	Eff.	Div.
Seq2Seq Wins	22.4	28.0	<b>31.1</b>	<b>34.0</b>
VAE Wins	20.0	6.0	26.0	4.0
RL Wins	<b>41.6</b>	<b>40.0</b>	30.0	30.0
Tie	16.0	26.0	12.9	32.0

Table 4: Human evaluation results. Table values are the percentage of the answers. Eff.: Effectiveness, evaluates how well the generated responses can mitigate hate speech. Div: Diversity, evaluates how many different responses are generated. Best results are in bold.

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 noisier 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 3, applying Reinforcement Learning does not lead to higher scores on the three automatic metrics. However, human evaluation (Table 4) shows that the RL model creates responses that are potentially better at mitigating hate speech and are more diverse, which is consistent with Li et al. (2016). There is a larger per-

formance 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 3 can hardly reflect the quality of the generated responses. As mentioned in Section 4, 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 4, 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 3, 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 4 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.

## 7 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.

## References

- Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Heiner Bielefeldt, Frank La Rue, and Githu Muigai. 2011. Ohchr expert workshops on the prohibition of incitement to national, racial or religious hatred. *Expert workshop on the Americas*.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21.
- Pete Burnap and Matthew L Williams. 2016. Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics. *EPJ Data Science*, 5(1):11.
- Pew Research Center. 2017. Online harassment 2017.
- Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM.
- Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.
- Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018b. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.
- Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.
- Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 774–782.

- Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233. ACM.
- George W. Kennedy III, Andrew W. McCollough, Edward Dixon, Alexie Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1192–1202.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.
- Abby Ohlheiser. 2016. Banned from twitter? this site promises you can say whatever you want. *The Washington Post*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018*.
- Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018a. Hierarchical cvae for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559.
- Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018b. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 118–123.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*.
- Anna Schmidt and Michael Wiegand. 2017. 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.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.
- Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and Fine-grained Classification of Cyberbullying Events. In *RANLP’15: International Conference Recent Advances in Natural Language Processing*, pages 672–680.
- William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *ACL’12: Proceedings of the 2nd Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.
- Zeeraq Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *IJCAI’16: Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3952–3958.