# Towards Automatic Online Hate Speech Intervention Generation using Pretrained Language Model

**Raj Ratn Pranesh**[1], **Ambesh Shekhar**[1], **Anish Kumar**[1]
[1]Birla Institute of Technology, Mesra
Ranchi, India
{raj.ratn18, ambesh.sinha, anishkr10052}@gmail.com

## Abstract

Social media harbors substantial toxic and hateful conversations today. Curbing them has emerged as a critical challenge for governments and organizations globally. Prior research has primarily concentrated on the detection of online hate speech while ignoring further action needed to discourage individuals from using hate speech in the future. Counter-speech is an effective way to tackle online hate, leaving freedom of speech untouched. The focus is to directly intervene in the conversation with textual responses that counter the hate content and prevent it from further spreading. In this paper, we propose a novel natural language generation task for hate speech intervention, where the goal is to automatically generate responses to intervene during online conversations that contain hate speech. We sequentially analyzed the performance and capability of various state-of-the-art pretrained language models dialogue generation model for automated hate speech intervention system using automatic metric and manual human evaluation. The results indicates that the generated intervention responses are very promising in terms of relevance and contextual meaning[1]

## Introduction

The growing popularity of online interactions through social media can be attributed to ease of information and opinion sharing with large masses in almost no time. While social media is invaluable in those aspects, it also paves the way for the propagation of online harassment, including hate speech. These negative experiences can hurt users mentally and emotionally. The increasing spread of hate speech through social media has drawn the attention of various governments and organizations. In May 2016, the European Commission agreed with Facebook, Microsoft, Twitter, and YouTube a code of conduct (Jourová 2016) on countering illegal hate speech online. Several other large companies have joined the code of conduct later.

One of the techniques that these organizations use to curb online hate speech is suspending or blocking the message or the user account itself. This method requires the identification of such user accounts. Thus, previous research works have focused on hate speech detection to address the growing problem of online hate ((Warner and Hirschberg 2012); (Waseem and Hovy 2016); (Waseem et al. 2017); (Schmidt and Wiegand 2017); (ElSherief et al. 2018a), (ElSherief et al. 2018b); (Qian et al. 2018a), (Qian et al. 2018b)). This standard approach enables the prevention of online hate spread by suspending these user accounts or deleting the hate comments from the social media platforms. However, identifying hate speech or suspicious users is not enough to prevent such recurring incidents. Such users are likely to do the same using other accounts.

The countries and organizations pondered over the countering of hate speech as an alternative to blocking (Gagliardone et al. 2015). Thus, few Non-Governmental Organizations (NGOs) had trained operators to intervene in online hate conversations by writing counter-narratives. A Counter-Narrative (CN) is a non-aggressive response that offers feedback through fact-bound arguments and is considered as the most effective approach to withstand hate messages (Benesch 2014); (Schieb and Preuss 2016). Still, this approach was not scalable. Hence, research in the field of automatic counter-narratives generation becomes a necessity. This approach is faster and scalable, more flexible, and responsive and capable of dealing with extremism from anywhere and in any language. It does not form a barrier against the principle of free and open public space for debate. Therefore, we propose a data-driven natural language generative approach for hate speech intervention to encourage strategies of countering online hate speech. We utilized 3 large pretrained language models namely- BART(Lewis et al. 2019), DialoGPT(Zhang et al. 2019) and BERT(Devlin et al. 2018) for building the jointly trainable encoder-decoder based automatic hate speech intervention generation models. We used various automatic performance evaluation metrics widely used for machine translation tasks as well as we used manual human evaluation for more accurate exploration and in-depth analysis. In our experiment, the models demonstrated good potential as the machine-generated intervention responses were very similar to the human-generated responses.

Rest of the paper is structured in the following manner: (i) section *Related Work* summarises the recent work done in the area of hate speech and mitigation, (ii) section *Dataset* discuses about the dataset used in detail, (iii) section *Meth-*



[1]This paper contains highly offensive words and phrases. They do not reflect the views of any of the authors and are intended to be purely demonstrative

*ods* talks about the various state-of-the-art models used in our work, (iv) section *Experiment* systematically elaborates the experiment setup, (v) section *Results and Discussion* provides analysis and insights about the models performance and lastly (vi) section *Conclusion and Future Work* talks about future work and concludes the paper.

## Related Work

In recent years, several datasets have been collected from various social media platforms such as Twitter, Facebook, Reddit, etc. for hate speech detection. Most of them combine the utilization of hate keywords and suspicious user accounts to build the dataset. (Waseem and Hovy 2016) collected 17k tweets based on hate-related slurs, (Waseem et al. 2017) used a similar strategy to extract 25k tweets from 85.4 million posts, sampled across a wider user-base (33,458 users). (Golbeck et al. 2017) focuses on online harassment on Twitter and proposed a fine-grained labeled dataset with 6 categories. (Founta et al. 2018) introduced a large Twitter dataset with 100k tweets, but with a relatively low (5%) ratio of hateful tweets. (Kennedy et al. 2017) introduced 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%). All these hate speech datasets were used for the classification or detection, employing features such as lexical resources (Gitari et al. 2015); (Burnap and Williams 2016), sentiment polarity (Burnap and Williams 2015) and multimodal information (Hosseinmardi et al. 2015) to a classifier. But these datasets ignore the context of the post and intervention methods required to pacify the users effectively. Several counter-narrative methods to counter hatred have been outlined and tested by (Benesch 2014), (Munger 2017), and (Mathew et al. 2019) since then. (Qian et al. 2019) proposed a dataset of hate speech conversations collected from Reddit and Gab, manually annotated. This dataset consists of labels, intervention suggestions, along with the conversations for the model to understand the context.

## Dataset

For the novel natural language generative task, we use A Benchmark Dataset for Learning to Intervene in Online Hate Speech, introduced by (Qian et al. 2019). It provides conversation segments, hate-speech labels, as well as intervention responses, written by Mechanical Turk Workers. The high-quality conversational data that include hate speech were retrieved from Reddit using the list of the whiniest most low-key toxic subreddits. The ten subreddits were: r/DankMemes, r/Imgoingtohellforthis, r/KotakuInAction, r/MensRights, r/MetaCanada, r/MGTOW, r/PussyPass, r/PussyPassDenied, r/The Donald, and r/TumblrInAction. The top 200 hottest submissions were retrieved for each of these subreddits using Reddit's API. Similar technique was used for Gab as well. To further focus on conversations with hate speech in each submission, hate keywords (ElSherief et al. 2018b) were used to identify potentially hateful comments and then reconstruct the conversational context of each comment. This context consists of all comments preceding and following a potentially hateful comment. This collection allows the model to understand the context before the generation of counter-hate comments. The dataset contains 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. Also, the dataset contains 11,825 conversations,consisting of 33,776 posts, collected from Gab. 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. There are 7,641 unique intervention responses in the aggregated Reddit dataset and 21,747 in the aggregated Gab dataset. The workers had certain strategies for intervention. The intervention strategies include identifying hate keywords and warning users not to use that word. For example, "The C-word and language attacking gender are unacceptable. Please refrain from future use." This strategy was often used when the hatred in the post is mainly conveyed by specific hate keywords. The other strategy was to categorize hate speech into different categories, such as racist, sexist, homophobic, etc. This strategy was 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." Lastly, a positive tone followed by transitions was used as one of the strategies. The first part would start 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 was used 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 was likely more acceptable for the users and be more likely to clam down a quarrel full of hate speech. Also, the workers would suggest proper actions as a strategy to curb the spread of online hate speech. For example, "I think that you should do more research on how resources are allocated in this country."

## Methods

In this section, we provided an detailed overview of existing well-established and state-of-the-art dialogue generation models. For our experiment, we used 3 deep learning encoder-decoder based models, i.e., BART(Lewis et al. 2019), DialoGPT(Zhang et al. 2019) and BERT(Devlin et al.

2018). We utilized the available Huggingface[2] pretrained models of BERT, DialoGPT and BART which were fine-tuned on our dataset.

For a given pair of conversation segment in the dataset consisting of an alternating sequence of hate speech and it's associated human generated intervention responses, i.e. (i) $D_1$: online user-generated hate speech and (ii) $D_2$: human-generated hate speech intervention response. So, the utterance pair of $\{D_1, D_2\}$ is used for training all of our dialogue generation models. Given an input $D_1$, the dialogue generation model outputs $D_2$.

## DialoGPT

In the paper(Radford et al. 2018), the author proposed a transformer based language model- GPT. Given a token sequence $x_1, ..., x_n$, in the language model the sequence probability was defined as: $p(x_1, ..., x_n) = \Pi_{i=2}^{n} p(x_i|x_1, ..., x_{i-1})$, where next token are predicted through the historical sequences. For GPT, transformer decoder was used to define $p(x_i|x_1, ..., x_{i-1})$. The decoder consists of stacked self-attention feed-forward layers(each accompanied by normalization layer) for encoding $x_1, ..., x_{i-1}$ and which was then used to predict $x_i$. As an improvement over GPT, author in paper(Radford et al. 2019) proposed GPT-2, a normalization layer was assigned to each of the sub-blocks input. Also an extra normalization layer was incorporated immediately next to the last self-attention block.

In our experiment, we used DialoGPT(Zhang et al. 2019) which was a GPT-2 based model trained on a very large corpus consisting of English Reddit dialogues. The corpus was consist of 147,116,725 instances of dialogues, collected over a period of 12 years. The model takes the dialogue utterances history $S$ and ground truth response $T = x_1, ..., x_n$, the DialoGPT model aims at maximizing the probability:$p(T|S) = p(x_1|S)\Pi_{i=2}^{n} p(x_i|S, x_1, ..., x_{i-1})$, where the transformer model defines the conditional probabilities. Through a maximum mutual information (MMI) function(Li et al. 2015), the model also gets penalized for generating uninteresting responses. In our experiment, we used $DialoGPT_{\text{small}}$ with 117 million weight parameters.

## BART

BART(Lewis et al. 2019) is a denoising autoencoder used in the pretraining of sequence-to-sequence models. It maps a corrupted document to the original document it was derived from, using a bidirectional encoder over a corrupted text and a left-to-right autoregressive decoder. The mask attention mechanism facilitates the training on sequence from left to right, generating texts based on the left part of the sequence.

We create a BART language model wrapper for this transformer-based dialogue system employing the API of the BART-large model from hugging face-transformers. This pretrained model has 400M trainable parameters with 6 encoding and decoding layers in each block, alongside 16 attention heads both at the encoding and decoding layer. Each encoder layer is represented as an $Encoder(.)$, which outputs the hidden state of the respective layer. The encoder block of the BART model is fed with a set of input id's from the dialogue history $Q$. Let the input for the first encoder layer be $h_e^0$. The $h_e^0$ is converted into an embedding matrix which passes through the $1^st$ layer's encoder function yielding a hidden state for the first layer. This step is repeated for each $l^{th}$ layer, where $l \in \{1, .., 12\}$. We get a hidden state $h_e$ for every $l^{th}$ layer by applying the $Encoder(.)$ function as shown in equation(1). The hidden state output $h_e^12$ from the final $12^{th}$ layer of encoder block is then used by the hidden state of the decoder layer for sequential decoding.

$$h_e^l = Encoder(h_e^{l-1}) \tag{1}$$

$$h_d^l = Decoder(h_d^{l-1} \cdot h_e^{12}) \tag{2}$$

Next we feed the set of target-response $T = \{x_1, x_2, ..., x_n\}$ to the decoder block. Similar to encoder block we represent each decoding layer as the $Decoder(.)$ function, which generates hidden state $h_d$ for each decoder layer. Again, let the decoder's input be $h_d^0$. We feed the model's decoder block with decoder's input ids along with the hidden state of $12^{th}$ encoder layer. With the help of the decoder block function it generates the hidden state $h_d$ for each layers as shown in equation(2).

We have included a linear layer in the BART's language model wrapper which generates output tokens probabilities(logits) by applying a normalized exponential function(softmax). This output helps in determining the words within a sequence. Our fine-tuned model aims at maximizing the likelihood as stated in equation(3) by training $\theta$ parameters on minimizing cross-entropy of BART model as stated in equation(4)

$$P(T|Q) = P(x_1|Q) \prod_{i=2}^{n} P(x_i|Q, x_1, ..., x_{i-1}) \tag{3}$$

$$\begin{aligned} \mathcal{L}_{xe}(\theta) &= -logP_\theta(T|Q) \\ &= -\sum_{t=1}^{N} logP_\theta(y_t|y_{t:t-1}, Q) \end{aligned} \tag{4}$$

## BERT

In the paper (Devlin et al. 2018), the author proposed BERT-Bidirectional Encoder Representations from Transformers. BERT has been successfully applied at various state-of-the-art Natural Language Processing (NLP) tasks and was able to achieve drastic improvement over previous state-of-the-art models in terms of performance. BERT architecture is consist of multiple layers of transformer encoder with each utilizing bidirectional self-attention mechanism for learning the contextual relations between words (or sub-words) in the text. In BERT's vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. For our experiment, we used $BERT_{\text{base}}$ uncased model having 12 encoder layer, 769 hidden state size, 12 attention head, 768 vocabulary size and 110 millions parameters.

[2]https://huggingface.co/models

We used $BERT_{\text{base}}$ to build a encoder and decoder model jointly trained for hate speech intervention response generation task. In our model, through BERT the input online hate speech text sequence, $D_1$ is encoded. Similarly, on the other hand we used BERT as decoder for generating the hate speech intervention response, i.e. $D_2$.

## Experiment

In this section, we elaborated the steps involved in dataset preprocessing and structuring. We also gave an overview of the models optimization strategies and hyperparameter setting used for training and validation of the dialogue generation model.

### Data Preprocessing

For our experiment, we used both Reddit and Gab hate speech intervention data(Qian et al. 2019). Some instances in dataset do not have any hate speech as a result no intervention was given, we removed these instances from the dataset and we were left with 15,016 instances. As described in the section , each pair in the dataset consisting of hate speech and it's response was used for our experiment. During the preprocessing of text data, we removed the unnecessary noisy symbols and all the emojis. All the unwanted extra spaces in the utterances were removed. The dataset was divided into 3 parts: train/validate/train, The percentage distribution of training, validation and testing dataset was 80%/10%/10% respectively. Hyperparameters fine-tuning was done using the validation dataset.

### Experiment Settings

For **BART-BART** based dialogue generation model in our experiment, we used $BART_{\text{base}}$. We utilized the Huggingface's transformer $BART_{\text{base}}$ model[3]. As stated in the paper(Lewis et al. 2019), we fine-tuned the BART model's parameters by training the model for 10 epochs with batch size 64. We utilized the Adam(Kingma and Ba 2014) optimizer along with linear warm-up scheduler and the initial learning rate was set at 4e-5. The encoder was fed with noised input tokens having a max sequence length of 500 with their corresponding attention mask tokens generated by the Byte-pair-encoding tokenizer. Similarly decoder input was tokenized with a max token length of 75 and was supplied into decoder with it's respective attention mask. During the model training the encoder-decoder sub-models were trained jointly. We train the model by calculating the cross entropy loss with label smoothing(factor = 0.1) from the logits. For **GPT-GPT** based model, we used $DialoGPT_{\text{small}}$(Zhang et al. 2019) and for **BERT-BERT** model we used $BERT_{\text{base}}$ uncased model. Similar to the BART model both models were fine-tuned on our dataset. The fine-tuning was performed for 10 epochs and the batch size was set as 64. DialoGPT uses Byte-Pair-Encoding tokenizer where as BERT uses Word-Piece tokenizer for generating token sequence from the encoder input sting. The max token sequence length for the encoder and decoder sub-model was set to 500 and 75 respectively. Adam(Kingma and Ba 2014) optimizer with along with liner learning schedule was employed and the initial learning rate was set at 4e-5. While training the models cross-entropy loss(label smoothing factor = 0.1) was calculated. For each of the model, we stored the weights based on their performance during the validation. For the comparative study and evaluation, we analyzed the models performance on the test data.

[3]Available at https://huggingface.co/models

## Results and Discussion

In this section we discuss about the performance of language models for the hate speech intervention generation task. We used two methods for the investigating the quality and accuracy of the generated dialogues : (i) evaluation using automatic metric and (ii) manual human evaluation. Below we have discussed both of the methods and results in details.

### Automatic Metric Evaluation

Our task was related to text generation we used metrics like perplexity score, BLEU-n(Papineni et al. 2002) score, METEOR score(Lavie and Agarwal 2007) and NIST-n score (Doddington 2002). These metrics are widely used in machine translation tasks where they calculate the similarity between the n-grams of ground truth with generated response from the model. In NIST scoring, the function calculates the informativeness of n-grams whereas in BLEU scoring, n-gram's precision is calculated by adding equal weights. Perplexity is a common scoring function to test models efficiency by calculating and comparing the smoothness and quality of produced responses.

In the table 1, we have summarized the performance result of all the dialogue generation model. The following are the observation we can take from the table. Firstly, out of all three models, BART achieved the lowest perplexity score of 13.20, whereas $DialoGPT_{\text{small}}$ and BERT achieved a score of 46.23 and 77.34 respectively. The biggest advantage of $BART_{\text{base}}$ was that it was trained on much bigger and diverse data in a way to reconstruct the texts from the corrupted documents, which therefore enhanced and increased BART capabilities as compared to other models. The observation that could be made was on the machine translation benchmark scores such as METEOR, BLEU, and NIST, the best performance was given by the $DialoGPT_{\text{small}}$ model. Since the $DialoGPT_{\text{small}}$ model was pretrained on a large Reddit dialogue dataset which gave the model more contextual understanding for handling our dataset and as a result, more related and relevant dialogue n-grams were produced by the model. In the table 2, we have provided the generated responses from all the dialogue generation models on an example dialogue from the test dataset. On average the generated dialogues length form, various models was approximately 99. As evident from the response generated by BART, it is more relevant, and human-like, compared to the ones generated by other baselines. BART's response correctly identifies the contextual hate-speech and asks the user to refrain from using such words. It raises the concern and demeaning nature of such comments. Being grammatically correct also gives it a human touch, which would push the user to introspect about his or her actions. In contrast, Di-

| | BERT | DialogGPT | BART |
|---|---|---|---|
| Perplexity | 77.34 | 46.23 | **13.20** |
| NIST-4 | 0.96 | **1.09** | 1.03 |
| BLEU-2 | 3.77% | 7.99% | **9.02%** |
| BLEU-4 | 0.20% | **2.14%** | 1.89% |
| METEOR | 18.55% | 19.15% | **22.97%** |

Table 1: Performance score of various models on automatic evaluation metrics.

| Model | Text |
|---|---|
| Hate Speech | Tell your best friend she's a cunt |
| Ground-truth (Human) | Using slurs doesn't make anyone seem very intelligent. That is a highly offensive term that shouldn't be used at all, let alone to a friend that is not so nice . |
| BERT | We can not think about a woman's. stop bashing terms. Sexist language in a productive or that are rude and unwelcome. Please refrain from these type of comments. Thank you. |
| BART | Use of the curse word is unacceptable in our discourse as it demeans and insults women. Women can have their own opinions regardless of their actual gender there needs to be more sensitivity regarding language associated with females. Your post uses speech that denigrates gender which is in violation of our community guidelines and terms of service agreement. Using a name calling an entire group of people based on a few parts isn't helping. Please do not use that harmful word |
| DialoGPT | No calling using that c speech just make a valid argument instead hey refrain using comments derogatory the are of productive people please lead from to type and comments other they. hurt you and . from you type admin comments be the. thanks you use admin kind be words you can. You from a such term .You you to to it this using.! |

Table 2: Grouth-truth and generated intervention response by various models for a given hate speech

aloGPT fails to identify the real issue and talks about derogatory comments in general. The response is grammatically incorrect and does not stop the user from making such similar remarks in future. BERT seems to correctly identify the issue, but is not as comprehensive in forcing the user to think about the actions taken.

### Manual Human Evaluation

(Liu et al. 2016) point out the unreliability in the automatic evaluation of the generated response. Hence, we also perform the human evaluation of these responses. Four graduate and undergraduate students were asked to rate the responses, from 1 to 5; 5 being the best. They were to rate each of the two aspects: (1) Relevancy: based on how relevant the response was to the conversation history; (2) Human-like: how far close the response sounded like a real human. The responses were de-identified to keep the response generation method anonymous to the annotators. Thus, the ground-truth replies from the doctor were also given ratings (in an anonymous way). The ratings from different annotators were finally averaged.

Table 3 shows human evaluation results. From this table, we can observe the following. First, pretrained models BART and DialoGPT perform better than the Transformers. These further illustrate the effectiveness of pretraining. Second, although DialoGPT achieves better scores on machine translation metrics, BART performs better than DialoGPT. It mirrors the issues raised in (Liu et al. 2016), pointing out the fact that machine translation metrics are not appropriate for evaluating dialogue generation. Third, BART achieves a human-like score that is very close to the ground-truth. It indicates that the auto-generated responses have high language quality. The relevancy rating of BART which is higher than 3 indicates a good level of relevance between the generated responses and conversation histories.

| | Transformer | DialoGPT | BART | Groundtruth |
|---|---|---|---|---|
| Relevancy | 2.34 | 2.93 | 3.26 | 3.92 |
| Human-like | 2.18 | 2.82 | 3.12 | 3.87 |

Table 3: Human evaluation results of various models

## Conclusion and Future Work

In this paper, we presented a analysis of various pretrained language model based encoder-decoder models for building an automatic online hate speech intervention responses system. We utilized human-generated hate speech and intervention responses to train 3 language models and measured their intervention generation capabilities via multiple automatic metrics. Our results shows that out of 3 models, $DialoGPT_{\text{small}}$ model produced very contextually correct and human-like responses against hate speech. Overall performance of all the models were very promising which paves the way for developing and building automatic online hate speech intervention systems using pretrained language models. We believe that our work would assist researcher and scientists working towards the goal of hate speech mitigation to some extent. In future, we aim at moving forward with the idea of developing multilingual online hate speech intervention responses system that would address the hate speech mitigation problem on a broader aspect.

## References


Benesch, S. 2014. Defining and diminishing hate speech. *State of the World's Minorities and Indigenous Peoples* 2014: 18–25.

Burnap, P.; and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2): 223–242.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, 138–145.

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Wang, W. Y.; and Belding, E. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint arXiv:1804.04257* .

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018b. Peer to peer hate: Hate speech instigators and their targets. *arXiv preprint arXiv:1804.04649* .

Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393* .

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

Gitari, N. D.; Zuping, Z.; Damien, H.; and Long, J. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10(4): 215–230.

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; Gnanasekaran, R. K.; Gunasekaran, R. R.; et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, 229–233.

Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* .

Jourová, V. 2016. Code of Conduct on countering illegal hate speech online: First results on implementation. *European Commission.[cit. 8. březen 2018]* .

Kennedy, G.; McCollough, A.; Dixon, E.; Bastidas, A.; Ryan, J.; Loo, C.; and Sahay, S. 2017. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, 73–77.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Lavie, A.; and Agarwal, A. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, 228–231.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* .

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* .

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of*

*the International AAAI Conference on Web and Social Media*, volume 13, 369–380.

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

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Qian, J.; Bethke, A.; Liu, Y.; Belding, E.; and Wang, W. Y. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251* .

Qian, J.; ElSherief, M.; Belding, E.; and Wang, W. Y. 2018a. Hierarchical cvae for fine-grained hate speech classification. *arXiv preprint arXiv:1809.00088* .

Qian, J.; ElSherief, M.; Belding, E. M.; and Wang, W. Y. 2018b. Leveraging intra-user and inter-user representation learning for automated hate speech detection. *arXiv preprint arXiv:1804.03124* .

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8): 9.

Schieb, C.; and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Schmidt, A.; and Wiegand, M. 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*, 1–10.

Warner, W.; and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, 19–26.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899* .

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* .