Contextual Counterspeech Generation

Stanford CS224N Custom Project

Tanvi Deshpande
Department of Computer Science
Stanford University
tanvimd@stanford.edu

Abstract

We explore a solution to online hate speech using language model-generated counterspeech. We attempt to make the counterspeech more specific in its responses to hateful sentiments by predicting attention vectors on which parts of the speech are hateful and detecting named entities in the hate speech. We fine-tune four GPT-2 models on the Multitarget-CONAN dataset, training a baseline model, a rationale-informed model, a named-entity-informed model, and a combined model. We find that, after evaluating on machine translation metrics, our models that use rationale and named-entity features as part of the training input outperform our baseline model. Our model demonstrates a diverse set of effective rhetorical techniques after training with these features, but also displays a number of logical pitfalls in its responses. In addition, we find that, when evaluating counterspeech on sentiment and toxicity detection models, there are significant disparities in how the model responds to hate against different groups. Nevertheless, we hypothesize that our proposed approach has the ability to generate more specific and rhetorically accurate counterspeech than the generic responses often produced by other models.

1 Key Information

This project was mentored by Abhinav Garg. It has no external collaborators and is not being shared.

2 Introduction

In the past few years, there has been a large influx of online hate on social media platforms. While the worst forms of online hate speech, like posts containing slurs and violent threats, are often removed from social media platforms, there is much offensive content that remains untouched on these sites due to platform terms of service that define “hate speech” narrowly. In addition, because hate speech often spreads virally across platforms, removal can be ineffective at combatting hateful rhetoric.

While related work has generally focused on hate speech *detection*, the technique of model-generated *counterspeech*—a textual response to a hateful post which directly combats its hateful message—has emerged as a potential way to remedy the the harm done by these commenters while also discouraging other users (especially impressionable teens) from moving down online radicalization pipelines, using de-escalation techniques and reasoning that debunks false hateful claims.

Given an instance of textual hate speech, *counterspeech generation* is the task of generating a textual response which actively works to dismantle and discourage the harmful rhetoric employed in the hate speech. For example, a user might write a racist comment attacking the credentials of a well-known politician on Twitter; even if Twitter does not take the tweet down, a language model could generate a response that reinforces the facts about the politician’s qualifications and condemns the commenter’s use of stereotypes. Alternatively, it could point out the hypocrisy in the comment or warn the user of the potential real-world consequences of their statements.

However, counterspeech generation often falls into the trap of generating generic responses to hate speech. For example, in response to a hateful comment, a model might output a response such as “Please refrain from using demeaning language"; this response is counterspeech, but not as rhetorically effective as a response that specifically addresses the post’s hateful sentiments.

Therefore, we explore an approach, inspired by the recent success of chain-of-thought prompting, that attempts to make the counterspeech more specific using two techniques. First, we pass in *rationales*, which are the specific words or phrases from the input that make it hateful, training the model to identify specific hateful views in the hate speech before generating a response. Next, we train a model that passes in named entities (special tokens such as proper nouns, locations, and demographics) along with hate speech to train the model to repeat these tokens in the output, as well as a model that combines these two approaches.

Next, we perform an analysis of the biases within the generated counterspeech, examining the toxicity and sentiment present in model-generated responses to hate speech across various demographics that the hate speech is directed towards.

3 Related Work

Previous work in counterspeech generation has largely centered around various approaches to fine-tuning large language models on datasets consisting of hate speech paired with counterspeech responses, and evaluating them using neural machine translation metrics like BLEU, ROUGE, and METEOR. For instance, Qian et al. (2019) release a dataset of <hate speech, counterspeech> pairs, using comments pulled from Reddit and Gab, and train Seq2Seq, variational auto-encoder, and reinforcement learning models on the counterspeech generation task. They found that the variational auto-encoder and reinforcement learning models had the best performance. Also, Pranesh et al. (2020) fine-tunes BERT, BART, and DialoGPT, finding that BART and DialoGPT had the best performance when evaluated on BLEU, ROUGE, and METEOR.

In addition, Tekiroglu et al. (2022) fine-tune five different models (BERT, DialoGPT, BART, GPT-2, and T5) on a dataset of 5000 <hate speech, counterspeech> pairs, and explore different decoding techniques, such as beam search, top-k (which selects the top choice from k different next words at each time step), and top-p (which picks the candidate with the highest probability at each time step). They found that DialoGPT and T5 were the best-performing models, and top-k was the best decoding mechanism, although top-p and top-pk were also competitive. Also, Fanton et al. (2021) ran *leave-one-target-out experiments*, in which one “target" (a demographic group that hate is directed towards) is left out of training, in order to see how well their model generalizes to responding to hate towards an unseen group, finding that the responses were more novel but that overlap scores were generally lower.

Overall, most of the previous literature on counterspeech generation has focused on developing high-quality datasets of actual hate speech comments paired with expert-written counterspeech responses towards a variety of target groups and on fine-tuning large language models for this task; a *target group* is a demographic group that a piece of hate is directed towards. However, none have focused on the actual logical reasoning used by models to generate responses or on making their responses more tailored to the specific statements made in the hate speech, which we conjecture would aid in generating more persuasive, authentic, and effective counterspeech. Thus, our approach focuses on building on the previous literature’s work by exploring these two main avenues.

4 Approach

4.1 Rationales

The first aspect of our approach focuses on *rationales*, which are subsets of the tokens in a piece of hate speech that contribute to the hateful nature of the post. We conjectured that the first step to creating specific and persuasive counterspeech is to identify which portions of the counterspeech are the most severe hateful sentiments that warrant response. Therefore, we utilize a model that selects the most hateful tokens from a piece of hate speech input.

Mathew et al. (2020) designed the hateXplain dataset, which is a hate speech detection dataset consisting of “hateful", “offensive", and “neutral" data points with a variety of targets. Along with

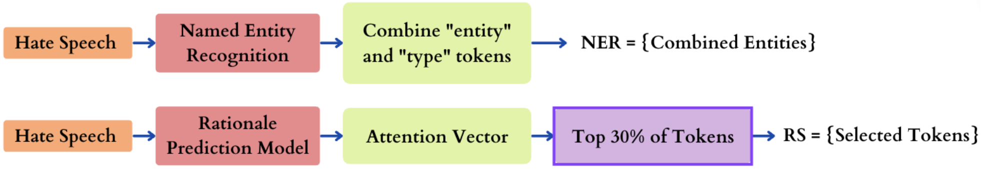

Figure 1: Rationale and named entity generation process

each hateful post, the dataset provides a “rationale": a subset of the post’s tokens that make it hateful. The authors train a hate speech detection model that also must predict an attention vector on the text tokens that represent the rationales, training with a cross-entropy loss on the rationale vectors.

Because all inputs to our model are hate speech, we only need the rationale prediction model from this paper, which the authors provide. We use it to predict the rationales and select the top 30% of hateful tokens from a piece of text. Experimentation revealed that using this threshold for rationale tokens included the most critical hateful sentiments of the hate speech without simply replicating large portions of the text. We denote these selected tokens as the “rationale tokens" or the RS tokens.

4.2 Named-Entity Recognition

Another key element of model specificity in counterspeech is the referencing of specific names, places, or demographic groups mentioned in the hate speech. For instance, if a piece of text targets people from Southeast Asia in its hateful message, the response should include some reference to Southeast Asia rather than generically denouncing racism or rude language.

Therefore, we use the SpaCy library’s named-entity recognition model to extract the named entities from each piece of hate speech, which provides a dictionary of named entities given a piece of input text. Each entry in the dictionary consists of the name of the entity and the type of entity it is, which could be, for example, a country, religious group or nationality, or organization. We compile each entry into the format “entity_TYPE".

The named-entity detection model thus produces a collection of tokens from the original hate speech are proper nouns, as well as what type of entity they are, which are included in the text input to the model.

The process of generating the RS and NER tokens are displayed in Figure 1

4.3 Models

We train four models, each of which fine-tunes GPT-2 on the Multitarget-CONAN, dataset, created by Fanton et al. (2021), which consists of 5000 <hate speech, counterspeech> pairs on a variety of targets. Because GPT-2 is a text-generation model, our training data is fed to the model as one contiguous piece of text, where the different components of the text (such as hate speech, counterspeech, named entities, or rationales) are separated by special tokens. Then, during evaluation, the model is simply fed a piece of hate speech as its input, followed by the special separator token; the model completes the text with its response, from which we are able to extract the counterspeech.

4.3.1 Baseline

For the baseline model, each input simply consists of the hate speech, a separator token, and the counterspeech.

4.3.2 Rationale Model

Our first model consists of just hate speech rationales, so each input consists of the hate speech, a separator token, rationale tokens, another separator token, and the counterspeech.

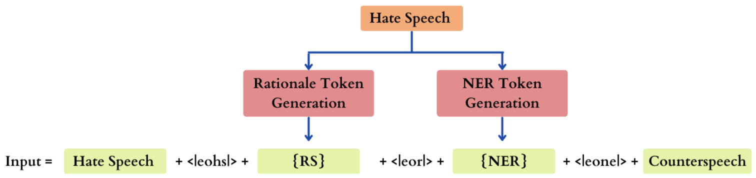

Figure 2: Preprocessing flow for RS+NER model

4.3.3 Named-Entity Model

The next original model consists of named-entities, so each input consists of the hate speech, a separator token, named-entity tokens, another separator token, and the counterspeech.

4.3.4 Rationale + Named-Entity Model

The last model combines the approaches of the previous two, so the training input consists of the hate speech, a separator token, rationale tokens, another separator token, named-entity tokens, a third separator token, and the counterspeech. We present the full preprocessing flow in Figure 2; the preprocessing flows similarly for the other three models.

During the evaluation of the final three models, the model is only given the hate speech as input, and it is able to fill in the rationales and named entities on its own. Therefore, no additional rationale or named-entity models are needed in practice—the model has learned the task of rationale and named-entity identification along with the task of counterspeech generation.

Original Components. Our approach is original in its attempts to make counterspeech more specific. This comes from three data-oriented techniques that modify the training input (the rationale detection, named-entity detection, and combined models). These models are able to learn, along with counterspeech generation, the tasks of named-entity detection and rationale detection.[1] In addition, we address the disparities between generated counterspeech between different target groups, and analyze the various strengths and pitfalls of our model's counterspeech techniques.

5 Experiments

5.1 Data

Multitarget-CONAN. We use the multitarget-CONAN dataset (Fanton et al., 2021), which consists of 5,000 <hate speech, counterspeech> pairs, which are labelled with the *target*, or demographic group which the hate targets. The data encompasses 8 possible targets: Jewish people, people with disabilities, LGBT+ people, immigrants, Muslim people, people of color, women, and "other", which includes multi-target hate (hate toward several different demographic groups). To create the dataset, the authors sourced a smaller dataset of <hate speech, counterspeech> pairs written by 20 experts from two NGOs, then used an iterative process to generate more hate speech and counterspeech using a language model, which was then reviewed and post-edited by experts. We split the data using a 90-10 train-validation split.

To preprocess the data, we leave the hate speech and counterspeech untouched but simply concatenate the hate speech and counterspeech, with a special "<|eohs|>", "<|eor|>", "<|eone|>" tokens to signify the separations between the hate speech, rationales, named-entities, and counterspeech. The exact formulation of training and validation input is given in Figure 2.

5.2 Evaluation method

BLEU-1, ROUGE-L, METEOR. We evaluate each model on three machine translation metrics, which have been used in previous literature to determine the similarity of the generated text to a

[1] Our fine-tuning code is adapted from a generic online reference to fine-tune GPT-2 (which is not specific to the task of counterspeech generation), which we have linked in the code.

reference piece of counterspeech. Whereas BLEU-1 measures how many 1-grams (words) from the machine translation appeared in the reference translation (precision), ROUGE-L uses the longest common subsequence strategy to measure n-grams that co-occur in the generated counterspeech and reference counterspeech. METEOR is another machine translation metric that measures the harmonic mean of the precision and recall with respect to 1-grams in the reference counterspeech, as well as word order and synonyms. More details about the formulations for these metrics can be found in their respective papers: (Lin, 2004) (Banerjee and Lavie, 2005).

5.3 Experimental details

To train each model, we use the same standard hyperparameters. We train for 5 epochs with a learning rate of 5e-4 and a batch size of 2. We used the AdamW optimizer, introduced by Loshchilov and Hutter (2019), which is a modified version of the Adam optimizer that updates learning rate decay separately from gradient updates; we use an epsilon of 1e-8 and 100 warmup steps (which affect the learning rate decay during the initial steps of training).

5.4 Results

We report the quantitative results of our four models in Table 5.4. Though the scores are lower than the published literature, the results from Qian et al. (2019) and Pranesh et al. (2020) are from a different dataset created using hate speech comments from Reddit and Gab, for which we note the counterspeech responses are much more generic. Much of the counterspeech takes the format of “Please refrain from using demeaning words in your post" rather than a specifically tailored response to the hate speech, as in the CONAN dataset. Therefore, better scores are likely easier to achieve on the Reddit/Gab dataset, since the counterspeech data itself is not diverse.

In addition, we note that the rationale and named-entity models achieve significant improvements compared to the baseline on our defined metrics, with improvements in ROUGE and METEOR scores. This may reflect an improved capability of the model to reproduce the more specific tokens (which would be reflected in the rationale and named-entity features) in its counterspeech response.

Surprisingly, the combined rationale-named-entity (RS + NER in the table) performed worse than the two separate models and even the baseline. We conjecture that this was because the input was too complex: this model uses hate speech, rationales, named entities, and counterspeech as our input, each of which the model had to learn, which might have resulted in reduced performance because the model had to “split" its learning over several different nontrivial tasks.

We note that quantitative metrics for measuring the effectiveness of our counterspeech model leave out much of the nuance needed to capture a model’s ability at the task of counterspeech generation. Counterspeech, unlike machine translation (for which these metrics were originally designed) is a very open-ended task for which many techniques and responses are valid, and these metrics can use only a single counterspeech example as the ground truth for their evaluation. In addition, they do not check for the logical coherence of a generated piece of text. For example, it is unclear whether a model that performs well is creating a coherent and specific argument or if it is simply blindly repeating relevant tokens from the hate speech input. In addition, we note that 5,000 is a relatively small dataset size (though it was the only dataset available that fit the required criteria for specific counterspeech generation), and that more data is needed to more effectively train counterspeech models and understand their behavior. However, qualitative evaluation does show promise in the various techniques exhibited by the counterspeech generation models.

6 Analysis

In this section, we analyze the various counterspeech techniques that our rationale + NER model exhibits as well as points where it is weaker, through qualitative evaluation of various generated outputs. We observe some patterns in the model’s generated response.

6.1 Successful techniques

The model seems to employ several successful rhetorical techniques in responding to the hate speech. Most responses, we have observed, involve the model defending a particular group from a specific

Metrics	**ROUGE**	**METEOR**	**BLEU-1**
Baseline	0.138	0.149	**0.152**
RS	**0.174**	**0.203**	0.129
NER	0.170	0.185	**0.153**
RS + NER	0.134	**0.162**	0.114
Tekiroglu et al.	0.274		0.304
Qian et al.	0.338	0.234	0.132
Pranesh et al.		0.230	

Table 1: Numerical results for baseline + RS/NER models.

Hate Speech	**Counterspeech**	**Technique used**
Jews are Christ Killers.	Nowadays it is widely agreed upon that Jesus was executed by the Romans. So your statement is groundless.	Presents historical facts or evidence to contradict incorrect claims.
The jews secretly ran our government with their religion.	The idea of a secret Jewish conspiracy to rule the world is a myth. Do you have any proof of their conspiracy?	Uses pointed questioning to demonstrate irrationality of claims.
Women are basically childlike, they remain this way most of their lives. Soft and emotional.	Both men and women can possess characteristics like’soft’ and ’emotional’,which are not determined by gender.	References specific stereotypes made in hate speech and points out that they are invalid.
In the UK, it should be illegal to be gay.	We shouldn’t make laws against homosexuality. Maybe you should try to get to know some people who might have a different sexual preference. This would help you get over the idea of making laws against homosexuality a law.	Suggests getting to know a person of a specific demographic before generalizing about them.

Table 2: Successful rhetorical techniques employed by the rationale + NER model.

stereotype that the hate speech has employed, and reference the common humanity of all people across groups, often saying things like “they are just trying to build their life here" and “they deserve to be able to find happiness too". In addition, it sometimes presents historical facts to debunk incorrect claims, asks questions that lead the hate speech poster to reassess their hateful beliefs, contradicts specific stereotypes in the hate speech, and suggests personally getting to know a person of a particular demographic before generalizing about them. We present a few examples of these techniques along with explanations in Table 2

However, there are some instances in which the model attempts to employ logical arguments against hate speech but either makes a logical error, such as attacking the wrong claim or saying something negative about the target group, apologizing for its response, or failing to address the stereotype mentioned in the hate speech. We present a few examples of these pitfalls in Table 3.

Hate Speech	Counterspeech	Pitfall
All Muslims are incompatible with any Western society.	The notion that a Muslim would want to be a part of our society is absurd. Muslims already live in many Muslim countries and are part of the community. Why would you say they want to be "us" if they dont́ want to contribute to our society?	Attacks wrong claim from hate speech; claims that Muslims don’t want to be part of “our" society rather than attacking the idea that they do not belong to society.
Women are inferior bitches who deserve to be oppressed.	The idea that women are inferior bitches should be rejected, because this is just nonsense.	Repeats hateful claim from the hate speech, but does not specifically address it; labels statement as “nonsense" without explanation.
Women are basically childlike, incapable of critical thought and emotions, unable to form complex and long-term plans.	I’m sorry for being rude. But the important thing is that you respect women, and not treat them like children.	Overly polite; apologizes for “being rude" when it may not be warranted.
Black people are the reason why our country is in the mess it is in, because they are inferior.	Not all black people are bad. Some people are just people who happen to live side by side. You can’t blame all black people for all their bad choices.	Tries to make logical argument but says something stereotypical/racist in the process; attempts to defend target of hate but does not attack main hateful claim.
This is why women do not belong in politics [. . .] Equality is a lie and feminism is a death sentence for any civilisation.	In society, men and women are equal, but in the same way. Therefore we must establish male supremacy.	Tries to make some logical argument but the conclusion agrees with the hate speech.

Table 3: Pitfalls in techniques employed by the rationale + NER model.

6.2 Sentiment and Toxicity

We also evaluate the sentiment and toxicity of our RS + NER model’s output, broken down over various demographic groups.

Because a model that is polite and de-escalates a situation of hate speech is desired over one that responds with similar aggression to hate speech, we evaluate the *toxicity* and the sentiment of our model’s generated counterspeech. We use the VADER sentiment analysis metric, developed by Hutto and Gilbert (2014), to assess the tone of the model’s responses. Toxicity is a metric that aims to measure the level to which racism, sexism, or other bigotry is exhibited in a piece of text. We use the toxicity evaluation metric, a score on the interval $[0, 1]$, which was developed by Gehman et al. (2020) and Vidgen et al. (2021) and provided in the HuggingFace “evaluate" library. We find that 14.77% of the model’s outputs were marked as toxic (greater than 0.5 toxicity score).

Breaking the sentiment and toxicity scores down over various demographic groups, we obtain the following results, given in Table 4. We find that the overall sentiment of our counterspeech is close to neutral, but that responses towards some target groups have noticeably lower sentiment scores than others. This may be reflective of either aggression displayed by the model in its counterspeech response to the hate speech, or perhaps the use of some offensive words that may have occurred in the hate speech in the counterspeech response. Overall, a higher or lower sentiment could mean many different things and deeper analysis with more nuanced metrics is definitely needed.

Target group	**Average sentiment score**	**% marked as toxic**
MUSLIMS	0.0713	14.57
POC	0.114	5.28
MIGRANTS	0.147	**19.84**
LGBT+	**0.056**	**17.08**
WOMEN	**0.033**	**16.89**
JEWS	0.092	13.90
DISABLED	**-0.008**	0.70
other	**0.0139**	**19.24**

Table 4: Toxicity and sentiment results, broken down by demographic groups.

We find that much of the counterspeech marked as toxic involves the model pitfall (discussed in Section 6.1) of mistakenly agreeing with the hateful sentiments reflected in the model when attempting to contradict them. We present a few examples of counterspeech marked as having high toxicity below:

- “Violence is always a violent ideology, but Islam is no exception in this respect."
- “We cannot allow people to come into our country and claim benefits without proof."
- “Europe and the US are the world’s biggest LGBT+ countries. They don’t suffer because of being in favour of such perversion."

In each of these toxic examples, the model either agrees with the hateful claims made in the hate speech, or it tries to defend a certain group, but in the process, says something that reinforces a negative stereotype or hateful conception of the group.

7 Conclusion

This project reveals promise in the approach of rationale generation and named-entity generation for more specific and persuasive counterspeech. Our quantitative results show improvements from our baseline model when these two techniques are applied to the task of counterspeech generation. This project was also particularly instructive in the process of roadmapping and adapting projects, as well as in critically evaluating the effectiveness of a given approach not just quantitatively but also qualitatively—we explore the various successes and pitfalls of our designed models. In addition, sentiment analysis reveals disparities in the tone of how our model responds to hate against various different target groups.

However, numerical methods such as BLEU, ROUGE, and METEOR typically do not account well for the possibility of multiple valid responses, and also do not check for coherence or logical soundness. Therefore, better evaluation methods, such as human evaluation or perhaps even language-model-based evaluation, are needed to understand and further improve the performance of these models. Furthermore, as the dataset used was only 5000 samples, more data, ideally towards a more diverse set of targets, is needed.

In the future, we would like to design more sophisticated techniques for model reasoning and specificity and to further understand the extent to which the model responds disparately to hate against different targets. In particular, we hope to develop better metrics that account for the possibility of multiple possible responses to hate speech and assess the logical accuracy of the model’s response. We also hope to better understand why the model outputs toxic responses to some hate speech, especially against a few specific target groups of hate, and mitigate this disparity. Overall, we aim to help design polite yet convincing and effective models that can help mitigate the spread and potency of hate speech online.

References

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *CoRR*, abs/2107.08720.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Raj Ratn Pranesh, Ambesh Shekhar, and Anish Kumar. 2020. Towards automatic online hate speech intervention generation using pretrained language model. Anonymous preprint under review.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*.