International Journal for Modern Trends in Science and Technology, 9(04): 357-362, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0904052
Available online at: http://www.ijmtst.com/vol9issue04.html

Generate, Prune, and Select: A Counterspeech Creation over Online Hate Speech

B. Teja Sree, Bhanu Prasanna Meka, Gandikota Durga Naga Sri, Chittineedi Vasavi Naga Sai Priya and Gottumukkala Jahnavi

Department of Information Technology, S.R.K.R. Engineering College(A) SRKR Marg, Bhimavaram

To Cite this Article
B. Teja Sree, Bhanu Prasanna Meka, Gandikota Durga Naga Sri, Chittineedi Vasavi Naga Sai Priya and Gottumukkala Jahnavi. Generate, Prune, and Select: A Counterspeech Creation over Online Hate Speech. International Journal for Modern Trends in Science and Technology 2023, 9(04), pp. 357-362. https://doi.org/10.46501/IJMTST0904052

Article Info
Received: 22 March 2023; Accepted: 16 April 2023; Published: 21 April 2023.

ABSTRACT

Keeping in view the proliferation of unpleasant language directed at minorities on social media, the development of counter-hate speeches (CHS) is regarded as an automatic approach to addressing this issue. The CHS generation is predicated on the idealistic belief thatany effort to stop hate speech on social media may have a beneficial impact on this situation. To that end, NLG has the potential to establish innovative solutions, however, off-the-shelf natural language generation methods tend to be sequence-to-sequence neural models and are constrained in that they only offer ineffectual or generic responses, regardless of the usage of hate speech, making them worthless for diffusing combative encounters. In order to significantly enhance the diversity and reliability, we devised a three-module pipeline technique that creates a variety of counter speech options including filtering the non-grammatical ones by a TF-IDF model, and then chooses the most extremely relevant counter speech conclusion by a novel retrieval- based method. In this study, using three example datasets, we propose to create a model that can provide a variety of pertinent counter- speech.

Keywords—Counter hate speech, NLG, RNN, TF-IDFModel, Cosine similarity, etc.

1. INTRODUCTION

Hate speech is any form of expression that spreads, incites, promotes, or justifies racial hatred, xenophobia, anti-Semitism, or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, prejudice, and hatred towards minorities, migrants, and individuals of immigrant background.

Fundamentally, hate speech is any expression or statement (written, verbal, digital, or otherwise) that targets someone or a group of people solely because they identify with a particular social, racial, or cultural group, frequently one that is already marginalised, excluded, or otherwise disadvantageous. Effective countermeasures demand that freedom of expression not be restricted by censorship or active moderation as a result of its enormous rise on the Internet. Here is an effective counter-speech that offers a positive response to hateful speech and promotes harmonious conversation on social media platforms by taking into account the range of appropriate responses and their applicability to the topic of hate speech. This enables coherent conversation as opposed to conversations that are off-topic or irrelevant.

Although NLG systems can generate text at scale, the q quality of the outputs is minimal in contrast to the

above mentioned parameters. In fact, the only high-quality research on counter-speech creation yet has shown its flaws: the answers are frequently irrelevant and generally commonplace. These constraints extend more widely to generic conversational language production problems, owing principally to the inherent end-to-end training nature of a single sequence-to-sequence architecture. Improved diversification or better relevance are two model improvements that have been independently addressed to adjust for these constraints. It is challenging to include these advances in a single model, though. That is what this research seeks to do. By proposing a three-module pipeline technique, Generate, Prune, and Select (abbreviated as "GPS"), to guarantee the created phrases conform to the requisite qualifications of diversification and relevancy, we approach the problem from a completely new aspect. First, using a generative model, the Candidate Generation module creates a wide array of different response candidates. As a result, a sizable candidate pool is made accessible for selection, contributing to increased diversification. Then, the Candidate Pruning module eliminates the candidates with grammar errors from the candidate pool. Therefore, he Response Selection module then selects the best suitable counter speech from the small pool of counter speech possibilities using a unique retrieval-based response selection technique for a specific instance of hate speech.

Hate Speech:	I am done with Islam and ISIS. All Muslims should be sent to their homeland. Britain will be better without their violence and ideology.
Counter-speech:	I agree that ISIS is an evil aberration, but to extend this to include up to 3 million people just in the UK is just plain silly.

Fig5: An Illustrative example from the CONAN dataset (Chungetal.2019)

We use a systematic comparison with other competing NLG systems to show the effectiveness of GPS, the first pipeline strategy for counterspeech formation, in providing varied and pertinent counterspeech through a methodical comparison with other competing NLG systems. By using both automatic and human evaluations, in three benchmark datasets, we show increased diversification and relevancy and generate brand-new, cutting-edge outcomes.

2. LITERATURE SURVEY (RELATED WORK)

A. Yi-Ling Chung et al. 2019. Conan-counter narratives using specialized sourcing: a multilingual collection ofcomments to combat online hate speech.

In Proceedings of ACL. Despite tremendous efforts to offer adequate responses to hate speech on social media platforms in terms of legislation and rules, dealing with hatred online remains a difficult subject. If hate speech is dealt through the customary procedures of content deletion or user suspension, charges of censorship and disproportionate blocking may be raised. The research community hasn't paid much attention to one alternative tactic, which is to counter hate material via counternarratives (i.e., informed textual responses). The first complete, multilingual, expert-based collection of hate speech and counter-narrative combinations was produced, and in this study, we look at how it was done. We also give extra annotations regarding expert demographics, hatred, and answer types, as well as data augmentation through translation and paraphrasing, along with the collected data. Lastly, we provide preliminary tests to evaluate the accuracy of our data.

B. Hui Su, Xiaoyu Shen et al. 2020. Using non-conversationaltext to diversify dialogue generating.

Proceedings of the Association for Computational Linguistics (ACL). When it comes to generating open-domain discourse, neural network-based sequence-to-sequence (seq2seq) models suffer greatly from the low-diversity problem. Due to the prevalence of boring and generic phrases in our daily conversation, avoiding them to produce more engaging replies necessitates sophisticated data filtering, sample strategies, or changing the training target. In this study, we provide a fresh way of looking at dialogue creation that makes use of non-conversational material. Non-conversational writing is more accessible, and diversified, and covers a wider variety of themes than bilateral conversations. We assemble a sizable non-conversational corpus from a variety of sources, such as forum posts, idioms, and passages from books. We further present a training paradigm to effectively incorporate these texts via iterative back translation. The resultant model is evaluated on two conversational datasets and is demonstrated to deliver noticeably more diversified replies without losing relevance with context.

C. Conversational Language Generation

As with conversational language production, most of the top systems for creating counter speech are based on neural models that have been trained sequentially. Although these models perform well, one of their well-known inherent flaws is the production of safe and predictable answers as a result of faulty objective functions, a lack of model variability, a weak conditional signal, and model overconfidence. This tendency served as the driving force for the development of a wide range of approaches to improve diversity, including the optimization of multiple loss functions, alteration of the latent space, use of adversarial learning, and use of non-conversational information. Our work is distinct from all others previously stated in that we employ a pipeline strategy that increases diversity by creating a diverse candidate pool. As a result, it lacks the inherent flaw of a sequence-to- sequence paradigm that was previously noted.

3. SYSTEM IMPLEMENTATION (METHODOLOGY)

NLG is regarded as the second component of NLP. It is described as the process through which a machine produces NL as an output. The machine's output should be logical, which means that any NL it generates should also be logical. Several NLG systems employ fundamental facts or knowledge-based representation to provide logical output.

A large portion of the data that you could be examining is unstructured and contains text that can be read by humans. Before we can inspect the data programmatically, we must first pre-process it.

A. Abbreviations and Acronyms

- NLG- Natural Language Generation
- GPS- Generate, Prune, and Select
- ACL- Association for Computational Linguistics
- RNN- Recurrent Neural Network
- TF-IDF- Term Frequency-Inverse DocumentFrequency
- GRU- Gated Recurrent Unit
- MMI- Maximum Mutual Information
- CoLA- Corpus of Linguistic Acceptability

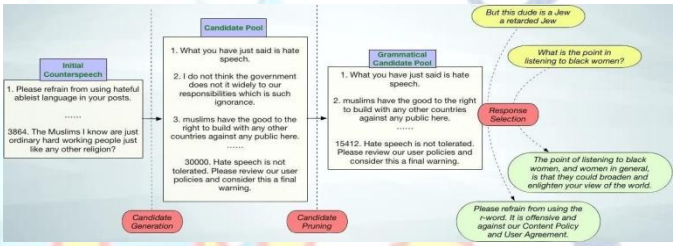
B. Algorithms UsedTF-IDF:

Term Frequency (TF) and Inverse Document Frequency (IDF) are two commonly used metrics in information retrieval and natural language processing to determine the importance of a term in a document.

Term Frequency (TF) is the frequency of a term in a document. It is calculated by dividing the number of times a term appears in a document by the total number of terms in the document. The idea behind this metric is that the more frequently a term appears in a document, the more importantit is to that document.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Inverse Document Frequency (IDF) measures the rarity of a term in a corpus of documents. It is calculated by



dividing the total number of documents in the corpus by the number of documents that contain the term. According to this metric, a term's significance to the texts it appears in increases as it becomes more uncommon in the corpus.

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

The TF-IDF score is the product of the TF and IDF scores for a term. It measures the relevance of a term to a document in the corpus. The higher the TF-IDF score, the more important the term is to the document.

There are several uses for TF-IDF, including text categorization, document clustering, and search engines. It is a useful method for removing the most pertinent data from huge document collections.

$$TF\text{-}IDF = TF * IDF$$

RNN Based Variational Encoder:

A recurrent neural network (RNN) based variational autoencoder (VAE) is a type of generative model that

combines the strength of RNNs with the probabilistic nature of VAEs.

In this model, the RNN is used to model the temporal dependencies in the data, while the VAE is used to learn a probabilistic latent space that captures the underlying structure of the data. The RNN provides a natural way to model sequences of data, such as time series or text, while the VAE provides a way to generate new data that is similar to the training data. The architecture of an RNN-based VAE generally consists of an encoder network that converts the input sequence to a latent space, a decoder network that creates a new sequence from the latent space, and a loss function that promotes the learnt latent space to be organised and informative. During training, the decoder network creates a new sequence using a sample from this distribution after the encoder network maps the input sequence to a distribution over the latent space. The loss function contains a reconstruction loss that promotes similarity between the produced and input sequences and a KL divergence loss that promotes structure and information in the learnt latent space.

All things considered, the RNN-based VAE is a potent generative model that can capture the temporal relationships in sequential data and produce new sequences that are comparable to the training data. It has applications in many different industries, including as time series analysis, speech recognition, and natural language processing.

Cosine Similarity: Cosine similarity is a measure of similarity between two non-zero vectors in a high-dimensional space. It is often used in information retrieval and text mining to compare documents based on their content. Mathematically, the cosine similarity between two vectors, x and y, can be defined as:

cosine_similarity(x, y) = dot_product(x, y) / (norm(x) * norm(y))

where dot_product(x, y) is the dot product of vectors x and y, and norm(x) and norm(y) are the Euclidean norms of vectors xand y, accordingly.

One of the advantages of using cosine similarity is that it is not affected by the length of the vectors, only their direction. This makes it particularly useful when comparing textdocuments of different lengths.

DataSets Gathering: We make use of the benchmark datasets developed by Qian et al. (2019), which include 5,257 and 14,614 hate speech occurrences from Reddit and Gab, respectively, and are fully labelled hate speech intervention datasets. We employ the Qian et al. (2019) filtered conversation setting, which only keeps talks classified as hate speech and ignores all other interactions. Moreover, we make use of the english-language section of the CONAN dataset which includes 408 examples of hate speech counterspeech authored by specialists with training in doing so. For every hate speech, there are on average 2.66, 2.86, and 9.47 ground truth counterspeech in the Reddit, Gab, and CONAN datasets, respectively.

4. EXPERIMENTS & RESULTS

We presume that we have exposure to corpus of discussion pairings with labels D = (x1, y1), (x2, y2),..., (xn, yn), where xi is a hate speech and yi is the pertinent counter speech as decided by experts or by crowdsourcing.

The objective is to develop a design that includes hateful speech as input (x) and generates a counter speech as an output (y). In Table 1, a compelling illustration is presented. The creation of interesting and pertinent counter speech is our maingoal.

A. Candidate Generation

The major objective of this module is to develop a diversified pool of candidates for generating counter speech. Using the training dataset, we extract all of the counter speech instances Y = [y1, y2,..., yn] and use a generative model to increase the counter speech pool. In order to generate candidates, we specifically use an RNN-based variational autoencoder that combines the global distributed latent representations of all utterances. To offer robust training, we use two highway network layers, where both the encoder and the decoder contain two levels of 512 nodes each. Like other generative models, its goal is to maximise the lower bound of the likelihood L of producing the trainingdata Y,

$$L = -KL(q\theta(z|y)||p(z)) + Eq\theta(z|y)[\log p\theta(y|z)]$$

where θ is parameters of the generating model, p stands for the prior distribution, q for the posterior distribution, and KL stands for the KL-divergence. Z is a latent variable with a Gaussian distribution and a diagonal covariance matrix. We use the KL annealing approach throughout the training procedure to avoid the unwanted stable equilibrium problem. When training is complete, we create candidates by just decoding noise samples taken from a typical Gaussian distribution. Here, the generative model not only generates a variety of options but also fully captures the holistic qualities of sentences, including style, subject, and high-level syntactic elements.

B. Candidate Pruning

Candidates generated by an RNN-based variational autoencoder, while diverse, are not always grammatical. As a result, in this module, we exclude everything except the grammatical candidates from the list. The corpus of linguistic acceptability (CoLA), a dataset comprising 10,657 English phrases identified as grammatical or ungrammatical from linguistics journals, is used to train a grammaticality classifier in order to do this. The classification model we choose is TF-IDF, which we then hone on the CoLA dataset. The TF-IDF model is chosen because it has a higher computational efficiency and can capture both syntactic and contextual information the best.

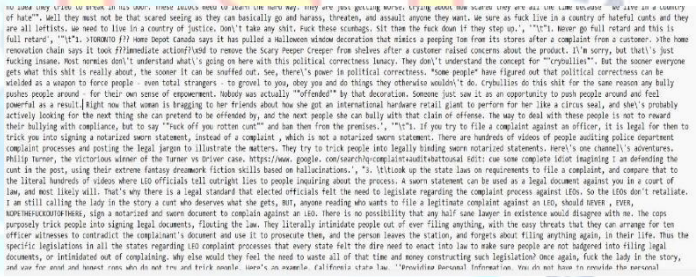
C. Response Selection

We now have a variety of grammatically correct counter speech replies. The goal is to choose the most pertinent answer to a specific incident of hate speech. We develop a pretrained response selection model for task-oriented conversation systems, fine-tune it using our dataset, and consider the genuinely small number of training cases that are available.

1. With the negative sampling strategy, train a response selection classifier: It depends on selecting undesirable applicants at random from the candidate pool. One hate speech incident typically includes several relevant counter speech incidents in our task, though. Other claims that might be used as successful counterarguments include the following: "You cannot hold a tiny number of people accountable for all of their actions." Outright prohibiting anything won't solve anything, or "Can anything ever be banned?" As a result, a number of wrongly chosen negative examples may have a significant detrimental influence on the response selection classifier's inductive bias.

2. Select based on cosine similarity: We point out that the replies (counter speech candidates) and the input (hate speech) do not share the same latent vector space, thus the cosine similarities of the learnt embeddings may not fully achieve the goal of tying the response to the input.

In order to enhance the second technique, we fuse the latent spaces of the input and the replies rather than directly adopting the two ways that are now accessible. We explicitly recommend learning a linear embedding mapping from the latent space of the responses to the latent space of the input and then using cosine similarity to select the best option. We empirically discover that the linear mapping is efficient, so we hold off on using more advanced mapping techniques until later.



5. CONCLUSION

For the development of counter speech against online hate speech, we suggested a three-module pipeline called Generate, Prune, and Select. An empirical analysis of three datasets shows that our approach is capable of generating a variety of pertinent counter speech.

FUTURE WORK

The future enhancement could go either of the following two directions: 1) Generating stylistically effective counter speech: For various hate speech concerns, different counter speech styles/strategies can be required; as a consequence, it would be intriguing to build new techniques for developing up with the appropriate counter speech style for each hateful subject. Given that we can use a style categorization in the Candidate Pruning function, we believe this might be an appropriate addition to our suggested model. 2) System

deployment: This area of research may be directly impacted by looking at the social effects of using computerized counter-speech production to lessen online hate speech via system deployment and actual performance monitoring.

ACKNOWLEDGMENT

We gratefully acknowledge the anonymous reviewers whose comments on earlier revisions have enhanced this research. Also, without the exceptional support of my superiors, B. Teja Sree madam and Dr. K. Kishore Raju sir, I could not have managed to finish this research. I am also grateful to the staff of Sagi Rama Krishnam Raju College for their valuable assistance and guidance during the project. It would have been challenging to finish the task on time without their prompt ideas and insights, as well as the chanceto collect data from other sources made possible by the organization's staff. Their efforts were essential to the project's successful completion.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019a. Spread of hate speech in online social media. In Proceedings of the 10th ACM Conference on Web Science.

[2] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. arXiv preprint arXiv:1812.02712.

[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gul- cehre, 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 Empirical Methods in Natural Language Processing (EMNLP).

[4] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan-counter narratives through niche sourcing: a multilingual dataset of responses to fight online hate speech. In Proceedings of ACL.

[5] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder decoder model for generating dialogues. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).

[6] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In Proceedings of NAACL HLT.

[7] HuiSu, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In Proceedings of the Association for Computational Linguistics (ACL). 20 Matthew Williams. 2019. Hatred behind the screens: Are port on the rise of online hate speech