DisCGen: A Framework for Discourse-Informed Counterspeech Generation

Sabit Hassan
School of Computing and Information
University of Pittsburgh, Pittsburgh, PA
sabit.hassan@pitt.edu

Malihe Alikhani
Khoury College of Computer Science
Northeastern University, Boston, MA
m.alikhani@northeastern.edu

Abstract

Counterspeech can be an effective method for battling hateful content on social media. Automated counterspeech generation can aid in this process. Generated counterspeech, however, can be viable only when grounded in the context of topic, audience and sensitivity as these factors influence both the efficacy and appropriateness. In this work, we propose a novel framework based on theories of discourse to study the inferential links that connect counter speeches to the hateful comment. Within this framework, we propose: i) a taxonomy of counterspeech derived from discourse frameworks, and ii) discourse-informed prompting strategies for generating contextually-grounded counterspeech. To construct and validate this framework, we present a process for collecting an in-the-wild dataset of counterspeech from Reddit. Using this process, we manually annotate a dataset of **3.9k** Reddit comment pairs for the presence of hatespeech and counterspeech[1]. The positive pairs are annotated for 10 classes in our proposed taxonomy. We annotate these pairs with paraphrased counterparts to remove offensiveness and first-person references. We show that by using our dataset and framework, large language models can generate contextually-grounded counterspeech informed by theories of discourse. According to our human evaluation, our approaches can act as a safeguard against critical failures of discourse-agnostic models.

1 Introduction

A promising countermeasure to hatespeech is *counterspeech* (Mathew et al., 2018b) —any response that counters hateful and offensive content, at times referred to as *Counter-Narrative* (Fanton et al., 2021). Counterspeech do not appear in isolation, but as an integral component of a broader discourse. In this work, we leverage theories of discourse (Asher and Lascarides, 2005) to capture the context that the counterspeech appears in.

The interpretation of a counterspeech is shaped by its relevance to the topic, its intended audience, and its sensitivity to the matter. For instance, a counterspeech posing a *Probing Question* might resonate differently with an audience compared to one providing a *Correction*. We propose a discourse-aware framework, **DisCGen**, to study these different types of counter speeches and how we can potentially generate them automatically. **DisCGen** consists of a taxonomy of counterspeech based on discourse relations and discourse-augmented prompting strategies. The taxonomy is derived from **Segmented Discourse Representation Theory (SDRT)** (Asher and Lascarides, 2005).

Since there is no existing counterspeech dataset with discourse relations, we construct the first dataset from Reddit to construct and validate our framework. We choose Reddit as the data source as in-the-wild data from Reddit is likely to have more diversity compared to Nichesourced or Crowdsourced datasets that contain counterspeech written by NGO workers (Chung et al., 2019) or Mechanical Turk annotators (Qian et al., 2019). Diversity in the dataset is important to demonstrate the flexibility of our framework.

Constructing an in-the-wild dataset, however, is challenging as the percentage of comments forming hatespeech-counterspeech pairs is very small on Reddit. As such, we follow a two-stage process for collecting *effective* counterspeech in-the-wild. We manually annotate a dataset **3.9K** Reddit comment pairs for the presence of hatespeech-counterspeech pairs. We manually annotate **250** positive pairs of hatespeech-counterspeech with SDRT relations. We paraphrase the counterspeech manually to remove profanity and first-person references while retaining the original content and linguistic style. We also annotate the positive samples with the **tar-**

[1] Our code and data can be requested from here: https://github.com/sabithsn/DisCGen

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 420–429
November 1–4, 2023.

geted group in the hateful comment. While the full dataset can be used for identifying effective counterspeech, the positive pairs can be used with our framework for counterspeech generation.

Lastly, we combine the proposed discourse-based taxonomy with prompting strategies in our framework. We compare Large Language Models (LLMs) under different settings. In the first setting, discourse relations are provided for both the examples in the prompt and for the inference text. In the second scenario, discourse relations are provided only for the prompt examples. These models are compared with a baseline discourse agnostic approach. Our analysis shows greater diversity in discourse relations preserved in the discourse-informed approach compared to discourse-agnostic prompting. We show that, in both discourse-informed settings, LLMs are capable of generating highly accurate counterspeeches (**>95%** cases) and also respect corresponding discourse relations (**74%** and **90%** for the two strategies respectively). Further, our human evaluation shows that our strategies can act as safeguard against critical failures that discourse-agnostic LLMs are susceptible to. Thus, the contributions of this paper are:

- A novel framework for **discourse-aware** counterspeech generation that comprises of: i) a discourse-based taxonomy of counterspeech and ii) discourse-informed prompting strategies.
- A process for collecting an **in-the-wild** dataset of effective counterspeech from Reddit.
- First dataset of **3.9K** pairs of Reddit comments annotated for hatespeech-counterspeech, with **250** positive pairs annotated for: i) taxonomy derived from **SDRT**, ii) **paraphrasing** removing offensiveness and first-person references, and iii) **target group** that the hateful attacks.

2 Related Work

While offensive content on social media has gained much recent interest (Ye et al., 2023), work on counterspeech is still under-explored. Benesch et al. (2016) conduct a field study of counterspeech on Twitter and list eight associated strategies from a social angle. The nichesourced CONAN (Chung et al., 2019) dataset and its subsequent variations (Fanton et al., 2021; Bonaldi et al., 2022), contain counterspeech written by NGO workers. Qian et al. (2019) inject counterspeech written by Mechanical Turk workers into conversations from Reddit and Gab. Our dataset is the first to contain counterspeech collected directly from Reddit, written by Reddit users. Mathew et al. (2018a) and Mathew et al. (2018b) collect counterspeech from Twitter and YouTube respectively. While these are written by social media users, they are not suitable for *generation* models due to presence of short and offensive counters. Our data is specifically curated for generation with removal of profanity and first-person references. Different from the existing works, ours is the first to present a discourse-based taxonomy and a dataset annotated with discourse relations.

A few recent works have studied counterspeech generation. Bonaldi et al. (2023) propose an attention based regularization with GPT-2 to generate more specific counter narratives. Zhu and Bhat (2021) first generate multiple candidates, filter out ungrammatical ones, and then selects the most relevant countersppech. Chung et al. (2021) study generation of knowledge-grounded counternarratives using an external database. Ashida and Komachi (2022) show the effectiveness of prompting large language models (LLMs) for generating counterspeech. Vallecillo-Rodríguez et al. (2023) show that GPT-3 is more capable of generating counter narratives compared to other large language models. Ours is the first work to provide a framework for discourse-aware counterspeech generation.

Discourse relations have been proposed as a mechanism for controlling generation, shown to aid summarization (Cohan et al., 2018; Xu et al., 2020), style transfer (Atwell et al., 2022), and question answering (Huang et al., 2021; Xu et al., 2022). Bosselut et al. (2018) show that discourse-aware models can generate more coherent texts. None of the aforementioned works however, target counterspeech generation. To our knowledge, our work is also the first to integrate discourse-based framework within prompting strategies for LLMs.

3 Framework

Our framework consists of a discourse-based taxonomy and two prompting strategies.

3.1 Discourse-based Taxonomy

The style and efficacy of counterspeech is often dictated by its context, which is typically offensive or hateful content. Thus, we aim to identify different types of counterspeech through the lens of discourse relations. For the remainder of the paper, we use hatespeech as an umbrella term for offensive/hateful content.

With the help of linguist annotators, we explored ways different discourse theories such as Penn Discourse Treebank (PDTB) (Prasad et al., 2008), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2005) might help us study the inferential links that connect hateful comments with counterspeech. Our initial investigation informed us that SDRT labels could closely model these inferential links. Thus, we decide to choose SDRT as primary source for our taxonomy.

We annotate 250 counterspeeches (Section 4) to construct a discourse-based taxonomy of counterspeech that contain 10 classes adapted from SDRT relations defined in Asher et al. (2016):

- *Acknowledgment* when the counterspeech signals an understanding. While Asher et al. (2016) include both understanding and acceptance as acknowledgment, acceptance is not considered in our definition as counterspeech should not agree with the hatespeech.

- *Clarification Question* when the counterspeech asks questions to clarify information presented in the hatespeech, analogous to Asher et al. (2016).

- *Comment* when the counterspeech provides an opinion or evaluation of the content in hatespeech, analogous to Asher et al. (2016).

- *Correction* when the counterspeech corrects an argument/fact presented in the hatespeech, analogous to Asher et al. (2016).

- *Elaboration* when the counterspeech expands on the scenario presented in the hatespeech. Differing from Asher et al. (2016), counterspeech does not elaborate on its own argument, but offers a broader perspective on the hatespeech.

- *Probing question*, when the counterspeech asks a question intending to acquire more information, similar to Q-Elab in Asher et al. (2016).

- *Explanation* when the counterspeech offers an explanation of a situation presented in the hatespeech, similar to Asher et al. (2016).

- *Parallel* when the counterspeech shows commonality between hatespeech and an external scenario, a special case of Asher et al. (2016).

- *Result* when the counterspeech connects the consequences to the content of hatespeech. The consequence is a special case of "effect" in Asher et al. (2016).

3.2 Prompting Scenarios

We propose prompting strategies for two scenarios that use discourse relations in our taxonomy.

Strategy 1: The LLM is provided with discourse relations only for prompt examples. The LLM is asked to decide an appropriate discourse relation for inference text, and then generate a counterspeech .
Strategy 2: The LLM is provided with discourse relations for prompt examples and also inference text. The LLM is then asked to generate counterspeech respecting discourse relation of the inference text.

Strategy 1 is to be applied when there is no prior information about the type of counterspeech that should be generated. The model learns from prompting examples the type of discourse relations it should maintain with respect to context. Strategy 2 is to be applied when the desired discourse relation is known beforehand.

4 Dataset

In this section, we first outline our data collection pipeline. Then, we describe our two-stage process for constructing our dataset, followed by annotation protocol and inter-annotator agreement. Lastly, we analyze distributions in our dataset[2].

4.1 Data Collection

We use Pushshift API[3] to collect comments from 14 subreddits (Appendix **??**) spanning topics of politics, personal views, gender rights, and question-answer across a period of six months, starting from June 1st, 2021. To filter out comments that do not contain hatespeech, we fine-tune a BERT (Devlin et al., 2019) on the fine-grained hatespeech data in Multitarget CONAN (Fanton et al.,

[2]Our data collection was approved by our institution's ethics review board (anonymized for blind review)

[3]https://github.com/pushshift/api

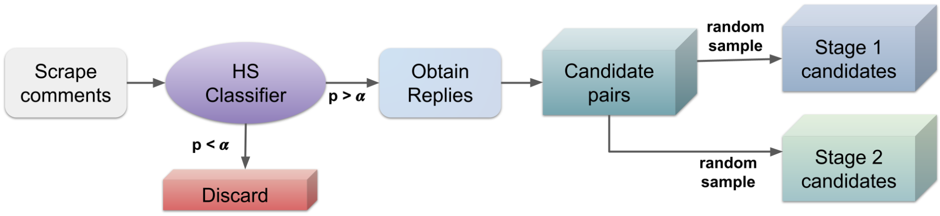

Figure 1: Our data collection pipeline. Comments and their replies are scraped from Reddit, and then run through a hatespeech classifier. If the classifier confidence falls below threshold α then they are discarded, else their replies are obtained to form pairs. These pairs are randomly split into two buckets for two stages of annotation.

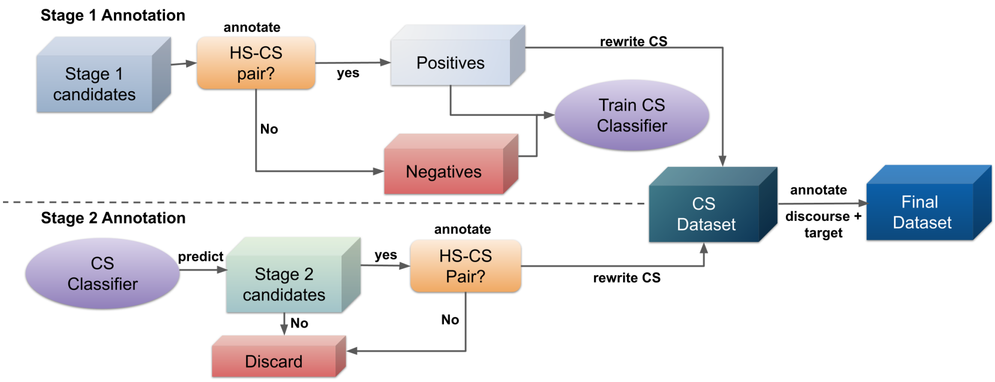

Figure 2: Our process for constructing in-the-wild dataset from Reddit. First set of random samples are annotated for hatespeech-counterspeech. Then we train a classifier to predict counterspeech, and then perform a final round of annotation. We rewrite the counterspeech if it is offensive or first-person. Finally, we annotate for discourse relations and target groups. In the figure, HS refers to hatespeech and CS refers to counterspeech.

2021). The hatespeech segment of the dataset contains 5K fine-grained annotations for the following classes: WOMEN, POC, LGBT+, DISABLED, JEWS, MUSLIMS, MIGRANTS and OTHER. We removed duplicates in the dataset and use 70-10-20 split for train, dev and test data. The classifier achieves an F1 score of **91.02** on the test set. We use the classifier to classify comments. We obtain replies using PRAW[4] for only those tagged with hatespeech with probability > 0.8. The threshold is empirically decided. Discarding comments without replies, we end up with 18K comments. We do not use the counterspeech data in the Multitarget CONAN dataset because using a classifier trained on Nichesourced data would not capture the natural diversity of in-the-wild data.

Lastly, We take a random sample of 500 from each target group (excluding the OTHER class, for total of 3.5K) and set aside for stage 1 annotation. We take another random sample of 1K from each target group (total of 7K) and set aside for stage 2. Figure 2 shows our annotation process for the two stages.

[4]https://praw.readthedocs.io/en/stable/

Model	Acc	Prec	Recall	F1
bert-base-cased	90.0	58.1	65.0	60.1
bert-base-uncased	**93.1**	**64.5**	**69.6**	**66.6**
roberta-base	**95.7**	47.9	50.0	48.9
xlnet-base	86.6	53.4	62.3	53.6
albert-base-v2	91.4	46.2	49.4	47.8

Table 1: Results of finetuning pretrained models to detect counterspeech. bert-base-uncased outperforms all other pretrained transformers.

4.2 First Stage Annotation

We manually annotate 3.5K comment pairs for presence of hatespeech and counterspeech. Due to our rigorous annotation protocol (described in section 4.4), we end up with 152 positive pairs. Using this data, we fine-tune a range of pretrained transformer models to detect counterspeech: bert-base-cased, bert-base-uncased (Devlin et al., 2019), roberta-base (Liu et al., 2019), xlnet-base (Yang et al., 2019) and albert-base-v2 (Lan et al., 2019). All models are fine-tuned with the same parameters: learning rate of 8e-5 and batch size of 16 for 5 epochs. The results are reported in Table 1. All positive pairs are annotated for discourse relation and target groups. The counterspeech is also rewritten, if necessary, to remove offensiveness and first-person references.

4.3 Second Stage Annotation

Using the best classifier from Stage 1 (bert-base-uncased), we tag the pool of data set aside for Stage 2. 360 samples are tagged as counterspeech. The pairs containing these 360 samples are manually annotated with the same protocol, yielding 98 more positive pairs. This shows that using this method, we can grow the size of dataset containing counterspeech with much fewer human annotations. This approach can be used in the future to construct larger datasets. Since our purpose in this paper is to construct a dataset for prompting, we consider 250 positive samples to be enough as large language models are prompted in a few-shot setting.

4.4 Annotation Protocol

We recruit two graduate annotators with linguistic background. The annotators were paid according to the approved rate by the human-subject review board.

To construct our dataset, we define protocol for four types of annotations: i) hatespeech-counterspeech pair, ii) paraphrasing counterspeech, iii) target group, and iv) discourse relations.

Hatespeech-counterspeech pair: To decide if a comment is hatespeech, we ask the annotator to identify if the comment is offensive and targets any of the seven groups in the data. To determine if a reply to the hatespeech is counterspeech, the annotators are asked to assess if the response counters the hatespeech. We ask the annotators to discard any counterspeech that simply uses profanity and are not constructive.

Target group: The annotators are asked to choose the target group (e.g., migrants/LGBTQ) that the hatespeech attacks. The target groups are defined in (Fanton et al., 2021).

Discourse annotation: We provide the annotators with the hatespeech and counterspeech and ask to determine which SDRT discourse relations is most applicable between the two. We provide them with the SDRT annotation manual by Asher et al. (2016), modified with examples from our dataset. Annotators were also able to choose "unknown/ no-discourse relation present". These instances are excluded from our dataset.

Paraphrasing counterspeech: Even after discarding counterspeech that just contain profanity, we observe that some constructive counterspeech contain profanity to a degree. Thus, we ask the annotators to remove such profanity. Since our goal is to build dataset for counterspeech *generation*, we also ask annotator to remove any first-person reference. Both types of edits are made with minimal modification while retaining the original meaning and linguistic style.

4.5 Inter-Annotator Agreement (IAA):

Since the percentage of counterspeech in our data is very small, taking a random overlap of the full dataset would not yield any useful information. Thus, we take a random sample of *positive* samples as overlap between two annotators.

Hatespeech-Counterspeech: In 90% of the cases, the annotators agreed that the given pair contained hatespeech and counterspeech. The disagreements were primarily due to one of the annotators misinterpreting context of the hatespeech.

Target group: The Cohen's Kappa (Cohen, 1960) for target group annotation was 0.83, showing a high degree of agreement. The only cases where the annotators disagreed were due to presence of multiple target groups in the hatespeech. For example, a hatespeech targetting black women is labeled as "POC" by one annotator and "WOMEN" by the other.

Discourse relations: The Cohen's Kappa for discourse annotation was 0.62, indicating substantial agreement (McHugh, 2012). The lower agreement for discourse annotation is expected due to the difficulty of the task (Asher et al., 2016). The primary source of disagreement was confusion between classes that appeared together. For example, a pair often exhibited characteristics of both Comment and Correction classes.

4.6 Dataset Analysis

Target group distribution: Although we started with the same number of candidates for each group, we observe that after annotation, the data has the highest hatespeech-counterspeech pairs for hatespeech targeting WOMEN. Our manual examination reveals that among the candidate pairs, the classifier often mistook discussions about LGBT+ or POC as hatespeech even though they are not offensive. However, as seen from Figure 3 that the distri-

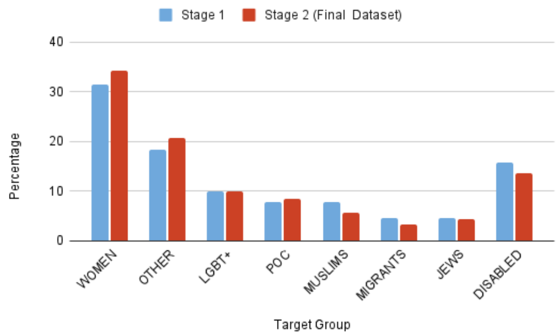

Figure 3: Women as the target group have the highest retention rate. The distribution remains similar across two stages.

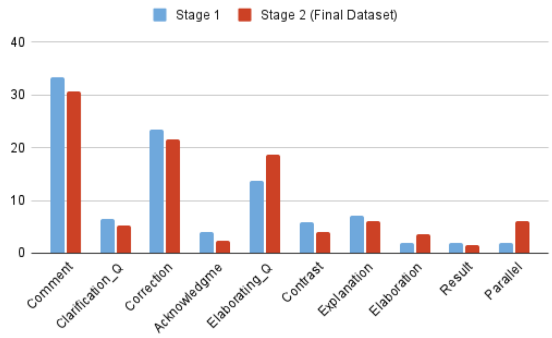

Figure 4: Comment and Correction are the dominating discourse relations. The distribution remains similar across the two stages.

bution remains similar across the two stages. Thus this skewness is primarily a limitation of the classifier trained on the MultiTarget CONAN dataset rather than the classifier used in Stage 2.

Discourse relation distribution: We observe that the discourse relations Correction and Comment are the most dominant ones in Figure 4. This is expected from a natural distribution because users are more likely to correct hateful content or denounce them than ask questions or show acknowledgment. Similar to earlier, we see that the distribution remains similar across the two stages.

5 Counterspeech Generation

In this section, we provide analysis and evaluation of proposed strategies. For all experiments, we use davinci-text-003 version of GPT-3. We use 50 randomly chosen samples as example in the prompts and evaluate the models on the remaining 200 samples of our dataset. In our experiments, baseline GPT-3 is instructed to generate counterspeech for given hatespeech and is compared with the two prompting strategies outlined in Section 3.2.

5.1 Evaluation:

Evaluation of generated counterspeech is considered a difficult task (Ashida and Komachi, 2022). Common generation metrics such as BLEU, BERtScore are not helpful for evaluating counterspeech. As such, we primarily rely on human evaluation similar to past works (Ashida and Komachi, 2022).

For human evaluation, we consider: i) does the generated text count as counterspeech, ii) is the generated text offensive[5], and iii) for prompting strategy 1 and 2, does the generated text respect discourse relation indicated. Table 2 show examples of counterspeech generated by different strategies.

Accuracy: We observe that in a few cases (6%), the baseline generated text that could not be considered counterspeech as they agreed with the input text instead of countering them (Table 2). Our proposed strategies, however, had fewer such failures (4% and 2% respectively for strategy 1 and strategy 2). This suggests, by explicitly instructing language models with discourse relations, we can avoid pitfalls of generating counterspeech by prompting large language models.

Offensiveness: Since we removed profanity from our dataset with paraphrases, the generated counterspeech did not display offensiveness toward any groups during human evaluation. We also used an independent classifier, a bert-base-cased model, finetuned on the OLID dataset (Zampieri et al., 2019b) to tag the generated counterspeech. The classifier tagged 20% of the generated counterspeech as offensive. However, with manual analysis, we observe in most cases, the classifier tagged sensitive topics and text about minority groups as offensive. This is consistent with the observations by Hartvigsen et al. (2022) that toxic language detection systems can falsely tag text with minority group mentions. Such limitations of classifiers need to be addressed for an unbiased large-scale evaluation of machine-generated counterspeech.

Diversity We observe that without any instructions about discourse relations, GPT-3 mostly generates counterspeech that are Comment or Correc-

[5]we refer to definition provided in (Zampieri et al., 2019a):"any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words"

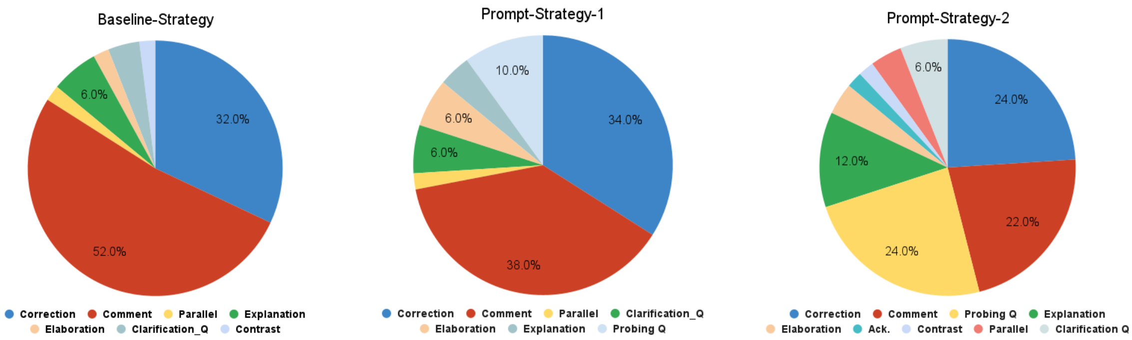

Figure 5: Distribution of discourse relations in generated counterspeech. Strategy 1 yields higher diversity than baseline while Strategy 2 generates the most diverse counterspeeches.

Hatespeech *No. School is for teaching my child how to read, write, do math, lessons about history, geography, etc. Its not for my 5 year old to learn about sexuality. That's my job to decide if and when that's appropriate.*
Baseline: Agreed. Schools should focus on the basics and leave the more complex topics to parents.
Strategy 1: Agreed. Schools should focus on the basics and leave more complex topics to parents. **[Acknowledgment]**
Strategy 2: Sex education is important for children to learn about their bodies and how to protect themselves from abuse. It's not just about sexuality, it's about safety and respect. **[Explanation]**

Table 2: Our proposed strategies can safeguard against critical failures, generate less generic response and have higher linguistic diversity. While Strategy 1 is more accurate than baseline, Strategy 2 is the most accurate.

tion. While these two are also the most frequent categories for our strategies, our strategies yield higher frequency of discourse relations such as Q-Elab or Elaboration. Higher diversity is observed when relations are explicitly mentioned in prompting strategy 2. Figure 5 shows the distribution of discourse relations in generated counterspeech.

Respecting discourse relations: We evaluated if the generated texts by the LLMs respect discourse relations. For the first strategy, 74% counterspeech were generated in the discourse relation that GPT-3 outputs. For the second strategy, 90% of the counterspeech were generated in the explicitly specified discourse style. This suggests specifying desired discourse relations during inference time can be respected well by language models compared to providing discourse relations only in the prompt examples.

6 Discussion

In this section, we discuss the challenges of counterspeech generation.

Evaluation: Evaluation of counterspeech remains a challenging task. As our examples show, language models such as GPT-3 mostly produce grammatically correct and coherent texts. As such, automated metrics of grammar and coherence (Marchenko et al., 2020) are not good indicators of the quality of generated counterspeech. Instead, there is a need for automated metrics that can measure the countering capacity of generated text. While the focus of this paper is constructing the framework, dataset and generation capabilities of language models, it is important to conduct a study to evaluate the efficacy of different content in counterspeech among real users in the future.

Classifier bias: We observed that the initial classifier, trained on the MultiTarget CONAN dataset (Fanton et al., 2021), that we used to identify hatespeech candidates, has certain limitations. Although the classifier boasted a 91% F1-score on the test set, it often tagged instances that are not hatespeech but talked about sensitive topics as hatespeech. Although we eliminated this bias by manually excluding them, care must be taken for building datasets using such classifiers. While there has been studies regarding gender and racial exhibited by classifiers recently (Ahn and Oh, 2021; Lu et al., 2020), further study is needed to quantify and mitigate this kind of bias. Bias reduction methods for classification tasks (Hassan and Alikhani, 2023) need to be explored in the context of generation.

Knowledge Grounding: The focus of this paper has been to construct the first-of-its-kind in-the-wild dataset annotated with discourse relations that can be used to control the content when prompting large language models. An aspect of counterspeech generation that is out-of-scope for this paper, but needs to be explored is how facts and knowledge can be injected into the generated counterspeech. Approaches that rely on an external database of knowledge (Chung et al., 2021) can be used in conjunction with our approach to generate knowledge-grounded counterspeech that respects discourse relations. Approaches such as use of topic phrases (Fan et al., 2019) can be explored for complementing our discourse-augmented prompting mechanism as well.

7 Conclusion

In this paper, we presented **DisCGen**, a framework for discourse-informed counterspeech generation. Within this framework, we presented a discourse-based taxonomy of counterspeech and two discourse-informed prompting strategies. Further, we outlined a process for the challenging task of collecting in-the-wild counterspeech from Reddit. Using our methodology, we collected a first-of-its kind dataset that contains ***hatespeech-counterspeech pairs*** annotated for ***discourse relations***, ***paraphrased*** version of counterspeech and the ***targeted group*** in the hatespeech . Our automated and human evaluation show that LLMs can generate accurate and inoffensive counterspeech with our dataset. Our analysis shows that by using our proposed discourse-aware framework, we can control the content of counterspeech generated by large language models with respect to the context. We also show that our proposed approach results in a higher diversity in terms of linguistic style and can serve as a safeguard against critical failures of discourse-agnostic approaches.

Limitations

Using classifiers to aid counterspeech classifier may result in bias. However, it is a necessary step for collecting a sizeable dataset as the percentage of counterspeech on social media is extremely low. It should be noted, however, that we manually verify all positive instances tagged by the classifier, eliminating any false-positive bias the classifier may exhibit.

It should also be noted that the scope of this paper is to present a framework for discourse-aware counterspeech generation, outline a methodology for collecting an in-the-wild dataset, publicly share the dataset, and evaluate LLM's capacity of generating counterspeech with the proposed framework. A large-scale user-study for evaluating the social impact of different types of generated counterspeech is not within the scope of this paper.

Ethics Statement

In certain scenarios, counterspeech can be insensitive to users. Inappropriate counterspeech can hurt the feelings of social media users rather than promote a safer environment. As such, counterspeech generation is aimed to reduce psychological pressure on human moderators and social media users, not replace them. Counterspeech generation should not be used indiscriminately across social media. In an ideal case, the generated counterspeech should be reviewed by a human before posted on social media or elsewhere.

Although text generated by large language models such as GPT-3 are coherent and relevant, they may exhibit bias toward certain groups such as feminine characters (Lucy and Bamman, 2021). If a generated counterspeech exhibits bias towards certain groups, it may have adverse effects. Although we did not observe such behavior with our models, these models needs to be carefully examined for specific use cases before deploying in the real-world.

Acknowledgment

This project was supported by DARPA grant prime OTA No. HR00112290024 (subcontract No. AWD00005100 and SRA00002145). We also acknowledge the Center for Research Computing at the University of Pittsburgh for providing computational resources. We would also like to thank the human annotators, the anonymous reviewers, and Katherine Atwell for their valuable feedback.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2005. Logics of conversation. In *Studies in natural language processing*.

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. *Counterspeech on Twitter: A field study*. Dangerous Speech Project.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. *ArXiv*, abs/2211.03433.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, W. Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Sabit Hassan and Malihe Alikhani. 2023. D-CALM: A dynamic clustering-based active learning approach for mitigating bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5540–5553, Toronto, Canada. Association for Computational Linguistics.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Oleksandr Marchenko, Olga Radyvonenko, Tetiana Ignatova, P. V. Titarchuk, and Dmytro Zhelezniakov. 2020. Improving text generation through introducing coherence metrics. *Cybernetics and Systems Analysis*, 56:13–21.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018a. Analyzing the hate and counter speech accounts on twitter. *ArXiv*, abs/1812.02712.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2018b. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Maria Estrella Vallecillo-Rodríguez, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71:227–245.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. *arXiv preprint arXiv:2203.11048*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *ACL*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Predicting the type and target of offensive posts in social media. In *NAACL*.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.