A Strategy Labelled Dataset of Counterspeech

Aashima Poudhar[1] and **Ioannis Konstas**[1,2] and **Gavin Abercrombie**[1]
[1]Heriot-Watt University [2]Alana AI
Edinburgh, Scotland
{ap2099, i.konstas, g.abercrombie}@hw.ac.uk

Abstract

Increasing hateful conduct online demands effective *counterspeech strategies* to mitigate its impact. We introduce a novel dataset annotated with such strategies, aimed at facilitating the generation of targeted responses to hateful language. We labelled 1000 hate speech/counterspeech pairs from an existing dataset with strategies established in the social sciences. We find that a *one-shot* prompted classification model achieves promising accuracy in classifying the strategies according to the manual labels, demonstrating the potential of generative Large Language Models (LLMs) to distinguish between counterspeech strategies.

1 Introduction

Over 60% of the world's population use social media platforms (Dean, 2024) and many interactions on these involve hateful and toxic language (Vidgen et al., 2019). While recent research has begun to investigate the use of counterspeech as an effective technique to mitigate hate while preserving the right to free speech (compared to traditional flagging and moderation), there is little natural language processing (NLP) research investigating counterspeech generation based on known, effective strategies.

There are, in fact, a wide range of strategies employed in counterspeech, from fact-checking to use of humour, and research on counterspeech deployed in real-life situations shows its effectiveness to vary significantly depending on the approach taken (Benesch et al., 2016; Chung et al., 2023).

Our contributions Focusing on English language interactions, we develop a nuanced understanding of counterspeech by annotating 1000 examples from the `Multitarget-CONAN` dataset of hate speech/counterspeech pairs (Fanton et al., 2021) with labels based on strategies developed by experts. We then conduct a benchmark classification experiment to investigate the capacity of LLMs to distinguish between the strategies used.

2 Background

de Gibert et al. (2018) define **hate speech** as "any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic." While hate speech may constitute only a small proportion of social media content, nearly one third of the population is affected by it (Vidgen et al., 2019), necessitating research into its prevention and mitigation.

Contrary to traditional content removal, which may be considered to impinge upon freedom of speech, the idea of responding with **counterspeech** has gained ground. Another advantage of this approach is that its use is unbound by the intricacies of what constitutes hate speech according to the disparate platform guidelines. Cepollaro et al. (2023) define counterspeech as "communication that tries to counteract potential harm brought about by other speech." Real-world studies report counterspeech as an effective technique to counteract hate speech (Mathew et al., 2019). For example, Buerger (2021) elicits improvements in discourse in online comment sections through the application of carefully drafted counterspeech, and social media platforms like Facebook are reportedly investigating its application (Osman, 2022).

Prior research has illuminated the varied effectiveness of counterspeech **strategies** in mitigating hateful conduct (see also Section 3.2) (Benesch et al., 2016; Chung et al., 2023). However, this work has thus far focused on empirical investigation of manually crafted counterspeech interventions (e.g. Hangartner et al., 2021).

We seek to introduce the strategies developed by social scientists and policy experts to the NLP

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 256–265

Strategy	Definition	Examples
Positive Tone, Empathy and Affiliation	This strategy involves connecting on a personal level, showing understanding or solidarity with the speaker or target. Look for friendly, empathetic language.	1. I understand why this topic is upsetting. Let's find a solution together. 2. Migrants need help. They flee to find better living conditions.
Fact-Checking	Addresses inaccuracies or false claims by presenting facts. Look for use of verifiable facts or simple corrections.	1. Statistics show crime rates have decreased. 2. From what I know only a minority of the Gypsy population live in shanty towns.
Humour/Sarcasm	Uses wit, jokes, or sarcasm to counter hate speech, often lightening the conversation's tone. Identify humour by the playful or ironic twist in the counterspeech.	1. If believing in equality makes me a 'snowflake', then I'm ready for a blizzard! 2. Really? I thought it was due to the salaries of the players. But of course it's the same old Jewish conspiracies fault.
Warning of Consequences	Highlights potential negative outcomes of hate speech, like social or legal consequences. Recognize it by alerts or cautionary advice.	1. Remember, spreading hate can lead to serious consequences, not just online but in real life too. 2. It is also quite dangerous to say something so strong without proof.
Denouncing	Expresses outright rejection of the hateful views and may call out the hate speech by directly labelling it as racist, sexist, cause for discrimination etc.	1. Hate has no place in our community. 2. The mere existence of a minority is not a reason to target it. There is no need to be racist.
Pointing Out Hypocrisy	Underlines logical flaws or double standards in the hate speech. Identifies and questions inconsistencies, or presents contradicting or hypocritical positions in the hate speech.	1. Ironic, you advocate for free speech but silence those who disagree with you? 2. Imagine if someone of another religion had power over you this way. Would you rather have that person's power over you or not?
Questioning	Asks questions that prompt reevaluation of the presented views or statements. Characterised by questions that challenge the assumptions or generalizations in hate speech or use of rhetorical or direct questions aiming to provoke thought or self-reflection.	1. What exactly is your fear about sharing public places with people of a different religion? 2. When you say niggas are enemies of the people, who exactly are 'the people'?

Table 1: Seven strategies to counter hate speech with definitions and examples. These also serve as (refined) annotation guidelines.

counterspeech research community by implementing a combination of manual and automated strategy annotations on the hate speech-counterspeech dataset presented by Fanton et al. (2021) (see also 4.1). We create seven label classes based on the strategies discussed in the literature (Benesch et al., 2016; Chung et al., 2023). Table 1 provides a summary of these strategies, along with examples. The choice of strategies is supported by the complexity and variety observed in niche-sourced (that is, expert-produced) counterspeech data (Tekiroğlu et al., 2020), akin to the one in our research.

We conducted an annotator feedback survey after the annotation pilot study which revealed that most annotators find *Denouncing* to be the most confusing strategy, frequently mistaking it for *Shaming and Labelling* due to similar elements of 'rejecting hate'. Therefore, we merge *Denouncing* and *Shaming and Labelling* strategies for the next phase of annotation. Moreover, from the strategies proposed by annotators in their feedback, our analysis identified the inclusion of *Questioning* as necessary, and consequently incorporated it into the strategies considered in our study. See also Section 4.2 and Appendix B.2 for details of the annotation process, including the changes made to the guidelines based on annotator feedback.

3 Related Work

Two recent works provide a comprehensive overview of the social and technical challenges of using counterspeech to counter toxic content.

The first, a systematic review of work from multiple fields by Chung et al. (2023) identified eight strategies that have been used in counterspeech studies in the social sciences and real-world policy-driven campaigns. They also summarised the evidence of the effectiveness and efficacy of these strategies, which suggests that some approaches may provide better results in certain circumstances, but that this is highly context dependent.

For a more technical perspective meanwhile, Bonaldi et al. (2024) survey NLP methods and datasets for counterspeech generation, finding a range of approaches to collecting data from crowdsourcing to nichesourcing responses—that is, har-

nessing the knowledge of experts trained in countering online hate.

One of the most widely used nichesourced datasets is that of Fanton et al. (2021) who present a dataset of 5003 hate speech/counterspeech pairs on multiple targets of hate curated using an innovative combination of language model generation and expert review and post-edit. We annotate a subsection of this data with strategy labels (see also Section 4.1). The only work we are aware of to have previously analysed the strategies present in a dataset is that of Chung et al. (2019), who recruited non-expert annotators to label the response types in the CONAN dataset. We extend this work by developing and testing an annotation scheme and guidelines and exploring automated identification of these strategies.

3.1 Application of Large Language Models

Qian et al. (2019) were among the first to experiment with automated "generative intervention" in hate speech using a Seq2Seq encoder-decoder model, a Variational Auto-Encoder model and Reinforcement Learning. Tekiroğlu et al. (2020) propose the use of NLG for automated intervention and depict large language models as a promising alternative to manual intervention through their use of the GPT-2 language model to produce counterspeech and the model fine-tuned on an expert-generated counterspeech dataset secured a higher novelty score. A notable aspect is that their experimental automatic classifier showed better results over human filtering.

Tekiroğlu et al. (2022) compare the performance of various language models to determine the most suitable model for counterspeech generation using the Multitarget-CONAN (Fanton et al., 2021). They find that automatic post-editing using machine translation with a fine-tuned GPT-2 model improves the quality of generated responses, eliminating the need for manual post-edit effort.

Ashida and Komachi (2022) use few-shot prompting to present quantitative analysis of length, diversity, and quality of counterspeech across several models. While they find GPT-3 to produce responses of relatively high quality, most outputs are found to present facts to counter hate. Therefore, they acknowledge the potential for generating strategic counterspeech and leave that for future work, which we begin to explore in our study.

3.2 Counterspeech Strategies

Most research to date is found in the social sciences and policy literature and focuses on real-world and usually non-automated (i.e. human-written) interventions. Hangartner et al. (2021) show the potential role of empathy in effectively mitigating hate speech. Other studies also provide results on relative efficacy of various counterspeech strategies (Bilewicz et al., 2021; Carthy and Sarma, 2023; Obermaier et al., 2023). Lasser et al. (2023) substantiate *Opinionating* without insults, sarcasm or negative tone in general to be effective in mitigating toxicity in online hate speech. Overall, evidence from these studies indicates that a strategy framework is important for effective counterspeech.

Thus far, there has been little exploration of these strategies in the NLP literature. The closest we find are those of Chung et al. (2019) (see above) and of Tekiroğlu et al. (2020), who refer to strategies as 'counterspeech argument types' and present a comparison of variety in argument types across crowd, niche, and crawl-sourced data. In niche (expert)-sourced data, they observe higher complexity and variety in arguments. Therefore, this study relies on niche-sourced data for counterspeech strategy identification.

Recent studies have highlighted the potential of LLMs as classifiers for text-based tasks. Møller et al. (2024) assessed LLMs for automated text annotation, finding promising results but lacking the depth of human annotations. Conversely, Zhang et al. (2024) demonstrated superior performance of LLMs over human efforts through iterative fine-tuning in text classification. Further investigations have applied LLMs to other tasks like news classification (Zhang et al., 2024; Zhao and Yu, 2024) and legal text annotation (Savelka, 2023). In our study, we extend these investigations to the complex and subjective challenge of classifying counterspeech into seven strategy labels. We employ human annotation to assess the intricacy of this task and to provide a benchmark for automated classification using GPT-3.5. Our goal is to evaluate the performance of the LLM in counterspeech classification. The human annotation primarily aims to gauge the complexity of the task, serving both as a benchmark for automated classification and as a dataset for future fine-tuning and strategy-guided counterspeech generation, rather than to compare human and automated labeling directly.

4 Method

4.1 Data

`Multitarget-CONAN` (Fanton et al., 2021), is a dataset of hate speech/counterspeech pairs with respect to eight targets of hate, curated using a human-in-the-loop generation-review pipeline in which reviewers were trained annotators who reviewed and/or post-edited the counterspeech interventions, which were then iteratively fed back to `GPT-2` as training data. Our preliminary analysis of the dataset found sufficient diversity and examples of the key strategies identified by Benesch et al. (2016) and Chung et al. (2023). We sampled 1000 examples (approximately 20% of the dataset), equally representing all targets of hate, for counterspeech strategy annotation. We make all data available on acceptance.

4.2 Counterspeech Strategy Annotation

Overview We formulated an annotation framework by consolidating the strategies delineated by Benesch et al. (2016) and Chung et al. (2023) with guidelines for each strategy including definitions, the key characteristics associated with each strategy, and examples drawn from the specifications of Benesch et al. (2016). In a pilot study, we initially recruited ten annotators to label 350 examples. Observing low agreement among non-expert annotators, we collected annotator feedback and refined the annotation guidelines (Table 1) and trained two of the annotators. The two trained annotators and the first author then labelled the full set of 1000 examples. This iterative approach resulted in the current dataset, validated through measures of inter-annotator reliability outlined in section 4.2 below.

Inter-Annotator Agreement Evaluation To measure inter-annotator agreement, we utilised (1) Cohen's *kappa*: a statistic for inter-annotator and intra-annotator reliability testing for pairs of annotators (McHugh, 2012); (2) Fleiss' *kappa*: adaptation of Cohen's *kappa* for three or more annotators (McHugh, 2012); and (3) raw agreement percentages for completeness. We also used Cohen's *kappa* to showcase the inter-annotator agreement per strategy. Tables 5 and 6 show the range of values and their reliability indication for Cohen's κ and Fleiss' κ.

Annotation process We recruited 10 annotators from among university peers and colleagues to label 350 examples, which were partitioned into sets of 50 and labelled by pairs of participants (see also Appendix B.2.1). See Appendix A for a full Data Statement.

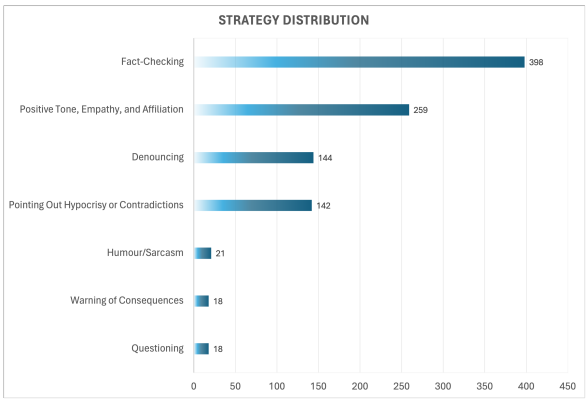

Figure 1: Distribution of strategies in our final dataset.

Observing low agreement (Cohen's $\kappa = 0.15$; 37.4%), we refined the final guidelines to produce Table 1 (see Appendix B.2 for details of these changes) and trained two of the non-expert annotators to address comprehension gaps in the key indicators for each strategy. The two trained annotators and one of the authors of this paper then labelled the full set of 1000 examples (see also Appendix B.2.3).

4.3 Automated Classification

To investigate the potential of generative large language models in classifying counterspeech strategies, we benchmarked the dataset with *one-shot* prompting of a `GPT-3.5` model. For this, we aggregate annotator responses by majority vote between the three trained annotators. We include the classification prompt in Appendix C.1 for reproducibility.

5 Analysis

Figure 1 illustrates the distribution of strategies in the dataset, where we can see clear preferences of the nichesourced reviewers/editors towards certain response types. *Fact checking* is the most prevalent strategy despite the fact that it is not thought to be effective due to people's cognitive biases.

Annotation We report Cohen's *kappa* (κ) and raw percentage agreement for annotator pairs, as well as per-strategy agreement.

Comparing the inter-annotator agreement between our two trained annotators on the 100 examples that they labelled both before and after receiving training and the adjustments to the labelling scheme and guidelines, we observe an

Strategy	Cohen's κ
Questioning	0.72
Hypocrisy & contradictions	0.61
Humour/sarcasm	0.59
Positive tone, empathy, affiliation	0.57
Warning of consequences	0.56
Fact checking	0.55
Denouncing	0.52

Table 2: Strategy-specific inter-annotator reliability.

Acronym	Strategy
FC	*Fact-Checking*
PEA	*Positive Tone, Empathy, and Affiliation*
DG	*Denouncing*
PHC	*Pointing Out Hypocrisy or Contradictions*
QG	*Questioning*
WC	*Warning of Consequences*
HS	*Humour/Sarcasm*

Table 4: Legend for Counterspeech Strategies

improvement in Cohen's κ from 0.12 to 0.58, highlighting the effectiveness of these interventions. For the full dataset, we observe agreement of $\kappa = 0.56$ (67.9%) between the trained annotators, commonly interpreted as 'moderate' agreement (McHugh, 2012). However, we observe large strategy-specific variations (Table 2). Additionally, we calculated Fleiss' *kappa* between all three annotator labelling, which yielded a value of 0.46, also indicating 'moderate' agreement. Results indicate that, while the annotation task is not trivial, consensus can be reached.

Automated Classification We report the performance of the `GPT-3.5` automated classifier based on three metrics: precision, recall, and F1 Score. The macro-averaged results are shown in Table 3 alongside the majority class baseline. For a breakdown of scores by strategy class, see Figure 2.

Metric	Majority Class	Classification
Precision	0.40	0.70
Recall	0.10	0.62
F1	0.57	0.62

Table 3: Comparing automated classification results alongside the majority class baseline metrics

Compared to the baseline, these results suggest a reasonable capacity to identify and categorise counterspeech strategies and suggest potential for LLM-driven counterspeech interventions.

To further understand which strategies are handled well by the model and which ones pose a challenge, we present a breakdown of scores by counterspeech strategy in Figure 2. The strategies are abbreviated as shown in Table 4.

6 Conclusion

We have conducted an exploratory study to enhance our understanding and application of counterspeech strategies in NLP. By annotating a dataset with seven prominent strategies, and investigating their classification with an LLM, we contribute to the ongoing research on combating hate speech online by providing a validated strategic counterspeech dataset for training and testing automated counterspeech techniques. Inter-annotator agreement analysis on the dataset indicate 'moderate' to 'substantial' agreement among annotators across the counterspeech strategies, validating the reliability of the annotated dataset. The evaluation of the automated classifier, employing a *one-shot* prompted `GPT-3.5` model yielded a promising F1 Score of 0.62. While the results indicate an encouraging start, they also highlight areas for improvement , particularly in increasing *recall* without compromising on *precision*.

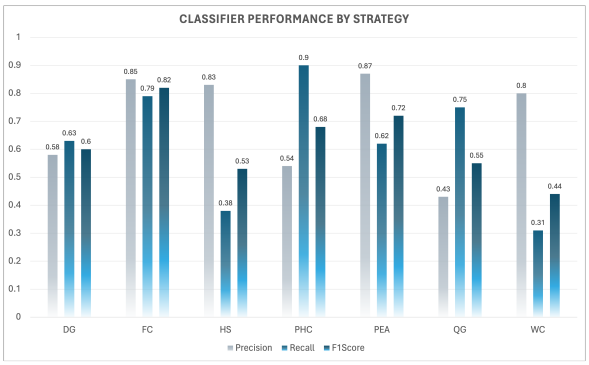

Figure 2: Performance by counterspeech strategy.

In future work, we aim to explore more sophisticated prompting strategies, expansion and enhancement of the strategic counterspeech dataset, and counterspeech generation using models fine-tuned on the dataset to generate nuanced and targeted strategy-driven counterspeech.

Limitations

Multi-annotator labelling revealed a low Cohen's κ score reflecting challenges in achieving consensus among annotators. Although subsequent refinements and training improved reliability, this observation underscores the difficulty of classifying counterspeech strategies. It potentially necessitates further refinement to create more nuanced guidelines and more extensive training for annotators.

Our dataset encompasses 1000 examples. The relatively limited size of the dataset may pose a challenge to the general applicability of our findings. While our sample was chosen to equally represent multiple targets of hate, some counterspeech strategies are under-represented in the resulting annotated dataset. While this likely reflects real-world occurrences, where certain strategies such as *fact checking* are more frequently utilised than others, this limitation presents a challenge for future research since generating nuanced strategy-driven counterspeech of adequate quality may require datasets with sufficient examples for each strategy. In addition, our current selection does not provide an exhaustive list of effective strategies. The evolving nature of online discourse calls for the expansion of counterspeech strategies.

The automated classification performance highlights potential for improvement in precision and recall. The model's performance reflects the current limitations of language models in capturing the intricacies of human language. This points to the ongoing need for enhancements in NLP technology and continual expert involvement in the development of automated solutions.

Our study focuses on the classification of counterspeech strategies without evaluating their relative efficacy in mitigating hate speech. The association between strategies and their effectiveness in different contexts is an important area for future NLP research.

We acknowledge that our use of closed-source commercial language models could impact reproducibility. However, these experiments are preliminary investigations into the application of language models for counterspeech strategy classification and future work will explore reproducible methods.

Ethical Considerations

Our study and experiments have been approved by our institute's Research Ethics Committee (reference on acceptance).

Since our experiments involved human exposure to potentially upsetting content, we took the following mitigation measures:

- Participants were informed about the nature of the task and warned about potential distress due to the offensive language in the data (1) in the Information Sheet and (2) in the Consent Form again.
- Participants had to provide consent and affirm that they had no physical disabilities, mental health issues, or any other conditions that might potentially negatively affect their well-being through participation in the study.
- Participants could withdraw from the study at any time.
- Each participant was allocated a small subset of the data, an average of 50 examples, and a generous time frame, averaging more than two weeks to mitigate prolonged exposure to potentially distressing language.

Chung et al. (2023) raise the concern of 'dual-use' in automated counterspeech where the same technology could be used against legitimate voices. To avoid this, hate speech detection algorithms should be accurate and unbiased. Also, counterspeech interventions should consider diverse parameters including speakers, recipients, and medium of communication, and evaluation should also assess social impact for a more comprehensive understanding of the potential impact of counterspeech (Chung et al., 2023).

7 Acknowledgements

Gavin Abercrombie and Ioannis Konstas were supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1).

The authors gratefully acknowledge the support and contributions of annotators, Abin Paul, Akshay Rajeev, Amrutha Purna Vadrevu, Gokul Kunnathuvilagam Padmakumar, Jane Bejoy, Kendal McDonald, Mariya Sebastian, Sachin Sasidharan Nair, and Shibin Jayaram Sajini, for their diligent efforts in the counterspeech strategy annotation pilot study, with particular appreciation to Abin Paul and Sachin Sasidharan Nair for their standout contributions to the trained-annotator labelling. We also thank Dr. Phil J. Bartie, for his resourceful insights on large language models.

References

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.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for successful counterspeech. *Dangerous speech project*.

Michał Bilewicz, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive behavior*, 47(3):260–266.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and *how-to* guide. In *Findings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Catherine Buerger. 2021. # iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media+ Society*, 7(4):20563051211063843.

Sarah L Carthy and Kiran M Sarma. 2023. Countering terrorist narratives: Assessing the efficacy and mechanisms of change in counter-narrative strategies. *Terrorism and Political Violence*, 35(3):569–593.

Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson. 2023. Counterspeech. *Philosophy Compass*, 18(1):e12890.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation. *arXiv preprint arXiv:2307.04761*.

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.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Brian Dean. 2024. Social media usage & growth statistics. https://backlinko.com/social-media-users#social-media-usage-stats. Online; Accessed 06 March 2024.

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.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jana Lasser, Alina Herderich, Joshua Garland, Segun Taofeek Aroyehun, David Garcia, and Mirta Galesic. 2023. Collective moderation of hate, toxicity, and extremity in online discussions.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. Association for Computational Linguistics.

Magdalena Obermaier, Desirée Schmuck, and Muniba Saleem. 2023. I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society*, 25(9):2339–2358.

Nawab Osman. 2022. Expanding Counterspeech Initiatives Into Pakistan and the UK. https://about.fb.com/news/2022/02/facebook-counterspeech-in-pakistan-uk/. Online; Accessed 06 March 2024.

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.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 447–451, New York, NY, USA. Association for Computing Machinery.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there. *Alan Turing Institute*, 11.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of LLM capacity for text classification.

Fengxiang Zhao and Fan Yu. 2024. Enhancing multiclass news classification through bert-augmented prompt engineering in large language models: A novel approach. In *The 10th International scientific and practical conference "Problems and prospects of modern science and education"(March 12–15, 2024) Stockholm, Sweden. International Science Group. 2024. 381 p.*, page 297.

A Data Statement

We collected annotator information to document the Data Statement for the counterspeech strategy classification undertaken as part of this study as recommended by Bender and Friedman (2018).

Curation Rationale The data used in our study is a subset of `Multitarget-CONAN` curated by Fanton et al. (2021). It was selected for the reasons outlined in 4.1.

Language Variety `en-UK`, `en-US`

Author Demographic Unknown

Annotator Demographic Annotator demographics for the counterspeech strategy classification, including individual annotation, are as follows:

- Age: 18 – 54
- Gender: Male: 6 (55%); Female: 5 (45%)
- Ethnicity: Asian 9: (82%); British: 2 (18%)
- Language Proficiency:
  - Fluent – Native: 7 (64%)
  - Intermediate – Advanced: 4 (36%)
- Training or experience in relevant disciplines: Yes: 2 (18%); No: 10 (82%)

Task Situation The annotations were conducted between February – March 2024.

Text Characteristics Hate speech and counterspeech pairs concerning eight targets of hate (see also 4.1), along with annotated counterspeech strategies.

Provenance Data statement was not available for the original dataset.

B Counterspeech Strategy Annotation

B.1 Annotation Framework

We provided a concise version (similar to Table 1) of the original comprehensive annotation framework, comprising the strategies – *Fact-Checking*, *Positive Tone, Empathy, and Affiliation*, *Denouncing*, *Shaming and Labelling*, *Pointing Out Hypocrisy or Contradictions*, *Warning of Consequences*, and *Humour/Sarcasm*, for the multi-annotator labelling pilot study. The following reasons underpinned this decision: **(1)** Peer annotators, primarily non-experts, with limited time, required concise guidelines to effectively engage in the task. **(2)** Condensed format provided quick and accessible reference, and expedited the initial training process. **(3)** The initial round of annotation aimed to elicit subjective perspectives and improve guidelines by incorporating feedback based on 'descriptive dataset paradigm' (Rottger et al., 2022).

B.2 Annotation Process

B.2.1 Multi-Annotator Labelling

We attribute the following potential reasons for *none-slight* agreement among annotators in our pilot study based on 350 examples:

1. Complexity of the task: ambiguity in class definitions or the highly subjective nature of the task may have contributed to divergent annotations.

2. Cultural and interpretational differences: diverse perspectives and cultural backgrounds may have influenced their understanding and classification of instances.

3. Expertise and training: limited expertise in or exposure to counterspeech may have led to inconsistencies in annotation.

4. Language fluency and communication: variations in English fluency levels and communication skills may have impacted their ability to accurately classify instances.

B.2.2 Annotator Feedback Survey

Key observations from the annotator feedback survey were:

1. Annotators expressed interest in the addition of specific strategies: *Questioning* (1), *Educating* (2), *Drawing Parallels* (1), and *Positive Tone* (1).

2. Annotators identified *Denouncing* as the most confusing, cited by six annotators, followed by *Shaming and Labelling* (4) , *Warning of Consequences* (2), and *Pointing Out Hypocrisy or Contradictions* (2).

3. Annotator preferences for counterspeech strategies in their application to mitigate hate speech: *Fact-Checking* (6), *Positive Tone, Empathy, and Affiliation* (5), *Warning of Consequences* (1) and *Denouncing* (1).

Based on this feedback, we: (1) incorporated *Positive Tone* alongside *Empathy and Affiliation*, and (2) consolidated the categories of *Shaming and Labelling* and *Denouncing* into a single category, retaining 'Denouncing' as the overarching classification that encompasses 'Shaming and Labelling'. Also, *Questioning* emerged as an indispensable category based on our analysis following annotator feedback and was later added to the framework.

B.2.3 Trained-Annotator Labelling

As part of training, confusing terms and definitions were translated into the native language of the annotators, accompanied with further clarifications. For instance, *Fact-Checking* was defined as countering hate speech with factual or general educational rebuttals, while *Denouncing* involved morally critiquing hate speech, with or without shaming and labelling its speaker.

Trained annotators were not directly provided with the *Questioning* label to avoid potentially mislabelling counterspeech framed as questions without due consideration of semantics. However, if annotators strongly believed that none of the existing strategies applied, they were encouraged to designate an external category such as *Educating*, *Questioning*, or use of *Hostile Tone*. This allowed for subsequent analysis to assess the necessity of such strategies and their incorporation.

B.2.4 Inter-Annotator Agreement Measures

Table 5 and Table 6 show the interpretation of Cohen's κ based on McHugh (2012) and Fleiss' κ based on Landis and Koch (1977) respectively.

Cohen's κ	Agreement Level
≤ 0	No agreement
$0.01 - 0.20$	None to slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost perfect

Table 5: Interpretation of Cohen's *kappa*

Fleiss' κ	Agreement Level
< 0.00	Poor
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost Perfect

Table 6: Interpretation of Fleiss' *kappa*

C Automated Classification

C.1 Automated Classifier Prompt

Table 7 shows the *one-shot* prompt used for the automated classification experiment with `GPT-3.5` for 700 examples of the dataset chosen for the experiment.

Section	Content
Preamble	You are a helpful assistant and an expert at strategy classification.
Example	Given the hate speech: "I can't believe migrants like you are allowed to be here." and the counterspeech: "Everyone has the right to be here. Discrimination based on identity is harmful and unjust.", classify the counterspeech into one of the following strategies: (A)Fact-Checking (B)Positive Tone, Empathy, and Affiliation (C)Humour/Sarcasm (D)Denouncing (E)Pointing Out Hypocrisy or Contradictions (F)Warning of Consequences (G)Questioning The answer should be only the letter corresponding to the correct strategy option. answer: (D)
Instruction	Given the hate speech: "Faggots are poisoning our children minds with their filthy depravity." and the counterspeech: "Isn't this just equating homosexuality and pedophilia? How can this be allowed? This is about intolerance and prejudice spreading it's negative influence throughout all parts of our society.", classify the counterspeech into one of the following strategies: (A)Fact-Checking (B)Positive Tone, Empathy, and Affiliation (C)Humour/Sarcasm (D)Denouncing (E)Pointing Out Hypocrisy or Contradictions (F)Warning of Consequences (G)Questioning The answer should be only the letter corresponding to the correct strategy option. answer:
Model Response	**(E)**

Table 7: *One-shot* prompt used in the automated classification experiment and corresponding response.