On the Effectiveness of Adversarial Robustness for Abuse Mitigation with Counterspeech

Content Warning: This paper contains examples of offensive language.

Yi-Ling Chung and **Jonathan Bright**
The Alan Turing Institute
ychung@turing.ac.uk, jbright@turing.ac.uk

Abstract

Recent work on automated approaches to counterspeech have mostly focused on synthetic data but seldom look into how the public deals with abuse. While these systems identifying and generating counterspeech have the potential for abuse mitigation, it remains unclear how robust a model is against adversarial attacks across multiple domains and how models trained on synthetic data can handle unseen user-generated abusive content in the real world. To tackle these issues, this paper first explores the dynamics of abuse and replies using our novel dataset of 6,955 labelled tweets targeted at footballers for studying public figure abuse. We then curate DynaCounter, a new English dataset of 1,911 pairs of abuse and replies addressing nine minority identity groups, collected in an adversarial human-in-the-loop process over four rounds. Our analysis shows that adversarial attacks do not necessarily result in better generalisation. We further present a study of multi-domain counterspeech generation, comparing Flan-T5 and T5 models. We observe that handling certain abuse targets is particularly challenging.

1 Introduction

Online abuse is a significant societal challenge, with public figures often bearing the brunt of its toxic impact. Being exposed to such abusive behaviour can have detrimental effects on the mental well-being of victims and even on bystanders who witness it (Saha et al., 2019; Siegel, 2020; Chung et al., 2023).

The approach of using *counterspeech* (or *counter-abuse*) to directly resist abusive or harmful content has received considerable interest for understanding its effective usage in real-life scenarios (e.g. Saltman and Russell, 2014; Carthy et al., 2020; Fraser et al., 2021) and automating techniques (e.g. Tekiroğlu et al., 2022; Ashida and Komachi, 2022).

While counterspeech is promising, the research field faces many challenges. From a theoretical perspective, most existing counter-abuse datasets are synthetically created rather than collected from real-world events (Chung et al., 2019; Fanton et al., 2021), making it hard to observe the interplay between abuse and responses in real-life scenarios, for instance, how abuse is responded to and what counterspeech constitutes (Garland et al., 2020). From a computational perspective, varied and challenging data is required to develop automated systems that can respond to abuse across targets, tropes, and domains. Models trained on synthetic content may not be able to deal with, characteristically different, user-generated (or authentic) content. We show different types of counterspeech in Table 1.

Abusive content: [Player] You are useless mate please leave our club, possibly the worst defence I have ever seen at my club.
Authentic counterspeech 1: Hold Dat. lol you proved that you don't know anything about football. People defending him as well mate. #bbcfootball
Authentic counterspeech 2: You obviously see what you want to see. For Christs sake he was sensational for us 90% of his time here... let him be happy.
Synthetic counterspeech: What exactly are you talking about? It is not appropriate to spread abuse! Let's create a safe space for everyone regardless of their background.
Adversarial counterspeech: I'm vegan, I don't shoot anything with a face. Bruh/sis, we ain't got no use fo' abuse 'round here. Let's build each other up wit love an' respect. #LoveNotAbuse #RespectEachOther.

Table 1: Abuse example and three types of counterspeech. Prior work addresses synthetic counterspeech that is characteristically different from authentic one. This paper explores authentic counterspeech compared to synthetic examples, and investigates the impact of adversarial attack on model robustness.

Prior studies show that existing models suffer from performance instability in real-world scenarios (Fortuna et al., 2021; Kiela et al., 2021), and adversarial model-attacking approach can improve model robustness on several tasks (Kiela et al., 2021; Kaushik et al., 2021; Wallace et al., 2022).

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6985–6999

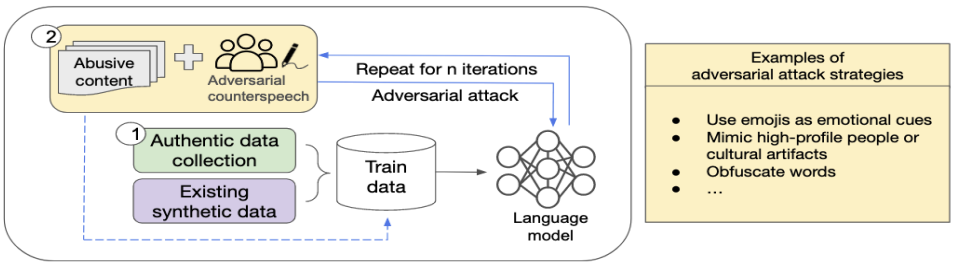


Figure 1: A diagram illustrating our method. We collect two types of counterspeech: (1) authentic and (2) adversarial data through adversarial attacks. The collected data is used to evaluate the transferability of counterspeech generation models across various abuse targets.

Considering the scenario of counterspeech classification, a challenging task in which authentic data is diverse and differs from the synthetic counterpart, we hypothesize that adversarial training could help models generalize beyond the original distribution.

In this work, we investigate the extent to which the robustness of counterspeech models benefits from adversarial examples. In collaboration with a civil organisation, we curate data and develop adversarial classifiers for counterspeech that can handle multi-domain abuse. The evaluation of the adversarial classifiers is conducted on a newly collected authentic dataset from Twitter/X, adversarial data, and existing synthetic multi-domain datasets. We also characterise the difference between synthetic and authentic datasets. To assess the transferability of counterspeech generation models across various abuse targets, we further analyse the performance of two large language models in generating counterspeech responses. We illustrate the process in Figure 1.

The main contributions are: (1) an analysis of how abuse against footballers is responded to based on three levels of annotations: whether a reply disagrees with the abuse, whether the reply supports the targets of abuse and whether the reply is abusive; (2) a **Dyna**mic adversarial **Counter**-abuse dataset, DynaCounter, covering nine new domains; (3) a series of adversarial attacks on counterspeech classifiers; and (4) an extensive automatic and human evaluation of multi-domain counterspeech generation to identify model weaknesses.[1]

2 Related Work

Counterspeech. aims to encourage opinions exchange (Benesch, 2014; Stroud and Cox, 2018) and can empower users to respond assertively to abuse (Bilewicz et al., 2021; Hangartner et al., 2021). Reviews on counterspeech studies are available, covering various aspects such as the impact (Carthy et al., 2020; Chung et al., 2023) and NLP approaches to counterspeech classification and generation (Alsagheer et al., 2022; Bonaldi et al., 2024).

Counterspeech datasets are generally curated in two ways: user-generated comments on social media (Mathew et al., 2018; Garland et al., 2020) and responses intentionally crafted by crowdworkers (Qian et al., 2019), experts (Chung et al., 2019) or language models (Fanton et al., 2021). Most of the extant popular datasets are collected synthetically and at the level of individual posts (single-turn dialogues, e.g. Chung et al., 2019; Fanton et al., 2021), making it hard to investigate the dynamic of abuse and responses. Recently, Yu et al. (2023) address this challenge by first proposing a taxonomy for annotating user-generated counterspeech and then collecting a dataset targeting men's rights, seduction, and gender issues on Reddit. The increasing emphasis on authentic counterspeech collection also indicates the importance of understanding how counterspeech is used in the real world. In this paper, we look into the use of counterspeech for abuse against public figures - footballers - on X.

Counterspeech Generation. has addressed various aspects such as politeness (Saha et al., 2022), personalisation through author profiling (Doğanç and Markov, 2023), generation in languages other than English via data augmentation (Chung et al., 2020; Furman et al., 2023; Vallecillo-Rodríguez et al., 2023), incorporating knowledge for informative responses (Chung et al., 2021b; Jiang et al., 2023), and intent-aware generation (Gupta et al., 2023). Additionally, Ashida and Komachi (2022) approaches counterspeech generation for implicit

[1]Dataset, guidelines and code are at https://github.com/Turing-Online-Safety-Codebase/counterspeech_adversarial

offensive text using prompt engineering. However, these generation models are built and tested on synthetic data, leaving open questions regarding their generalizability in real-life scenarios for effectively mitigating abuse.

Adversarial Attack and Testing. The core of adversarial attacks lies in iteratively probing model weakness in a human-and-model-in-the-loop setting. It has been shown to be effective in several domains including dialogue systems (Niu and Bansal, 2018; Dinan et al., 2019), abuse detection (Kiela et al., 2021) and sentiment analysis (Potts et al., 2021). The closest work to our adversarial robustness analysis is Fanton et al. (2021), who employ annotators to post-edit generated text for data collection. In contrast, our work focuses on testing model weaknesses and covering minority groups (e.g. Native American, Asian and Mexican).

3 PLF Dataset: Authentic Counterspeech Collection and Annotation

To our knowledge, there is no abusive language data that targets public figures and is paired with authentic counterspeech. We start our study by collecting such data, focusing on abuse against Premium League Footballers (PLF) in the UK on Twitter/X.[2] The data collection is conducted based on three steps, following the work done by Vidgen et al. (2022). Firstly, we collect 3,127,640 tweets directed at PLF in the period of 08/08/2021 to 01/04/2022, using Twitter API v1.1. The tweets are retrieved based on a list of player accounts from England's top football divisions (808 from the Men's Premier League).

Secondly, we apply an abuse classifier to automatically identify abusive tweets using Footballers Personal Attacks Classifier (Vidgen et al., 2022), which is suitable because it also targets abuse against footballers. The abusive tweets returned are reviewed by the authors for quality control, resulting in 4,556 abusive tweets in total.

The last step is to collect responses to the abusive tweets, i.e. a response as a direct reply to abuse. We collect the first twenty replies to each abusive tweet and discard the replies that are less than 10 words. Retweets and replies to replies are not included. This data allows us to characterise how abuse is responded to online.[3]

3.1 Dataset Annotation

Our annotation scheme is based on three levels (see instruction in Appendix A.1). Given an abusive tweet, we ask (1) the **strategy** of the reply (Disagree/Agree/Other), (2) whether the reply **supports** the victims (Yes/No) (3) whether the reply is **abusive** (Yes/No). Following Vidgen et al. (2022), abuse is defined as content which threatens, derogates (e.g. insults or the hateful use of slurs or negative use of stereotypes), dehumanises (compares individuals to insects, animals or trash), mocks or belittles an individual or their identity. A reply is not considered abusive if it only criticises the group or attacks abstract concepts and institutions.

We employ crowdworkers and expert annotators in order to achieve a balance between annotation quality and efficiency. Firstly, 2,154 crowdworkers from Appen were enlisted to annotate the replies in the dataset with each instance labelled by three annotators.[4] In cases where no consensus was reached, an additional two annotations were solicited, or one of the authors finalises the label if still no majority agreement was reached. The average inter-annotator agreement for each category was 79.83% (strategy), 88.15% (support) and 88.45% (abusive).

3.2 Dataset Analysis

In total, there are 6,956 pairs of abusive tweets and replies, with around 2 replies collected per tweet (see Table 2). The class distribution is highly imbalanced for each category. We find that most of the replies disagree with the abusive tweets (62%) using non-abusive language (93%) while not showing support for the targets of abuse, i.e. footballers (93%). Through our post-hoc analysis, we found that most replies challenge or denounce the perpetrators without directing the message to the targets (e.g. *Can you support football and stop being an abuser?*). While this finding suggests civil conversations online, it may be attributed to the unique nature of football's popularity.

4 DynaCounter Dataset: Dynamic Adversarial Counterspeech Collection

Our main goal is to assess if adversarial examples can improve the robustness of counterspeech classifier across multiple domains. Accordingly, we

[2]For simplicity, we refer to the dataset as *PLF dataset* throughout the paper.

[3]Due to institutional guidelines concerning privacy issues surrounding the release of social media data, we are unable to release the PLF dataset.

[4]These are crowdworkers who passed a 100-question qualification test with 80% accuracy.

Aspects	Count	Percentage (%)
# abusive tweet	4556	-
# replies	6956	-
# replies per thread_avg	2.64	-
# replies per thread_std	3.56	-
reply strategy		
disagree	4296	62%
agree	2363	34%
other	297	4%
reply support		
yes	478	7%
no	6478	93%
reply abusive		
yes	453	7%
no	6503	93%

Table 2: Main statistics of the dataset.

explore dynamic adversarial attacks over multiple rounds on binary counterspeech detection: to determine whether a reply disagrees or agrees with an abusive comment.[5] Each round of adversarial attack consists of three main elements: (1) a trained binary counterspeech classifier, (2) a pool of abusive text as context and (3) a team of expert annotators. Given an abusive text selected from the pool, annotators are tasked with testing the classifier's capability by composing responses that can trick the classifier (i.e. writing an example of counterspeech that the model misclassifies as non-counterspeech), as shown in Appendix A.2. The resulting dynamic data collected after each round is divided into train/test splits of .9/.1 for model evaluation, respectively. With this dataset, we can analyse the characteristic difference between synthetic and authentic data (§3).

4.1 Task Setup

Binary counterspeech classifier. We consider five datasets targeting seven abuse targets to train a counterspeech classifier: CONAN (Chung et al., 2019), MTCONAN (Fanton et al., 2021), KN-CONAN (Chung et al., 2021b), Gab+Reddit (Qian et al., 2019) and PLF dataset (§3). The seven targets include footballers, women, mental disability, migrants, Muslims, Jews, and LGBT+. Note that these datasets except PLF are synthetically created. The datasets are selected because they are the ones available for counterspeech detection.

Each dataset is first divided on a .8/.2 split and then concatenated together respectively as train/test set which is randomly shuffled. Since these datasets mostly contain only counterspeech examples, we further created 10,000 pairs of abuse and pseudo non-counterspeech instances that support (i.e. abusive comments randomly selected from the dataset) or are irrelevant to the abusive text, following the work by Chung et al. (2021a). For the irrelevant examples, we randomly sampled English instances from Wikilingual (Ladhak et al., 2020) that focused on how-to guides, topics unrelated to abuse. This is to ensure that models encounter non-counterspeech, avoiding model overfitting and reflecting real-world scenarios full of noisy data.

We conduct four rounds of collection, following the work of Vidgen et al. (2021). Each round is equipped with a model in the loop that is trained on the train set plus adversarial data collected in the previous rounds. For round 0 (R0), we finetune DistilBERT (Sanh et al., 2019) on the train set as M0. For round 1, model M1 is trained on the train set with data collected in R0. All models are finetuned on a Tesla K80 GPT for 3 epochs with the default set of hyperparameters from Transformers (Wolf et al., 2020): a learning rate of 1e-3, a maximum text length of 512 tokens and a batch size of 32.[6] To better assess the effect of adversarial training, we compare our approach with a baseline *Base* which uses DistilBERT finetuned only on PLF dataset.

Abusive text pool. Since our aim is to develop a generalised counterspeech classifier, we use TOXIGEN (Hartvigsen et al., 2022), which is a synthetic dataset generated by language models, to collect adversarial examples. We employ TOXIGEN rather than PLF as context for the following three reasons. First, using synthetic datasets such as TOXIGEN generated by language models does not violate our institution guidelines. There are concerns on reposting copy-righted data with labels (e.g. Tweets labelled as abusive) that can be attributed back to individuals, who may contest the accuracy of the label. Reposting labelled social media data on a third platform (even accessed only by a few people for research purposes) would raise privacy harms based on our institution guidelines. Second, TOXIGEN consists of over 135k toxic statements covering 13 minority identity groups not included in the five datasets used for training counterspeech classifiers, allowing for expanding the coverage of

[5]The experiment is developed using Streamlit: https://streamlit.io/

[6]All experiments in the paper use the same hyperparameters and are done in a single run unless indicated otherwise.

the trained classifier to handle abuse against new targets. Third, generating adversarial examples for unseen domains can provide insights into whether adversarial training improves over the baseline.

To create context for adversarial counterspeech collection, in each turn we randomly select statements about groups not included in the train sets for counterspeech classifiers (see Table 3). If a statement has received an adversarial example, we will not show it to annotators again. Generated by language models, the statements in TOXIGEN can contain grammatical errors or be incomplete. However, we do not remove such suboptimal statements as this reflects discussions on social media in which errors and typos are common.

Annotators. We collaborated with ten members (5 females and 5 males) of a civil society organisation that specialises in the area to compose adversarial examples. All annotators are English native speakers and deal with abuse countering on a daily basis. Before the experiments, we set up a session with annotators to introduce the task and the platform. In the same session, they were asked to try all the functions of the platform. Each annotator was then given one hour per round, resulting in 40 hours for the entire data collection.

4.2 Common Adversarial Attack Strategies

We summarise the strategies used by the annotators for tricking models across four rounds.

- Include emojis as emotional cues or replacements for words.
- Use hashtags such as #noAbuse, #noDiscrimination, and #IndigenousPeopleRock (e.g. *Michael Jordan is Lit #mjalltheway)*.
- Express counterspeech in multiple languages.
- Use acronyms, slang or idioms, such as Bruh/sis, and dat ain't cool talkin (e.g. *Ditto dude, NIMBY*).
- Use abusive slurs/derogatory terms and describe how these terms are not appropriate. This will help draw limits between abuse and counterspeech (e.g. *Are you happy if I call you* `[derogatory terms]`*?*).
- Mimic or quote high-profile people or cultural artifacts such as books, songs and movies (e.g. *"Sorry, not sorry, but hate speech or abuse is cancelled. It's time to spread love and acceptance." - Demi Lovato*).
- Obfuscate slurs or other words to subvert the models (e.g. *Our str@ngths lie beneath the surface. H8n aint gr8, so let's all just celebr8 and appreci8 each other.*).
- Combinations of some or all above.

4.3 Analysing Adversarial Models

In total, 1911 adversarial examples are collected in DynaCounter, covering nine new targets. Statistics of adversarial data by targets are presented in Table 3. Each target receives similar amounts of replies. Similar to Vidgen et al. (2021), we assess how models perform on the gathered data in two scenarios: one during data collection via *model error rate* against the collected adversarial examples in each round, and the other via post hoc *test set performance* assessed on two test sets.

Model error rate. Table 4 shows the statistics of collected adversarial data and the model error rate in each round. We calculate the percentage of examples wrongly predicted by models as model error rate. In R1 the model classifies all entries as counterspeech. As the rounds advance, the model error rate decreases with R3 at the lowest (2% of counterspeech and 9% of non-counterspeech tricked). Annotators reported that fooling the models is easier in the early rounds as opposed to later rounds, thus showing that models have learned to solve the task, in line with previous work (Wallace et al., 2022). Overall, examples using emojis or hashtags had the highest model error rates (47%), whereas obfuscation had the lowest error (5%).

Group	Count	Percentage (%)
Mexicans	242	12.66
Chinese	235	12.30
Asian	213	11.15
Mental Disability	212	11.09
Black	208	10.88
Native Americans	206	10.78
Middle Eastern	205	10.73
Physical Disability	199	10.41
Latinos	191	9.99
Total	1911	100

Table 3: Statistics of DynaCounter across groups. Percentage denotes the ratio of a group as a fraction of 100.

Category	R1	R2	R3	R4
N				
Counterspeech	266	563	414	333
Not	181	42	42	70
Error rate				
Counterspeech	0%	15%	2%	5%
Not	40%	6%	9%	15%
Total	40%	21%	11%	20%

Table 4: The number (N) of adversarial data collected for DynaCounter and model error rate in each round.

Test set performance. In comparison with the baseline trained on PLF dataset (Base), we report in Table 5 the macro F1 of adversarial models evaluated on three test sets: PLF, non-adversarial test sets (i.e. *non-adversarial* which include all five test splits specified in §4.1) and DynaCounter (the test split of adversarial data collected in each round). By testing with different types of datasets, we show how adversarial learning boots or deteriorates classification performance.

Non-adversarial. The baseline achieves scores lower than but close to M0 (0.70 vs. 0.76), showing that models trained on authentic counterspeech (PLF data) can bring a step towards identifying counterspeech written by experts. Generally, adversarial models (M1-M4) achieve higher macro F1 than baseline, while we do not observe consistent performance gains over iterations.

PLF. Considering PLF dataset is different from synthetic data in nature (see discussion in §6 and Appendix A.3), we further single out the results of PLF test data from non-adversarial test sets to better evaluate the performance of adversarial models. Overall, the scores for PLF are in the low 60s, compared to the mid-70s in non-adversarial test sets, demonstrating that PLF data is more challenging.

DynaCounter. The baseline obtains an average F1 score of 0.46 across 4 rounds of test data in DynaCounter. This provides evidence that PLF can be considered adversarial in nature. Adversarial models achieve overall lower performance on test sets in the later rounds. This finding confirms that adversarial data collection results in progressively challenging datasets over time. For instance, the highest score for R1 is 0.80, compared to the high 0.68 for R3 and low 0.45 for R4. M1 and M4 obtain the same and best average scores across the four test sets (0.64). This suggests that a small amount of adversarial data is efficient and cost-effective for improving existing models, and that more data could lead to performance drops. We also found that adversarial models achieve lower scores on R2 data as opposed to the baseline, suggesting that R2 data can be more challenging.

Effect of adversarial attack. While adversarial attack provides an opportunity for improving counterspeech detection models by collecting robust and diverse data, such improvement is not consistent. Moreover, adding adversarial data provides limited model generalisability. One possible reason is the little data collected in each round compared to the amount of data used in training. Additionally, adversarial training is sensitive to distributional shift in data and perform poorly on out-of-domain evaluation sets (Zhang et al., 2019; Kaushik et al., 2021). Our adversarial context addresses abuse targets differently from the ones in training with the aim of expanding the coverage of target domains. While this intuitively aids generalisability, such distributional shift may come at the cost of robustness and require more samples to mitigate the tradeoff (Schmidt et al., 2018; Raghunathan et al., 2019).

		Non-	DynaCounter				
	PLF	adversarial	R1	R2	R3	R4	Mean
Base	0.61	0.70	0.38	0.52	0.47	0.46	0.46
M0	0.62	**0.76*****	0.37	0.48	0.48	0.45	0.45
M1	0.63	0.75	**0.80*****	0.47	**0.68****	0.61**	0.64
M2	0.61	0.73*	**0.80*****	0.48	**0.68****	0.45	0.60
M3	0.62	0.74	0.78***	0.48	0.48	0.45	0.55
M4	0.61	0.74	0.76***	0.48	0.64**	**0.68****	0.64

Table 5: Macro F1 of counterspeech classifiers tested on PLF, non-adversarial test sets and DynaCounter. Base denotes the baseline model trained only on PLF train data. Mean denotes the average scores over the test splits of four rounds in DynaCounter (R1-R4). The number of asterisk symbols indicates the p-value of McNemar's test (McNemar, 1947) - one: p<0.05, two: p<0.01, and three: p<0.001 - with respect to the Base model.

5 Counterspeech Generation

The last set of experiments is to assess generation performance across targets and datasets. We adopt the same data used in the adversarial attack on counterspeech detection (§4.1) for the task of counterspeech generation. We conducted the experiments with T5 (Raffel et al., 2020) and Flan-T5 (Chung et al., 2022). The training inputs are structured as [ABUSE] *response* : [COUNTERSPEECH].

		Human evaluation							Automatic evaluation			
	N	UWM↑	UAI↑	Usa.↑	Eng.↑	Spe.↑	Com.↑	Pol.↑	Tox.↓	RR↓	R-L↑	Ber.↑
Flan-T5	9483	**0.76**	0.31	**0.72**	2.54	2.67	2.44	2.66	0.19	12.52	0.12	0.18
Football.	799	0.55	0.10	0.34	1.48	1.38	1.59	1.55	0.24	6.49	**0.25**	**0.35**
Women	164	0.78	**0.48**	**0.91**	2.66	**3.13**	**2.85**	**3.07**	0.21	7.11	**0.15**	0.18
MD.	40	0.83	**0.38**	**0.90**	**2.84**	**3.15**	2.63	**2.98**	**0.10**	8.62	0.14	0.18
Migrants	186	**0.83**	0.23	0.68	2.83	2.77	2.46	2.78	0.11	6.59	0.13	**0.19**
Muslims	1545	0.73	0.43	0.79	**2.65**	2.54	**2.51**	2.60	0.18	6.25	0.12	0.15
Jews	100	**0.83**	0.33	0.78	2.75	3.03	2.59	2.66	0.20	**5.43**	0.10	0.09
LGBT+	1397	**0.80**	0.23	0.68	2.54	2.70	2.45	2.98	0.19	11.25	0.10	0.11
T5	9483	0.69	0.27	0.66	2.19	2.31	2.18	2.46	0.19	13.14	0.12	0.18
Football.	799	0.60	0.03	0.42	1.53	1.54	1.54	1.98	0.20	7.21	**0.26**	**0.36**
Women	164	0.63	**0.30**	0.66	**2.32**	**2.50**	2.06	**2.40**	**0.09**	8.07	**0.14**	0.17
MD.	40	**0.80**	**0.48**	**1.00**	2.79	3.10	3.05	**3.31**	0.11	5.02	**0.14**	0.20
Migrants	186	**0.85**	0.30	0.73	2.45	2.55	2.45	2.83	**0.10**	6.17	0.13	0.16
Muslims	1545	0.68	0.28	0.63	2.09	2.14	**2.18**	2.30	0.18	6.27	0.11	0.14
Jews	100	0.55	0.18	0.48	1.66	**1.94**	1.50	1.88	0.26	6.54	0.13	0.10
LGBT+	1397	0.73	**0.35**	**0.73**	**2.48**	2.39	**2.47**	2.50	0.27	**4.73**	0.11	0.14
K-alpha	-	-	-	0.43	0.11	0.24	0.16	0.14	-	-	-	-

Table 6: Counterspeech generation results. Abbreviation: N for the size of test set, Football. for footballers, and MD. for mental disability. The scores of toxicity (Tox.), Rouge-L (R-L.), and BERTScore (Ber.) are between 0 and 1.

5.1 Results for Counterspeech Generation

Automatic Metrics. A multifaceted assessment of models is considered. We measure lexical and semantic similarity between references and generation using ROUGE-L (Lin, 2004) and BERTScore (Zhang* et al., 2020), respectively. We also evaluate lexical diversity and toxicity of the generation using repetition rate (Cettolo et al., 2014) and Perspective (Google Jigsaw, 2022).

The results are reported in Table 6 with the breakdown of seven abuse targets. Regarding ROUGE-L and BERTScore, both models attain similar scores for each abuse target. Nevertheless, in terms of toxicity and repetition rate, the performances of the two models vary across abuse targets. The generation of Flan-T5 is less toxic than the one of T5 for Jews and LGBT+, while the inverse behaviour is observed for footballers and women. The generation of Flan-T5 is more repetitive than the T5 one for the domain of mental disability and LGBT+. Both models achieve the highest scores of ROUGE-L and BERTScore for footballers.

When comparing the performance across targets, none of the abuse targets obtains the best scores in all evaluation metrics. The toxicity in the Flan-T5 generation for certain targets (i.e. footballers, women, Muslims, Jews and LGBT+) is almost double the one for mental disability and migrants. Similar behaviour is observable for T5. In terms of repetition rate, for LGBT+ the Flan-T5 generation is twice the T5. One possible explanation could be that the training size varies across targets and models. Future work can investigate further to pinpoint the best condition for counterspeech generation.

Human Evaluation. Three annotators (1 female and 2 males) from the same civil organisation as in §4 assessed the quality of counterspeech generation. They were first tasked with evaluating counterspeech **Usability** (Usa.), i.e. how usable the response is with regard to the abusive comment on a scale of 0 to 2, with 0 being not usable at all, 1 being *Usable With Modification* (UWM) in which only minor post-editing is needed for being used (such as typo, tone and grammaticality), and 2 being *Usable As Is* (UAI), in which the generation can be used without modification, inspired by Tekiroğlu et al. (2022). If the usefulness is scored with 1 or 2, annotators were then asked to evaluate the generation with respect to four aspects on a scale of 1 to 5, 1 being the worst, 5 being the best: (1) **Engagingness** (Eng.), how engaging/interesting the counterspeech is, (2) **Specificity** (Spe.), how specific/tailored the counterspeech is as a response with respect to the given abusive comment in terms of topicality and style, (3) **Comprehensibility** (Com.), how understandable and natural the counterspeech is, and (4) **Politeness** (Pol.), how polite, respectful, or safe the generation is as a response. We deliberately adopt this two-step evaluation to lessen annotators' workload and focus on the generation quality of usable counterspeech. Each annotator was given 560 pairs,

consisting of 40 pairs randomly selected per target per model from the test set. The entire evaluation took around 30 hours.

Results are reported in Table 6. In addition to reporting the usability of generation, we calculate the ratio of UWM and the ratio of UAI. For both models, most generation is rated usable, either with modification (76% for Flan-T5 and 69% for T5) or as is (31% for Flan-T5 and 27% for T5). Regardless of abuse targets, annotators generally consider Flan-T5 generation as more usable than T5 counterpart. If a generation is considered usable, two models are comparable across all aspects of evaluation in terms of engagingness, specificity, comprehensibility and politeness. Annotators deemed Flan-T5 generation slightly more engaging, specific and comprehensible and T5 generation more polite. Flan-T5 can produce reasonable and, sometimes, outstanding counterspeech responses for almost all targets. In particular, over 90% of the generation of Women and Mental Disability is usable with a score above 2.6 on all evaluation aspects. On the contrary, T5 struggles with handling topics related to Jews: over 50% of the generation is regarded as not useful. Additionally, we find that both models were particularly bad at generating responses for footballers. We provide generation examples from two models in Appendix A.4.

We also calculate Krippendorff's alpha coefficient (K-alpha, Hayes and Krippendorff (2007)) to measure inter-rater agreement towards the model performance for each aspect. The low correlations are expected for three reasons. First, the idea of suitable responses to abuse is complex and intertwined closely with biases and prior knowledge. Second, within this context, quantifying the differences between responses (on a five-scale scale in our case) can be even more subjective, thus, making disagreement unavoidable. Poor agreement has been reported in both subjective and objective tasks (Cabitza et al., 2023). Third, not all generations receive the same amount of annotations as annotators are asked to only assess the engagingness, specificity, comprehensibility and politeness of the generation that is usable. Missing values in these aspects might account for the low agreement. In summary, our results speak to the need for further research on, for instance, adopting a perspectivist framework analysing collective opinion towards suitable replies to abuse (Cabitza et al., 2023).

6 Discussion

Based on the results presented in the previous sections, we discuss four high-level key observations.

Authentic counterspeech differs from synthetic ones in terms of style, language and nature. Online users often adopt informal language and conversational styles, reflecting the diverse linguistic expressions prevalent on the internet. Their counterspeech tends to be short, emotive, and direct and may incorporate internet slang to engage with a focused target audience. On the contrary, with the aim of catering to a wide audience, experts tend to employ a formal and educational tone, using evidence-based arguments to substantiate their claims. The language in expert counterspeech is usually structured and precise (see Appendix A.3). Despite these differences, it is unclear what responses the two types of counterspeech elicit and how they compare in fostering constructive dialogue and having a positive impact on the target audience. In future work, it will be imperative to consider the nuances and the dynamics of abuse and counterspeech while developing abuse intervention tools (e.g. user studies).

Adversarial attack does not necessarily result in better generalisation. While we attempted to collect adversarial examples for training better-performing and robust counterspeech classifiers, we did not observe striking improvements over the models trained on existing datasets, coherent with previous work on model-in-the-loop data collection (Raghunathan et al., 2019; Huang et al., 2020). We consistently discussed with annotators to refine their comprehension of the instructions and adjust tactics for deceiving the model over iterations. However, it is hard for annotators to trick the model after the second round.

Counterspeech generation tools hold promise for aiding in abuse mitigation efforts. In our experiments focused on counterspeech generation across various domains, the results indicate that language models yield approximately 69-76% of responses that are considered usable. This finding implies that automation tools have the capability to substantially alleviate moderators' workload by as much as 60-70% during abuse-countering tasks.

Some abuse targets are difficult to deal with. The experiments of counterspeech generation show

that both Flan-T5 and T5 fail at producing satisfactory responses for footballers. T5 struggles with handling topics related to Jews. Two findings can be drawn. First, authentic and expert counterspeech are characteristically different, and the generation for user-generated abuse is far from perfect. Second, generation in multi-domain learning is challenging. An alternative approach can be adopting dynamic models that allow for adaptation according to domains (Li et al., 2021); for instance, exploring adapting generation performance across domains and where such adaptation helps or hurts performance in source domains.

7 Conclusion

We study the adversarial robustness of counterspeech classifiers by collecting four iterations of datasets, DynaCounter, covering multiple domains. Our analysis indicates that the characteristics of authentic counterspeech are distinct from synthetic/expert alternatives, which points towards future opportunities for investigating the role of different types of counterspeech for abuse mitigation. Our extensive evaluation of counterspeech generation shows that while Flan-T5 and T5 can produce decent responses to abuse against various targets, certain pitfalls should be addressed.

Limitations

This work has several limitations. Firstly, while we employ a qualification test and provide detailed instructions for annotating replies to abuse on X/Twitter, annotators' perceptions of suitable responses to abuse may still vary. The labels of each entry are obtained through a majority vote or based on expert judgement. Secondly, the datasets used to train classification models are synthetically created and have not been validated on actual abuse mitigation in real-world scenarios. Such a research agenda is reserved for future work. Finally, while our adversarial attack is done over four rounds (same as Kiela et al. (2021)), adversarial data collection at scale may also be conducted to examine model performance in the long term. For instance, Wallace et al. (2022) conduct dynamic adversarial data collection for 20 rounds. Additionally, performing adversarial attacks is costly and requires domain expertise. Our DynaCounter consists of 1,911 entries to abusive comments.

Ethical Considerations

We carefully managed the potential societal and ethical considerations raised in this study. We also acknowledge that all experiments, data collection protocols and data release policies are approved by the internal ethics review board at our institution. In consideration of fairness and workers' well-being, all annotators were informed about task descriptions and guidelines as well as extensive content warnings. We also encouraged them to take a rest whenever they experienced any feelings of stress, discomfort or being overwhelmed. We compensated respectively crowdworkers and expert annotators with £12 and £16 per hour, which are above the hourly rate for the national minimum wage in the UK. No identifiable information about annotators is stored. DynaCounter dataset can be used for research purposes.

There are no easy and fast solutions for abuse mitigation. While the purpose of counterspeech generation is to resist the harms and polarization of extremist narratives (Braddock and Horgan, 2016; Stephens et al., 2021), our methods based on large language models pose potential misuses in which, for instance, malicious actors exploit the methods to spread false information that could instead elicit hatred. It is thus crucial to have open discussions on best practices in the deployment of mitigation systems among researchers, practitioners, and policymakers.

Acknowledgements

We express our gratitude to Brent Carey, Sean Lyons, Lynne Fennell and the staff at Netsafe for their continued work in assessing model performance and for insightful feedback on model-attacking strategies. We are grateful to Liam Burke-Moore for his help in setting up the experiment website, Angus R. Williams for the support in data collection, and Bertie Vidgen for valuable feedback in the early stage. This work was supported by the Ecosystem Leadership Award under the EPSRC Grant EPX03870X1 & The Alan Turing Institute.

References

Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

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.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: US Holocaust Memorial Museum*.

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 Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Kurt Braddock and John Horgan. 2016. Towards a guide for constructing and disseminating counternarratives to reduce support for terrorism. *Studies in Conflict & Terrorism*, 39(5):381–404.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Sarah L Carthy, Colm B Doody, Katie Cox, Denis O'Hora, and Kiran M Sarma. 2020. Counternarratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3):e1106.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. 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. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Online.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. 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.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Mekselina Doğanç and Ilia Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 1–12, Prague, Czechia. 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.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Inf. Process. Manage.*, 58(3).

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023.

High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Google Jigsaw. 2022. Perspective API. Accessed: 14 June 2023.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. 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, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 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.

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.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

William Huang, Haokun Liu, and Samuel R. Bowman. 2020. Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 82–87, Online. Association for Computational Linguistics.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. 2021. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. 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 6618–6633, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Yunsheng Li, Lu Yuan, Yinpeng Chen, Pei Wang, and Nuno Vasconcelos. 2021. Dynamic transfer for multi-source domain adaptation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10993–11002.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv:1812.02712*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496, Brussels, Belgium. Association for Computational Linguistics.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. 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 2388–2404, Online. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 255–264, New York, NY, USA. Association for Computing Machinery.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech.

Erin Marie Saltman and Jonathan Russell. 2014. White paper–the role of Prevent in countering online extremism. *Quilliam publication*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. 2018. Adversarially robust generalization requires more data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5019–5031, Red Hook, NY, USA. Curran Associates Inc.

Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

William Stephens, Stijn Sieckelinck, and Hans Boutellier. 2021. Preventing violent extremism: A review of the literature. *Studies in Conflict & Terrorism*, 44(4):346–361.

Scott R. Stroud and William Cox. 2018. *The Varieties of Feminist Counterspeech in the Misogynistic Online World*, pages 293–310. Springer International Publishing, Cham.

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.

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.

Bertie Vidgen, Yi-Ling Chung, Pica Johansson, Hannah Rose Kirk, Angus Williams, Scott A. Hale, Helen Zerlina Margetts, Paul Röttger, and Laila Sprejer. 2022. Tracking Abuse on Twitter Against Football Players in the 2021 – 22 Premier League Season. Available at SSRN 4403913.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Eric Wallace, Adina Williams, Robin Jia, and Douwe Kiela. 2022. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

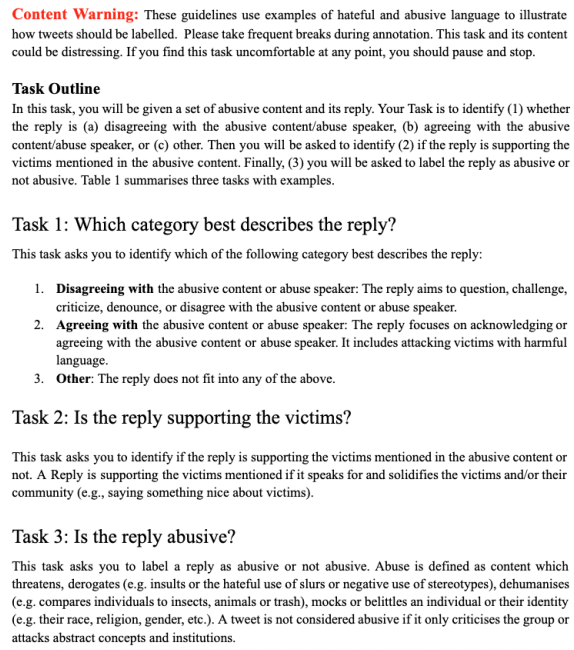
Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289, Singapore. Association for Computational Linguistics.

Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. 2019. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Annotation Instruction for PLF Dataset



Content Warning: These guidelines use examples of hateful and abusive language to illustrate how tweets should be labelled. Please take frequent breaks during annotation. This task and its content could be distressing. If you find this task uncomfortable at any point, you should pause and stop.

Task Outline

In this task, you will be given a set of abusive content and its reply. Your Task is to identify (1) whether the reply is (a) disagreeing with the abusive content/abuse speaker, (b) agreeing with the abusive content/abuse speaker, or (c) other. Then you will be asked to identify (2) if the reply is supporting the victims mentioned in the abusive content. Finally, (3) you will be asked to label the reply as abusive or not abusive. Table 1 summarises three tasks with examples.

Task 1: Which category best describes the reply?

This task asks you to identify which of the following category best describes the reply:

1. **Disagreeing with** the abusive content or abuse speaker: The reply aims to question, challenge, criticize, denounce, or disagree with the abusive content or abuse speaker.
2. **Agreeing with** the abusive content or abuse speaker: The reply focuses on acknowledging or agreeing with the abusive content or abuse speaker. It includes attacking victims with harmful language.
3. **Other**: The reply does not fit into any of the above.

Task 2: Is the reply supporting the victims?

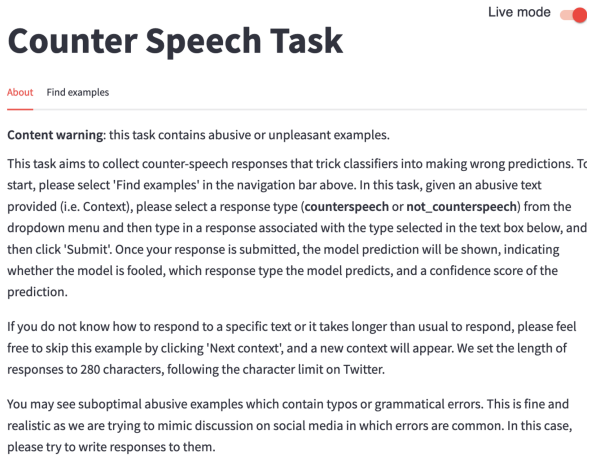
This task asks you to identify if the reply is supporting the victims mentioned in the abusive content or not. A Reply is supporting the victims mentioned if it speaks for and solidifies the victims and/or their community (e.g., saying something nice about victims).

Task 3: Is the reply abusive?

This task asks you to label a reply as abusive or not abusive. Abuse is defined as content which threatens, derogates (e.g. insults or the hateful use of slurs or negative use of stereotypes), dehumanises (e.g. compares individuals to insects, animals or trash), mocks or belittles an individual or their identity (e.g. their race, religion, gender, etc.). A tweet is not considered abusive if it only criticises the group or attacks abstract concepts and institutions.

Figure 2: A screenshot of instruction provided to annotators of PLF dataset.

A.2 Annotation Instruction for DynaCounter Dataset



Live mode

Counter Speech Task

About Find examples

Content warning: this task contains abusive or unpleasant examples.

This task aims to collect counter-speech responses that trick classifiers into making wrong predictions. To start, please select 'Find examples' in the navigation bar above. In this task, given an abusive text provided (i.e. Context), please select a response type (**counterspeech** or **not_counterspeech**) from the dropdown menu and then type in a response associated with the type selected in the text box below, and then click 'Submit'. Once your response is submitted, the model prediction will be shown, indicating whether the model is fooled, which response type the model predicts, and a confidence score of the prediction.

If you do not know how to respond to a specific text or it takes longer than usual to respond, please feel free to skip this example by clicking 'Next context', and a new context will appear. We set the length of responses to 280 characters, following the character limit on Twitter.

You may see suboptimal abusive examples which contain typos or grammatical errors. This is fine and realistic as we are trying to mimic discussion on social media in which errors are common. In this case, please try to write responses to them.

Figure 3: A screenshot of instruction provided to annotators of DynaCounter dataset.

A.3 Characteristics of Adversarial, Authentic and Expert Counterspeech

We conduct linguistic statistical analysis and provide specific examples for authentic counterspeech and synthetic counterparts. We compute the length of responses, percentage of negation in responses, and percentage of sentiment in authentic and synthetic responses. As shown in Table 7, we observe that the authentic responses (PLF) are shorter than synthetic ones (DynaCounter). We further include the breakdown of the length of each type of response (agree, disagree, other). In terms of negation operation, DynaCounter exhibits more negation cues than PLF. Regarding sentiment analysis, we find that DynaCounter employs a greater level of neutral and disapproval tones in responses as opposed to PLF. Conversely, the language used in PLF is found to be slightly more diverse in terms of sentiment distribution, such as anger, curiosity and amusement.

We further summarise key aspects of adversarial, authentic and expert counterspeech.

- While both authentic and adversarial counterspeech are short, the former can contain uncivil language and the latter can contain suboptimal speech for tricking the models (§4.2).
- Expert counterspeech addresses abuse targets at minority groups/identities with stereotypical associations, while authentic counterspeech uses supporting evidence or examples related to individual abuse targets (e.g. *[player] is great as he scores a goal.*)
- Authentic counterspeech addresses specific abusive claims using personal experience, event memory, endorsement, or specific event.

Category	DynaCounter	PLF
response_tokens$_{\mu}$	19.58	14.61
response_tokens$_{sd}$	7.6	13.15
response_agree_tokens$_{\mu}$	-	13.64
response_agree_tokens$_{sd}$	-	12.20
response_disagree_tokens$_{\mu}$	-	15.63
response_disagree_tokens$_{sd}$	-	13.63
response_other_tokens$_{\mu}$	-	8.24
response_other_tokens$_{sd}$	-	10.85
negation (%)	34.75	22.60
negation_agree	-	20.38
negation_disagree	-	25.14
negation_other	-	5.05
sentiment analysis (%)		
neutral	59.34	45.37
anger	2.51	7.52
disapproval	8.58	2.95
curiosity	2.25	6.19
amusement	0.78	4.05
sadness	0.94	2.69

Table 7: Linguistic and sentiment analysis of DynaCounter and PLF dataset.

A.4 Examples of Counterspeech Generation

HS: Feminism should be dead and give place to a male-dominated world.
Human: Feminism has always been in favour of women's rights. Now is time for us to open our hearts and minds and stop giving in to sexism.
T5: Feminism has always existed as a way to create harmony between men and women. That is why, for the most part, it is an effort that gives women, as well as people, a place to find happiness, fulfilment and respect.
Flan-T5: Feminism does not do anything to lead to a man-dominated world. We need to work towards better conditions of women.
HS: Children with Down Syndrome are a burden to society: they should be isolated, so that at least they do not hamper normal children education.
Human: And who would be 'normal'? I wonder how you would define 'disabled' people?
T5: Children with Down Syndrome can learn and grow to love the world around them, whether in school, at home, or in school.
Flan-T5: The education of a child with Down Syndrome is a direct result of the education system being corrupted by the government, not by the society itself. All children with Down Syndrome have the right to be educated properly.

Table 8: Samples of generations from T5 and Flan-T5.