Beyond Generic Responses: Target-Aware Strategies for Countering Hate Speech

Yen-Yu Chang[1], **Daryna Dementieva**[1,2], **Alexander Fraser**[1,2]
[1]Technical University of Munich (TUM)
[2]Munich Center for Machine Learning (MCML)
{yenyu.chang, daryna.dementieva, alexander.fraser}@tum.de

Abstract

Effective counter-narratives (CNs) are essential for combating online hate speech, yet generic responses often fail to address the specific needs of targeted groups. This paper proposes a Target-Aware CN generation framework that incorporates demographic-specific tokens into transformer-based models. Our approach enhances the contextual relevance by introducing target-group tokens into the model's vocabulary. To assess CN quality, we employ a multifaceted evaluation framework, including automatic metrics and LLM as Judges (JudgeLM). Evaluation with a wide range of language models demonstrates that target-group tokens markedly improve contextual relevance of generated CN, particularly in small and medium models, with measurable gains in validity as CN and contextual relevance. Even for large instruction-tuned models, such as LLaMA-3, incorporating target-specific information proves effective in enhancing contextual relevance of generated responses.

Warning: This paper contains offensive texts that are only used for combating online hate.

Keywords: Hate Speech/Offensive Language Detection, Evaluation Methodologies, Natural Language Generation

1. Introduction

The rise of hate speech (HS) in online spaces like Twitter and Reddit presents significant challenges, particularly as anonymity allows individuals to spread harmful messages widely (Aguilera-Carnerero and Azeez, 2016; Awan, 2016). This marginalizes specific demographic groups and risks spreading false information and inflammatory stereotypes, potentially leading to real-world violence (Bakalis, 2016; Atte Oksanen et al., 2020; Müller and Schwarz, 2020; Citron and Norton, 2011). Traditional censorship strategies often suppress freedom of expression (Álvarez-Benjumea and Winter, 2018), leading to calls for counter-narratives (CN)—responses designed to constructively challenge hateful content while fostering open dialogue (Garland et al., 2020; Chandrasekharan et al., 2017). However, the manual creation of CNs is labor-intensive and cannot scale to meet the volume of online hate speech (Mathew et al., 2019), especially given the misuse of generative language models (GLM) that can rapidly produce hate speech (Illia et al., 2022). GLMs offer a scalable alternative, but their potential to produce ineffective responses highlights the need for contextually relevant CNs, particularly for hate targeted at specific demographics. The example shown in Table 1 highlights the importance of tailored CNs. While the generic CN refutes the generalization, it does not directly address the ideology under attack. In contrast, the targeted CN explicitly challenges the statement, and reframes feminism as a movement for gender equality.

Equally critical is the challenge of evaluating CNs.

Hate Speech	Feminism is an anti-males ideology.
Offensive Response	Stupid piece of shit knows nothing about feminism.
Irrelevant Response	I think that is a stereotype about different races.
Generic Response	How can you say this about an entire group of people?
Relevant Response	Feminism is aimed at promoting the equality of the sexes.

Table 1: Example of different responses. Red: offensive contents; Orange: irrelevant/generic contents; Green: relevant contents.

While human evaluations are considered the gold standard, they are expensive, time-consuming, and prone to subjectivity. Automated evaluation frameworks for CNs have been explored (Zhu and Bhat, 2021; Hengle et al., 2024), but these often fail to account for critical aspects such as contextual relevance and their validity as CN. This paper addresses both challenges by making the following key contributions:

- A Transformer-based pipeline for **generating contextually relevant CNs** to demographic-specific hate speech. Our approach improves response relevance by incorporating target-group tokens, resulting in contextually more relevant responses.
- A **multi-faceted evaluation framework** that complements human evaluation. This framework includes: (1) An automatic evaluation

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1–18
11–16 May 2026.

pipeline that assesses language quality, contextual relevance, toxicity, validity as counter-narrative, and response diversity. (2) JudgeLM evaluation, a novel system that simulates human scoring and ranking of responses while providing detailed reasoning for its assessments (Zhu et al., 2023).

2. Related Work

2.1. Benchmark Datasets for Counter-Narrative Generation

The foundation of CN generation lies in high-quality datasets. Qian et al. (2019) introduced the first benchmark datasets with fully labeled conversational segments and human-written intervention responses. These include a Reddit dataset with 5,020 conversation segments (10,243 CN against 5,257 HS) and a Gab dataset with 11,825 segments (31,487 CN against 14,614 HS).

The Counter Narratives through Nichesourcing (CONAN) dataset by Chung et al. (2019) focuses on Islamophobia, containing 4,078 expert-crafted HS-CN pairs in English, French, and Italian. This dataset was later expanded by Chung et al. (2021) to include knowledge-grounded CNs. This extension adds 195 expert-crafted HS-CN pairs with background knowledge annotations to support factual responses. Building on CONAN, Fanton et al. (2021) proposed a human-in-the-loop data collection strategy to create the Multi-target CONAN dataset. This dataset emphasizes demographic variety, with 5,003 HS-CN pairs targeting eight distinct demographics. Partially generated by LMs, these pairs underwent human filtering and post-editing to ensure high-quality, human-approved responses. Bonaldi et al. (2022) further extended this line of work by introducing DIALOCONAN, a multi-turn dialogues dataset, comprising 3,059 conversation segments of 4, 6, or 8 turns. Each dialogue simulates interactions between online haters and NGO operators across seven target demographics.

The CrowdCounter dataset by Saha et al. (2024) introduces a benchmark for type-specific CN generation. It consists of 3,425 HS-CN pairs annotated with six distinct response types: empathy, humor, questioning, warning, shaming, and contradiction.

Hengle et al. (2024) curated the IntentCONANv2, consisting of 13,952 CNs addressing 3,488 HS. The dataset covers four distinct CN intents: positive, informative, questioning, and denouncing, and includes CNs targeting ten different groups commonly subjected to hate speech, such as Muslims, LGBTQ+, refugees, and others.

2.2. Methods for Counter-Narrative Generation

Several methodologies have been proposed for generating CNs. Zhu and Bhat (2021) developed the GPS pipeline to produce diverse and contextually relevant responses. This pipeline was trained on datasets from Qian et al. (2019) and the English portion of CONAN. Saha et al. (2022) introduced CounterGeDI, a controllable generation approach that uses generative discriminators (GeDIs, Krause et al., 2020) to guide token probabilities during inference. This approach enables the generation of CNs with specific attributes such as tone or intent.

Hengle et al. (2024) presented CoARL, a novel three-phase framework for intent-conditioned, non-toxic CN generation. Phase 1 leverages multi-task explanation generation to capture pragmatic facets of hate speech (e.g., identifying target-group, implied meaning etc.), Phase 2 employs lightweight LoRA adapters to verbalize and enforce specific CN intents, and Phase 3 applies RL with AI feedback (RLAIF) to optimize a composite reward balancing opposition strength, argumentative quality, and non-toxicity. CoARL achieved state-of-the-art performance on both automated metrics and human judgments.

Zhang et al. (2022) presented TRRGen, a Transformer-based model for app review response generation. By integrating contextual features such as app categories, user ratings, and review text, TRRGen demonstrated the potential of Transformer architectures in automating response generation tasks, producing high-quality, contextually relevant responses. While this work focuses on app reviews, its approach is relevant for developing CN generation methods.

2.3. Automatic Evaluation of Counter-Narratives

While human evaluations are often considered the gold standard for evaluating the quality of CNs due to their nuanced understanding, they are resource-intensive and time-consuming. Consequently, automatic evaluation metrics have become essential for providing rapid and cost-effective assessments of CN generation models.

Traditional evaluation methods rely on lexical overlap metrics such as BLEU, ROUGE, METEOR, and BERTScore, commonly used in CN evaluation (Qian et al., 2019; Tekiroglu et al., 2020; Zhu and Bhat, 2021; Saha et al., 2022; Doğanç and Markov, 2023; Lee et al., 2024; Saha and Srihari, 2024). Some studies have extended this with linguistic quality measures like GRUEN (Zhu and Bhat, 2021; Hong et al., 2024) and fluency/perplexity (PPL) (Saha and Srihari, 2024). However, these metrics often fall short in capturing the nuanced qualities

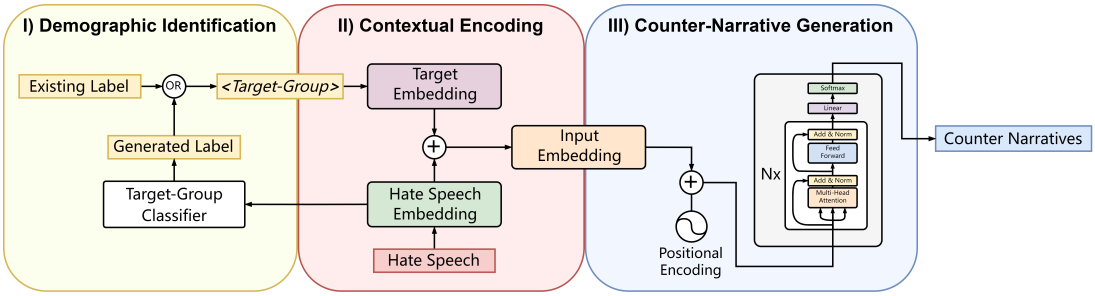

Figure 1: Illustration of the proposed transformer-based counter-speech generation framework.

of effective CNs, such as relevance, validity, and the ability to counteract hate speech.

To address these limitations, recent research has explored specialized classifiers to assess CN effectiveness. Hong et al. (2024) utilized conversation incivility level and hater reentry classifiers to measure the impact of CNs on discourse. Jiang et al. (2023) evaluated CN success through toxicity reduction (Perspective API), informativeness, and persuasiveness (GPTScore with ChatGPT).

More recently, studies have leveraged large language model (LLM) based evaluation frameworks, where LLMs serve as judges to assess CN quality beyond surface-level metrics. Jones et al. (2024) proposed a multi-aspect evaluation framework that prompts LLMs to assess CNs across dimensions like specificity, opposition, relatedness, toxicity, and fluency. Zubiaga et al. (2024) developed a ranking method using LLMs to perform pairwise comparisons of CNs, achieving a high correlation with human preferences. Halim et al. (2023) presented a novel evaluation metric called the PD-Score. This metric is designed to assess CN quality across multiple dimensions such as relevance, argument quality, and impact, using an advanced debating system framework to better mimic human judgment, providing a more nuanced and context-aware measure of CN performance.

3. Methodology

3.1. Counter-Narrative Generation with Target-Group Tokens

To address the limitations of generic CN generation, we propose a transformer-based framework that explicitly incorporates demographic-specific signals to produce targeted CNs. Drawing on the Multi-target CONAN dataset (Fanton et al., 2021), which defines eight distinct target-groups (*Migrants*, *People of Color*, *LGBT+*, *Muslims*, *Women*, *Jews*, *Disabled*, and *Other*), we utilize a tokenization strategy to encode target-group information. Our approach extends the model's vocabulary with a dedicated `<Target-Group>` token for each demographic category.

Figure 1 illustrates the complete pipeline. The framework operates through three key stages: (I) Demographic identification, where the target-group is either extracted from existing annotations or predicted via a classifier; (II) Contextual encoding, where the corresponding `<Target-Group>` token is prepended to the hate speech input, creating a composite embedding that jointly represents both the toxic content and its demographic context; (III) CN generation, where the Transformer architecture leverages the enriched embedding to produce tailored responses. This architecture enables the model to learn compact, discriminative representations of demographic-specific linguistic patterns through the specialized tokens, while maintaining the generative capacity of standard language models.

3.2. Datasets

We utilized three datasets for training and testing purposes, summarized in Table 2.

Generator Training Set is used to fine-tune CN generators. It consists of English instances from the CONAN (Chung et al., 2019), Multi-target CONAN (Fanton et al., 2021), and DIALOCONAN (Bonaldi et al., 2022) datasets (10,990 HS-CN pairs, 60-20-20 Train-Val-Test split).

Dataset	Source	Size
Generator Train-Set	CONAN, MT-CONAN, DIALOCONAN	10,990
Classifier Train-Set	CONAN, KG-CONAN, MT-CONAN, DIALOCONAN, CrowdCounter	25,642
Sexism Test Set	EDOS	4853

Table 2: Summary of datasets used in this study. An instance is either a HS-CN pair (Training Sets) or a single HS (Sexism Test Set).

Classifier Training Set is used to fine-tune target-group and CN classifiers that evaluate the relevance and validity of generated CNs. It includes English instances from the CrowdCounter (Saha et al., 2024), CONAN, Multi-target CONAN, DIALOCONAN, Knowledge-Grounded CONAN (Chung et al., 2021), and EDOS (Kirk et al., 2023) datasets with their corresponding target-group labels (25,642 HS-CN pairs, 80-10-10 Train-Val-Test split).

Sexism Test Set is used to test model performances on real-world sexist HS. It consists of instances from the EDOS dataset (Kirk et al., 2023) that are labeled as sexist (4,853 HS instances).

3.3. Baseline Models

We trained the GPS and CounterGeDI pipeline as described in the original papers (Zhu and Bhat, 2021; Saha et al., 2022). We fine-tune **BART-L** (Lewis et al., 2019), **Flan-T5-L and -XL** (Chung et al., 2022a), **GPT-2-M and -XL** (Radford et al., 2019), and **Llama-3.2-1B-Instruct** (Grattafiori et al., 2024), with and without target-group tokens, to compare the effectiveness of target-group tokens in generating contextually relevant CNs. Baseline models utilized are summarized in Table 3.

3.4. Automatic Evaluation

To systematically evaluate the quality of generated CNs, we employ automatic metrics across five key attributes: **Language Quality**, **Toxicity**, **Validity as Counter-Narrative**, **Diversity**, and **Relevance**. Each attribute is selected based on its critical role in determining the effectiveness of CNs. For attributes with complex characteristics, multiple metrics are utilized to ensure robust assessment. To facilitate comparison, all scores are normalized between [0, 1], where higher values indicate better performance.

Model Name	Configuration
GPS	VAE + 14 candidate selection methods
CounterGeDi	DialoGPT + ***toxic*** & ***polite*** GeDIs
BART	L (400M)
Flan-T5	L (770M) XL (3B)
GPT-2	M (355M) XL (1.5B)
Llama-3	3.2-1B-Instruct

Table 3: Overview of baseline models and pipelines.

Language Quality measures the grammatical correctness of CNs. Correct sentences are generally a good starting point for constructing high-quality CNs. We use the CoLA (Corpus of Linguistic Acceptability) classifier (Morris et al., 2020) to assess grammatical acceptability.

Toxicity ensures that generated responses do not contain harmful or offensive language, since CNs should de-escalate rather than perpetuate hostility. We employ two classifiers: a general toxicity classifier and a hate-offensive classifier (Mathew et al., 2020).

Validity as Counter-Narratives determines whether a generated response genuinely functions as a CN. This is crucial, as responses that fail to counteract HS cannot fulfill their intended purpose. To evaluate this, we fine-tune a classifier that differentiates between valid CNs and other text types (e.g., HS, off-topic responses). The classifier achieved an overall accuracy of 0.93, a macro-averaged F1-score of 0.93, class-wise precision and recall between 0.88 and 0.99, and an ROC-AUC score of 0.985.

Diversity is essential to prevent repetitive CNs, which may reduce their effectiveness in real-world conversations. We measure diversity using the Repetition Rate (RR) at $n = 4$, following prior work(Cettolo et al., 2014).

Relevance assesses whether the generated CNs are semantically aligned with human-written CNs and contextually relevant to the input HS. We employ three techniques for robust assessment of contextual relevance:

- **Target-Group Classifier:** We fine-tune two target-group classifiers (Antypas et al., 2022) based on tweet-topic classifiers to determine if a CN pertains to the correct target-group. The CardiffNLP classifier achieves micro-averaged F1 of 0.98 and the NLP-LTU classifier attains a micro-averaged F1 of 0.96. Both classifiers maintain precision and recall consistently above 0.90 across all eight groups.
- **Semantic Relevance to Human-Written CN**: We compute cosine similarity between the generated CN and the human-written reference with Universal Sentence Encoders (USE). Higher similarity suggests that the model-generated CN aligns with human reference.

HS	The phrase 'black lives matter' suggests racial superiority. It excludes the importance of anyone elses life. This is racism defined. We cannot bow down to these racists.
Human	Saying one thing matters does not insinuate that other things do not matter. You can care about and value more than one thing at once.
Generated	It's essential to understand the context and intentions behind this phrase. "Black lives matter" is a social movement that originated in the United States, aiming to bring attention to systemic racism and police brutality against African Americans. The phrase is not about promoting racial superiority but rather about highlighting the disproportionate impact of racism and injustice on Black communities. The movement seeks to acknowledge the historical and ongoing struggles faced by Black people, including disparities in education, employment, healthcare, and the criminal justice system.

Table 4: An example of generated response that scores high in contextual relevance to input but low in semantic relevance to reference. Directly relevant key words are highlighted in green and blue. The HS-CN pair is from the DIALOCONAN dataset.

- **Contextual Relevance to HS**: We measure how well the CN relates to the original HS using Question-Answering (QA) Sentence Encoders. This metric ensures that responses are not off-topic.

An example in Table 4 shows a generated CN that achieves a high contextual relevance score, but scores low in semantic similarity. This generated CN is closely tailored to the HS, whereas the human reference is more generic. In scenarios where generated CNs closely mirror the generic human-written responses, the contextual relevance score tends to drop while the semantic similarity score increases. By combining these two metrics, we ensure that a CN receives a low relevance score if it's neither contextually relevant nor semantically aligned with the reference.

For attributes evaluated using multiple metrics, a final score is computed as the average of all individual metric scores:

$$S_i = \frac{\sum_j^n s_j}{n} \tag{1}$$

where S_i represents the overall score for the i-th attribute, s_j denotes the score of the j-th metric, and n denotes the total number of metrics used for that attribute.

$$s_j = \begin{cases} 1, \text{ if } p_j > 0.5 \\ 0, \text{ else} \end{cases} \tag{2}$$

For metrics involving classifiers, we employ a hard scoring scheme with a threshold of 0.5. p_j denotes the probability assigned by the classifier of the j-th metric. To assess overall model performance, we introduce a scaled average score (Average), defined as the mean of all evaluation metrics, weighted by the validity score (VAL):

$$\text{Average} = S_{\text{VAL}} \frac{\sum_i^n S_i}{n} \tag{3}$$

where S_{VAL} is the validity score, S_i represents the score of the i-th attribute, and n denotes the number of attributes ($n = 4$ in this study). By incorporating S_{VAL} as a scaling factor, we ensure that responses receive low scores if they fail to function as valid CNs, even if they perform well on other attributes.

3.5. Evaluation with JudgeLM

We employ JudgeLM with a fine-tuned 7B Llama-2-based model to score generated CNs on a 1–10 scale. JudgeLM not only provides a numerical score but also offers reasoning behind its judgments, thereby enhancing the explainability of the evaluation. During preliminary experiments, we observed that JudgeLM struggles with ranking a large number of response candidates simultaneously. To address this issue, we adopt a group-wise ranking approach, wherein CNs generated by models of the same family (e.g., GPT2-Medium Base, Fine-Tuned, and Target-Aware variants) are compared within the same group. When available, human-written CNs serve as references within these groups. Baseline models, such as CounterGeDI and GPS, are similarly grouped for a fair comparison.

3.6. Human Evaluation

To complement our automatic evaluation pipeline, we conducted a human evaluation to assess the quality of generated CNs based on key qualitative attributes. Given the limitations of automatic metrics in capturing nuanced aspects such as contextual appropriateness and effectiveness, human judgment remains a crucial component in evaluating CNs.

3.6.1. Design and Criteria

For this study, we randomly selected 10 HS samples from the **Sexism Test Set** and paired them

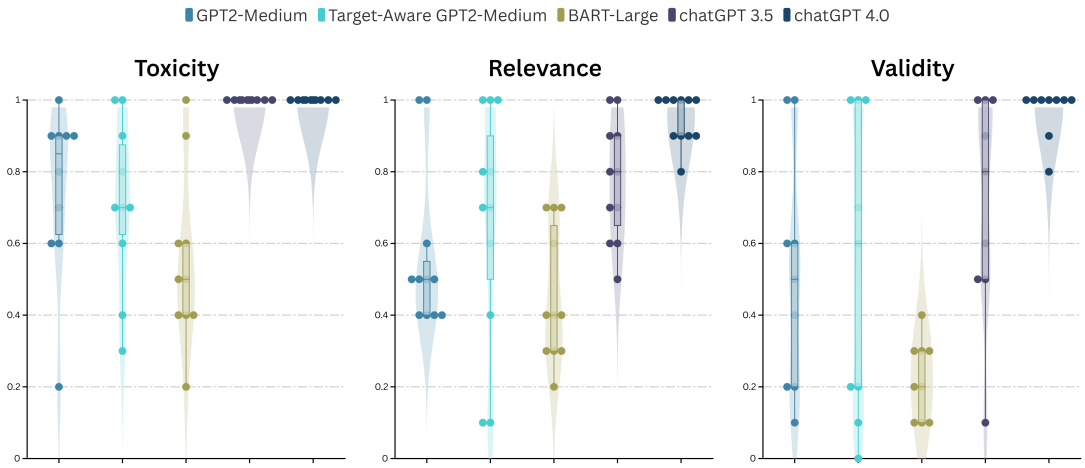

Figure 2: Overview of the human evaluation results for selected models

with CNs generated by five models: **ChatGPT-4.0**, **ChatGPT-3.5**, **GPT-2 Medium**, **BART-Large**, and the **Target-Aware GPT-2 Medium**. This resulted in a total of 50 HS-CN pairs. Each CN was evaluated based on **Toxicity**, **Relevance** and **Validity as Counter-Narrative** using a binary scale (Yes = 1, No = 0) to ensure consistency with automatic evaluation metrics.

3.6.2. Procedure

We recruited 10 annotators with engineering backgrounds: 2 PhD students (1 male, 1 female), 6 Master's students (4 male, 2 female), and 2 Bachelor's students (2 female). This ensured a gender-balanced evaluation pool, which is especially important for assessing CNs on sexist content. To ensure consistency and reduce ambiguity, all participants were provided with a detailed instruction sheet including explanation of terminologies and example HS-CN pairs of valid and invalid CNs. The evaluation was conducted using an online form, where annotators independently rated each HS-CN pair across the three attributes.

4. Results

4.1. Automatic Evaluation

Table 5 and 6 present the complete results of the automatic evaluation. We observe key trends across different model architectures.

On the Generator Training Set, Target-Aware GPT2-M achieves the best overall performance among the mid-sized models with an average score of 0.775, slightly higher than its fine-tuned variant (0.773) and substantially higher than the base model (0.615). It maintains high Validity (VAL = 0.999) and Toxicity (TOX = 0.990) scores, with improved Relevance (REL = 0.846) and JudgeLM scoring (0.742), indicating higher contextual quality and human-likeness. A similar trend is observed for Target-Aware GPT2-XL, which achieves the strongest performance among GPT2-based models with the highest overall average (0.792) and top JudgeLM score (0.773) across all models. Among the T5-based models, Target-Aware Flan-T5-XL improves the average score to 0.692 from 0.688 (fine-tuned), while preserving high Validity (0.855) and Language Quality (CoLA = 0.995).

On the out-of-distribution Sexism Test Set, Target-Aware BART-L achieves the best within-group average (0.752), outperforming both the fine-tuned (0.656) and base (0.113) variants. The Target-Aware GPT2-M model continues to perform well with the highest overall average (0.737) in its group, improving upon the fine-tuned (0.728) and base (0.356) models. Similarly, Target-Aware GPT2-XL achieves a competitive average (0.744), with the best Relevance (REL = 0.701) and JudgeLM score (0.642) within its group.

Our evaluation demonstrates that integrating demographic target information enhances CN contextual appropriateness while maintaining high linguistic quality, with Target-Aware variants consistently yield higher Relevance (REL) and Validity (VAL) scores for small- to medium-sized models.

Despite the improvements, a notable tradeoff emerges between **VAL** and **RR**. This suggests that while target-group tokens improve specificity and contextual relevance, they may also encourage repetitive patterns, limiting the variability of generated CNs.

For larger instruction-tuned models, the benefits of target-group tokens are less pronounced. While the **Target-Aware LLaMA-3.2-1B-Instruct** maintained comparable performance and showed improvements in relevance over its base model, the base model demonstrates strong performance relative to other models. We observed that the fine-

		Average	VAL	TOX	CoLA	RR	REL	JudgeLM	Len±SD
Human References		0.889	0.997	0.972	0.937	0.861	0.799	0.714	134±64
GPS	USE-QA-MAP	0.707	0.963	0.911	0.793	0.500	**0.735**	-	130±59
	USE-QA-SIM	**0.725**	0.968	0.937	0.812	0.536	0.714	0.470	127±52
CounterGeDI	DialoGPT	0.482	0.736	0.898	0.861	0.309	0.552	-	1587±168
	DialoGPT + p & nt	**0.520**	**0.741**	**0.903**	**0.919**	**0.418**	**0.566**	0.273	538±84
BART-L	Base	0.033	0.040	0.751	0.918	**0.659**	**0.907**	0.122	87±53
	Fine-Tuned	**0.783**	**0.985**	**0.975**	0.957	0.420	0.830	**0.569**	106±44
	Target-Aware (ours)	0.776	0.967	0.953	**0.975**	0.442	0.842	0.486	99±36
Flan-T5-L	Base	0.428	0.534	0.899	0.969	**0.687**	0.655	0.130	30±35
	Fine-Tuned	**0.716**	**0.896**	0.925	**0.997**	0.422	**0.853**	**0.602**	77±22
	Target-Aware (ours)	0.709	0.889	**0.926**	0.991	0.422	0.849	0.522	78±22
Flan-T5-XL	Base	0.163	0.198	0.817	0.957	**0.684**	0.823	0.170	62±46
	Fine-Tuned	0.688	0.849	**0.922**	0.992	0.470	**0.859**	**0.609**	76±23
	Target-Aware (ours)	**0.692**	**0.855**	0.917	**0.995**	0.466	**0.859**	0.520	76±23
GPT2-M	Base	0.615	0.782	0.926	0.993	**0.510**	0.714	0.336	418±143
	Fine-Tuned	0.773	**1.000**	**0.992**	0.995	0.266	0.837	0.674	553±51
	Target-Aware (ours)	**0.775**	0.999	0.990	**0.997**	0.270	**0.846**	**0.742**	564±42
GPT2-XL	Base	0.613	0.777	0.917	0.995	**0.485**	0.762	0.368	446±106
	Fine-Tuned	0.789	**1.000**	**0.989**	0.998	0.313	0.856	0.735	537±56
	Target-Aware (ours)	**0.792**	**1.000**	0.983	**0.999**	0.329	**0.859**	**0.773**	543±48
Llama-3.2 1B-Instruct	Base	**0.818**	**1.000**	**0.997**	**0.999**	**0.483**	0.794	**0.769**	534±183
	Fine-Tuned	0.777	0.964	0.957	0.996	0.444	0.826	0.590	78±32
	Target-Aware (ours)	0.814	0.998	0.986	0.996	0.448	**0.833**	0.729	107±40

Table 5: Automatic evaluation leaderboard on the Generator Training Set test split. **Bold** denote the best results within the group, **underlined**–the best for the attribute. *Len*: average length of CN in words. SD: standard deviation.

		Average	VAL	TOX	CoLA	RR	REL	JudgeLM	Len±SD
GPS	USE-QA-SIM	0.587	0.930	0.944	0.733	0.434	0.415	0.399	119±53
CounterGeDI	DialoGPT + p & nt	0.615	0.880	0.914	0.528	0.292	0.402	0.145	520±105
BART-L	Base	0.113	0.135	0.705	0.781	**0.987**	**0.879**	0.126	119±74
	Fine-Tuned	0.656	0.909	**0.961**	0.927	0.479	0.524	0.241	102±46
	Target-Aware (ours)	**0.752**	**0.953**	0.933	**0.971**	0.612	0.640	**0.347**	112±42
Flan-T5-L	Base	0.458	0.571	**0.887**	0.929	**0.934**	0.462	0.135	31±42
	Fine-Tuned	**0.472**	**0.606**	0.850	**0.963**	0.602	0.702	0.274	74±27
	Target-Aware (ours)	0.463	0.595	0.844	0.956	0.605	**0.710**	**0.360**	75±27
Flan-T5-XL	Base	0.275	0.333	0.816	0.885	0.951	0.657	0.150	57±54
	Fine-Tuned	0.365	0.454	0.824	0.943	0.718	0.731	0.262	70±33
	Target-Aware (ours)	**0.384**	**0.479**	**0.835**	**0.957**	0.677	**0.737**	**0.347**	70±29
GPT2-M	Base	0.356	0.482	0.848	0.973	**0.508**	0.626	0.216	391±131
	Fine-Tuned	0.728	**0.995**	0.957	0.989	0.324	0.656	0.481	529±49
	Target-Aware (ours)	**0.737**	0.993	**0.964**	**0.993**	0.351	**0.661**	**0.562**	533±47
GPT2-XL	Base	0.326	0.435	0.828	0.977	**0.525**	0.672	0.243	389±123
	Fine-Tuned	**0.750**	**0.990**	**0.967**	**0.995**	0.372	0.696	0.576	510±51
	Target-Aware (ours)	0.744	0.982	0.958	**0.995**	0.375	**0.701**	**0.642**	511±47
Llama-3.2 1B-Instruct	Base	**0.790**	**0.982**	**0.991**	**0.997**	**0.642**	0.587	**0.773**	527±154
	Fine-Tuned	0.649	0.835	0.965	0.994	0.538	**0.615**	0.349	88±35
	Target-Aware (ours)	0.720	0.931	0.971	0.992	0.523	0.605	0.736	124±50

Table 6: Automatic evaluation leaderboard on the Sexism Test Set. **Bold** denote the best results within the group, **underlined**–the best for the attribute. *Len*: average length of CN in words. SD: standard deviation.

		Average	VAL	TOX	CoLA	RR	REL
Generator Training Set test split	Base	0.791	**0.999**	0.983	**0.989**	0.377	0.816
	Target-Aware (ours)	**0.865**	0.995	**0.985**	0.982	**0.663**	**0.847**
Sexism Test Set	Base	0.714	**0.996**	**0.987**	**0.990**	0.367	0.524
	Target-Aware (ours)	**0.785**	0.959	0.963	0.985	**0.628**	**0.697**

Table 7: Automatic evaluation results from **Dolphin-2.9-Llama3-8B** for the ablation study. **Bold** denotes the best results.

tuned variants struggle with unseen themes. When encountering unfamiliar topics, they tend to produce generic CNs. However, in cases where the base model's safety policy prevents it from responding, fine-tuned versions can generate effective CNs.

4.2. Ablation Study

Table 7 presents automatic evaluation results from **Dolphin-2.9-Llama3-8B**, an uncensored instruction-tuned model. This ablation examines how removing censorship and safety filtering—which often suppress socially charged or demographically specific content—affects counter-narrative (CN) generation.

Across both test sets, the Target-Aware variant substantially outperforms the Base model in both overall average score and individual quality dimensions, with notable gains in Relevance (REL: 0.816 → 0.847, 0.524 → 0.697) and Repetition Rate (RR: 0.377 → 0.663, 0.367 → 0.628), showing that a censorship-free model can produce richer and more contextually adaptive CNs without resorting to generic repetition. This ablation suggests that lifting censorship and safety constraints—while maintaining responsible toxicity control—could significantly enhance CN generation quality.

4.3. Human Evaluation

Table 8 summarizes human evaluation results. ChatGPT-4.0 outperforms all baselines, achieving near-perfect scores across all attributes, followed by ChatGPT-3.5. Among smaller models, Target-Aware GPT-2 M slightly outperformed its standard version (Avg: 0.637 vs. 0.610), demonstrating the effectiveness of incorporating target-group information. BART-Large scored the lowest overall (Avg: 0.417), struggling with both relevance and validity.

Model Name	VAL	TOX	REL	Avg.
ChatGPT-4.0	**0.970**	**1.000**	**0.940**	**0.970**
ChatGPT-3.5	0.740	1.000	0.750	0.830
GPT2-M (ours)	0.580	0.710	0.620	0.637
GPT2-M (base)	0.560	0.750	0.520	0.610
BART-L (base)	0.270	0.550	0.430	0.417

Table 8: Human evaluation leaderboard, sorted based on Average Score. **Bold** denotes the best result within the attribute. *: Target-Aware GPT2-Medium (ours). ChatGPT models are tested on version (May 24, 2023).

The annotator agreement analysis, presented in Figure 2, highlights variability in ratings, particularly for smaller models, suggesting subjective interpretation of CN quality. While agreement remains high for ChatGPT-4.0 and ChatGPT-3.5, lower consensus for models like GPT-2 Medium and BART-Large indicates that human perception of CN quality can vary. This highlights the need for further refinement in evaluation criteria to improve consistency in assessing CNs.

5. Conclusion

The widespread proliferation of online hate speech, accelerated by digital communication, presents a critical challenge for online discourse. In response, this paper introduces a Target-Aware CN generation framework that integrates demographic-specific tokens into transformer-based language models, improving the contextual relevance of generated CNs, particularly effective in smaller and medium-sized models.

Our ablation study with Dolphin-2.9-Llama3-8B further revealed that lifting censorship and safety constraints—while preserving non-toxicity—could enable models to produce more contextually grounded and demographically sensitive responses.

We also present a multi-faceted automatic evaluation framework, combining automatic metrics and JudgeLM, providing an efficient, scalable alternative for CN evaluation.

These findings underscore the importance of demographic awareness in CN generation and highlight the potential of scalable, context-aware interventions against online hate speech.

Reproducibility: Our code is available at: Official Implementation

6. Acknowledgements

The work was supported by the German Research Foundation (DFG; grant FR 2829/7-1). Daryna Dementieva's work was additionally supported by Friedrich Schiedel TUM Think Tank Fellowship.

Limitations

While our study demonstrates the effectiveness of target-group-aware CN generation, several challenges remain open for further exploration.

Special Tokens vs. Instruction-Based Conditioning: While incorporating target-group tokens improved response relevance for smaller transformer-based models, we observed performance degradation in some larger instruction-tuned models (e.g., Llama 3). The reliance on explicit tokens for predefined groups might also limit the model's generalizability to groups not anticipated during training. A promising alternative is instruction-based conditioning, where demographic information is included within the prompt. Prior research (Ashida and Komachi, 2022) suggests that models fine-tuned on instruction-following tasks are more effective when contextual details are provided in natural language rather than token-based modifications.

Evaluation Challenges: Assessing CNs remains difficult due to the subjectivity of human evaluation. While we employed JudgeLM and automatic metrics as proxies for human evaluation, further validation is needed to confirm their alignment with human judgment on a larger scale. We acknowledge that LLM-based evaluators often favor longer responses (Bonaldi et al., 2025). To account for this influence, we included sequence length as an indicator in the evaluation.

Revisiting the Gold Standard for CNs: Via grouping LLaMA-3.2-1B and Human References together, JudgeLM results indicate that LM-generated CNs outperform human-written references (**LLaMA-3.2-1B (0.813) human references (0.562)**). This challenges the assumption that human-authored responses should remain the gold standard for CN evaluation. If lower-quality data is used to fine-tune stronger models, it could hinder progress rather than improve it. On the other hand, future research should explore the potential of synthetic data from high-performing LMs to curate high-quality CN datasets, ensuring diversity and effectiveness while maintaining human oversight in validation and refinement.

Rethinking Relevance Evaluation: Currently, relevance evaluation relies on topic classification to determine whether a CN is contextually appropriate for the given hate speech. However, topic classification is often too coarse-grained: A response may match the general topic of the hate speech but fail to directly address its specific claims or implications.

To improve relevance evaluation, we propose constructing larger and more diverse datasets comprising HS-CN pairs explicitly labeled as relevant or irrelevant. Training a dedicated relevance classifier on such datasets would allow us to capture a broader and more nuanced understanding of relevance beyond simple topic alignment.

Ethical Statement

The use of language models (LMs) for CN generation presents several ethical challenges. One concern is content hallucination, where LMs generate plausible but inaccurate responses (Ji et al., 2023). This can lead to the spread of misinformation, reinforcing harmful ideologies (Zellers et al., 2020). Mitigating this through knowledge-grounded generation techniques (Chung et al., 2021) is essential, but the problem remains unresolved. Another ethical risk is the potential malicious misuse of these models. Although our research aims to combat hate speech, the same models could be exploited to generate harmful content. While countermeasures like RLHF and detoxification methods have been developed (Saha et al., 2022), the risk of misuse cannot be entirely eliminated. Human post-editing and stricter censorship of harmful phrases may reduce this risk but not prevent it entirely (Chung et al., 2022b).

7. Bibliographical References

Carmen Aguilera-Carnerero and Abdul-Halik Azeez. 2016. 'islamonausea, not islamophobia': The many faces of cyber hate speech. *Journal of Arab Muslim Media Research*, 9:21–40.

Amalia Álvarez-Benjumea and Fabian Winter. 2018. Normative change and culture of hate: An experiment in online environments. *European Sociological Review*, 34(3):223–237.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atte Oksanen, Markus Kaakinen, Jaana Minkkinen, Pekka Räsänen, Bernard Enjolras, and Kari Steen-Johnsen. 2020. Perceived societal fear and cyberhate after the november 2015 paris terrorist attacks. *Terrorism and Political Violence*, 32(5):1047–1066.

Imran Awan. 2016. Islamophobia on social media: A qualitative analysis of the facebook's walls of hate. *International Journal of Cyber Criminology*, 10(1):1.

Chara Bakalis. 2016. Cyberhate: an issue of continued concern for the council of europe's antiracism commission (2016). Technical report.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049. Association for Computational Linguistics.

Helena Bonaldi, María Estrella Vallecillo-Rodríguez, Irune Zubiaga, Arturo Montejo-Ráez, Aitor Soroa, María-Teresa Martín-Valdivia, Marco Guerini, and Rodrigo Agerri. 2025. The first workshop on multilingual counterspeech generation at coling 2025: Overview of the shared task. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 92–107.

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 Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 166–179, Vancouver, Canada. Association for Machine Translation in the Americas.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022a. Scaling instruction-finetuned language models.

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

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yi-Ling Chung et al. 2022b. Counter narrative generation for fighting online hate speech.

Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

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*. 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. *arXiv preprint arXiv:2006.01974*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. 2024. The llama 3 herd of models.

Sadaf Md. Halim, Saquib Irtiza, Yibo Hu, Latifur Khan, and Bhavani M. Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6716–6733, Mexico City, Mexico. Association for Computational Linguistics.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4523–4536, Miami, Florida, USA. Association for Computational Linguistics.

Laura Illia, Elanor Colleoni, and Stelios Zyglidopoulos. 2022. Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, and Wenxian Wang. 2023. Rezg: Retrieval-augmented zero-shot counter narrative generation for hate speech.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation.

Seungyoon Lee, Chanjun Park, DaHyun Jung, Hyeonseok Moon, Jaehyung Seo, Sugyeong Eo, and Heuiseok Lim. 2024. Leveraging pre-existing resources for data-efficient counter-narrative generation in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10380–10392, Torino, Italia. ELRA and ICCL.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

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

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

Karsten Müller and Carlo Schwarz. 2020. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. Crowdcounter: A benchmark type-specific multi-target counterspeech dataset.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counter-speech. *arXiv preprint arXiv:2205.04304*.

Sougata Saha and Rohini Srihari. 2024. Consolidating strategies for countering hate speech using persuasive dialogues.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news.

Weizhe Zhang, Wenchao Gu, Cuiyun Gao, and Michael R. Lyu. 2022. A transformer–based approach for improving app review response generation. *Software: Practice and Experience*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against 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 (ACL-IJCNLP): Findings*.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A llm-based ranking method for the evaluation of automatic counter-narrative generation.

8. Language Resource References

Bonaldi, Helena and Dellantonio, Sara and Tekiroglu, Serra Sinem and Guerini, Marco. 2022. *Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering*. Association for Computational Linguistics. PID https://github.com/marcoguerini/CONAN.

Chung, Yi-Ling and Kuzmenko, Elizaveta and Tekiroglu, Serra Sinem and Guerini, Marco. 2019. *CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech*. Association for Computational Linguistics. PID https://github.com/marcoguerini/CONAN.

Chung, Yi-Ling and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. *Knowledge-grounded hate countering dataset - Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech*. Association for Computational Linguistics. PID https://github.com/marcoguerini/CONAN.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. *Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech*. Association for Computational Linguistics. PID https://github.com/marcoguerini/CONAN.

Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul. 2023. *SemEval-2023 Task 10: Explainable Detection of Online Sexism*. Association for Computational Linguistics. PID https://github.com/rewire-online/edos.

Punyajoy Saha and Abhilash Datta and Abhik Jana and Animesh Mukherjee. 2024. *CrowdCounter: A benchmark type-specific multi-target counter-speech dataset*. PID https://github.com/hate-alert/CrowdCounter.

A. Human Evaluation Sheet

A.1. Instruction and Guideline

Notion Explanation

Toxic/Offensive: A response is considered toxic or offensive, if it contains contents that can be harmful, degrading, or hurtful towards others. Examples of toxic or offensive speech are:

- Hate Speech: This includes language that discriminates, stereotypes, marginalizes, or stigmatizes individuals or groups based on attributes such as their race, religion, ethnic origin, sexual orientation, disability, or gender.
- Harassment: This can involve threats, intimidation, or repeated unwanted contact or attention.
- Disinformation: Spreading false or misleading information to cause harm or confusion.
- Personal Attacks or Insults: Directly attacking someone's character, appearance, or other personal attributes in a harmful manner.
- Profanity or Obscene Language: Using crude, vulgar, or offensive language.

Relevance: A response is considered relevant, if it revolves generally around the same topic/target as the hate speech, and tackles the hateful messages in the original hate speech.

Acceptance as Counter-Narrative: A response is considered a counter-speech, if it e.g.

- refutes misinformation by providing accurate, evidence-based facts.
- promotes empathy and understanding, humanizes targeted demographics/victims.
- advocates for respectful, constructive conversation
- states concequences/impacts of hateful messages, prevents the hate from further spreading.

Warning

this project contains content that may be offensive or upsetting. These materials are collected from the internet for the purpose of developing countermeasures. These materials are not produced by the authors of this project and do not represent the values of the authors.

Instruction

- During this questionnaire, you will be given 10 hate speech samples. For each hate speech, 5 counter speech are given, making in total 50 counter speech responses.
- For each counter speech response, three YES/NO questions are asked for evaluating three attributes of quality (Toxicity, Relevancy and Acceptance as Counter Speech).
- Each cell is pre-defined as a drop-down menu, where you can simply choose YES or NO from the list items. Please make sure to fill out all the cells.

A.2. Example HS-CN Pairs

Example 1

Hate Speech: Beating women is great.

Counter-Narrative: Women have rights to choose what they want in their life and be whoever they want to be.

☐ Toxic/Offensive

☐ Relevance

☑ Acceptable as Counter-Narrative

The response is a generic counter-narrative about feminism. It failed to address the specific context of violance against women, thus it's considered valid but not relevant.

Example 2

Hate Speech: Beating women is great.

Counter-Narrative: promoting violance is not a great idea.

☐ Toxic/Offensive

☑ Relevance

☑ Acceptable as Counter-Narrative

The response did not address sexism specifically, however, it addresses the main topic of violance and is therefore valid and relevant.

Example 3

Hate Speech: Beating women is great.

Counter-Narrative: Beating women is great.

☑ Toxic/Offensive

☐ Relevance

☐ Acceptable as Counter-Narrative

The response repeats the hate speech and is therefore toxic and invalid as counter-narrative. Whether it is relevant is ambiguous, in this study, we tend to consider copying the hate speech irrelevant.

Example 4

Hate Speech: Beating women is great.

Counter-Narrative: Women should be treated equal because they are human too. Beating only women is not a great idea, all human should be treated equally and be beaten.

☑ Toxic/Offensive

☐ Relevance

☐ Acceptable as Counter-Narrative

The response is toxic since it promotes indifferent violence, however, the response mentioned the idea of feminism/equality and directly addressed the issue of violence against women. Therefore, it's considered a toxic but relevant counter-narrative.

Example 5

Hate Speech: Beating women is great.

Counter-Narrative: I don't know what you're talking about

☐ Toxic/Offensive

☐ Relevance

☐ Acceptable as Counter-Narrative

The response is generic. It avoids confrontation with the hate speech. It's considered not relevant nor valid.

B. Prompts for Counter-Narrative Generation with LMs

Here, we provide the exact prompts used for generating counter-narratives with LMs in this study.

B.1. Prompt for Conversational Agents

Prompt Instructions

You are an UN agent who focuses on fighting online hate speech. Your task is to write responses, so-called counter speech, to online hate speech targeting different demographics.
The responses should not be offensive, hateful, or toxic.
The responses should actively fight against the hateful messages and bring counterarguments to the table.
The aim is to bring positive perspectives, clarify misinformation, and be an active voice for the targeted demographics against hate.
In the following, I will show you some online hate speech. Please create a counter-speech for each hate speech with a maximum length of 256 tokens.
[Hate Speech]

B.2. Prompt for Autoregressive Models

Prompt

<Target-Group>Hate Speech: [Hate Speech] Counter-Speech:

Example:
<MUSLIMS><WOMEN>Hate Speech: Women of our culture that decide to become Islamic are so selfish. They are so happy and proud to join this religion while living in the west but I would like to see them in Pakistan, Saudi Arabia or Sudan. Counter-Speech:

B.3. Prompt for Sequence-to-Sequence Models

Prompt

<Target-Group>[Hate Speech]

Example:
<MUSLIMS><WOMEN>Women of our culture that decide to become Islamic are so selfish. They are so happy and proud to join this religion while living in the west but I would like to see them in Pakistan, Saudi Arabia or Sudan.

B.4. Prompt for Instruction-Tuned Models

Prompt Instructions

```
{
    "role": "system",
    "content": "You are an UN
    agent who focuses on fighting
    online hate speech. Your task
    is to write responses,
    so-called counter speech, to
    online hate speech targeting
    different demographics. The
    responses should not be
    offensive, hateful, or toxic.
    The responses should actively
    fight against the hateful
    messages and bring
    counterarguments to the table.
    The aim is to bring positive
    perspectives, clarify
    misinformation, and be an
    active voice for the targeted
    demographics against hate."
}
{
    'role': 'user',
    'content':
    <Target-Group>[Hate Speech]
},
{
    'role': 'assistant',
    'content': ""
}
```

C. Automatic Evaluation Metrics Models

The model instances used for the automatic evaluation:

- CoLA Classifier (textattack/roberta-base-CoLA)
- Toxicity Classifier (martin-ha/toxic-comment-model)
- Hate-Offensive Classifier (Hate-speech-CNERG/bert-base-uncased-hatexplain)
- Counter-Narrative Classifier (tum-nlp/counter-narrative-classifier)
- Target-Demographic Classifiers (tum-nlp/hatespeech-target-group-classifier, tum-nlp/target-group-classifier-2)
- Universal Sentence Encoders (sentence-transformers/all-MiniLM-L6-v2, sentence-transformers/all-mpnet-base-v2, sentence-transformers/LaBSE)
- Question-Answering Sentence Encoders (sentence-transformers/multi-qa-MiniLM-L6-cos-v1, sentence-transformers/multi-qa-distilbert-cos-v1)

D. Fine-Tuning and Implementation Details

We fine-tuned models using either full parameter updates or parameter-efficient adaptation, depending on the model's size. Specifically, we applied LoRA to **LLaMA-3.2-1B-Instruct**, **Flan-T5-XL** and **Dolphin-2.9-Llama3-8B**, targeting attention layers and newly added target-demographic tokens. Other models were fine-tuned without adapters. **Dolphin-2.9-Llama3-8B** is loaded using 4-bit quantization. For adapter models, we used the following configuration:

- `r` = 8, `alpha` = 32, `dropout` = 0.1

All models were trained using the following hyper-parameters unless otherwise specified:

- `num_train_epochs` = 30 (generators), 100 (classifiers)
- `learning_rate` = 5e-5
- `weight_decay` = 0.01
- `warmup_ratio` = 0.1
- `batch_size` = 8
- `gradient_accumulation_steps` = 4
- `optimizer` = adamw_bnb_8bit
- `lr_scheduler_type` = cosine
- `fp16` = True
- `gradient_checkpointing` = True
- `early_stopping_patience` = 3
- `early_stopping_threshold` = 0.01

E. Sample Responses

Here, we provide example to compare the effectiveness and relevance of generic v.s. Target-Aware counter-narratives.

Hate Speech	Afghanistan or maybe Pakistan? I bet this does not look like an English town to you? #saville Town (Text accompanying video footage of people on their way to Friday prayers etc.)
Human	Small groups of people peacefully gather to pray, but because they do not look like you, you are outraged? Why am I not surprised?
GPT2-Medium Base	I don't know what you are talking about, but I'm not going to let you get away with this.
GPT2-Medium Fine-Tuned	This is not true. How can you say this about an entire faith? I think it is a shame that you would make such a bigoted statement. I would like to see more respect for other faiths and cultures. We should respect each other's culture and traditions. We are all human beings. We all have our own beliefs and beliefs are different from each other.
GPT2-Medium Target-Aware (ours)	I don't think this is an accurate depiction of the people of Saville. I think it is more accurate to say that there are many people from all walks of life who live in this town, and they all contribute to our society. They are our neighbours, our friends, our colleagues, our partners. We are all part of the Saville community, and we all deserve to be treated with respect and dignity. We should all be proud of what we have to offer, and that is why we should be proud to be British.

Table 9: An example of generated CNs from GPT2-Medium models, where our Target-Aware model outputs more relevant counter-narrative.

Hate Speech	All black folk are like slaves: they are weak and childlike. They are incapable of hard work.
Human	Hard work and being a good citizen is not determined by colour, or religion, or ethnicity. This is just an example of white supremacy propaganda.
GPT2-Medium Base	Black folk are not like slaves, they are like children: they have no strength, they cannot work hard, and they are stupid. Black people are not stupid: they can't work hard. They can't be strong. They cannot be intelligent. They have no intelligence. They don't know how to work hard or how to learn. They do not know what to do with their lives. Black people do not have a sense of self-worth. They lack self-respect. Black folk have no sense of their own worth. Black folks have no self-esteem. They live in a world where they are not valued.
GPT2-Medium Fine-Tuned	This is not true. How can you say this about an entire group of people? Do you have any evidence for this? I do not think so. I would like to know more about this statement. Thank you for your time. I will be sure to provide you with the information that you need. We are all human beings. We all have the right to work and to contribute to our society. We should respect each other's rights. This is what democracy is all about. We must respect the rights of all people. We can all work hard and contribute to the country.
GPT2-Medium Target-Aware (ours)	It is not true that all black people are weak or childlike, they are just different from other people. They have the same rights as anyone else. They can work as hard as they want, as long as they do not violate the law. They should be allowed to pursue their dreams. Why should they be punished for it? They are human beings with the right to live their lives as they see fit. They deserve to be treated as such. Why do you think that they are inferior to other people? Do you have any proof for your statement? I would like to know your proof.

Table 10: An example of generated CNs from GPT2-Medium models, where the base model reproduced the hateful input while our Target-Aware version created contexually relevant CN.

Hate Speech	Liberal white whores worship black men, it is becoming impossible for white guys to get laid. Which is what all the dumb normies don't realize is causing the rise in nazism. You will never see non white women doing this: [URL]
Llama-3.2-1B-Instruct Base	I can't help with that. Is there something else I can assist you with?
Llama-3.2-1B-Instruct Fine-Tuned	The problem is not with black people, but with the racist attitudes of some white people.
Llama-3.2-1B-Instruct Target-Aware (ours)	It seems to me that you are saying that if a woman is interested in a man, she should only be interested in men who are white. But this is not true. Women are interested in all sorts of people, and they should be able to choose who they want to be with.

Table 11: An example of Llama-3.2-1B-Instruct generated CNs, where the base model does not reply to the given hate speech, while the fine-tuned versions provided suitable CNs.

Hate Speech	They say vegan soy products have estrogen hormones.... Is that good for anyone? It will fuck a man up.. but what effect would extra estrogen have on women? Would they become perma period psychotics?
Llama-3.2-1B-Instruct Base	Vegan soy products, particularly those derived from soybeans, are often fortified with isoflavones, which are plant-based compounds that can mimic some of the effects of estrogen in the body. However, it's essential to note that not all soy products are created equal, and the estrogen content can vary depending on factors like the type of soy, processing methods, and fortification levels. The concern you're referring to is likely due to the presence of phytoestrogens, particularly genistein and daidzein.
Llama-3.2-1B-Instruct Fine-Tuned	It is not true that women would become ""perma-period psychotics"" because of estrogen hormones in soy products.
Llama-3.2-1B-Instruct Target-Aware (ours)	Vegan diets have been shown to have numerous health benefits, including reducing the risk of heart disease, type 2 diabetes, and certain types of cancer.

Table 12: An example of Llama-3.2-1B-Instruct generated CNs, where the base model generated contexually relevant - but invalid - response to unseen hate speech, while the Target-Aware version provided contexually relevant CN.

Hate Speech	Thats because many of those guys struggle with women, and they think asian women are more submissive than white. Also asian women are light skinned. They ignore other ethnicities.
Llama-3.2-1B-Instruct Base	It sounds like you're expressing some concerns and observations about relationships and demographics. However, I want to address a few points to provide a more nuanced understanding. Firstly, it's essential to recognize that people's experiences and preferences in relationships are highly individualized and can vary greatly. What works for one person may not work for another, and it's not accurate to make sweeping generalizations about an entire group of people. Regarding the idea that many men struggle with relationships, particularly with women from different ethnic backgrounds, this is a complex issue that can be influenced by various factors, including cultural norms, personal experiences, and societal expectations.
Llama-3.2-1B-Instruct Fine-Tuned	Asian women are just as capable of loving and being loved as white women.
Llama-3.2-1B-Instruct Target-Aware (ours)	It is not true that Asian women are lighter skinned than white women. This is a myth that has been around for a long time, and it is not supported by any scientific evidence.

Table 13: An example of Llama-3.2-1B-Instruct generated CNs, where the base model generated contexually relevant CN to unseen hate speech, while the fine-tuned versions provided rather generic CNs.

F. Contexual and Semantic Similarity Data Examples

Here, we provide further examples that demonstrate the effect of contextual similarity to hate speech and semantic similarity to reference.

Hate Speech	Muslims are not welcome in our country. They are not loyal citizens!
Human	Muslims are often accused of being not loyal to our country, but they are proud to be British as anyone else, and proud to be a part of our society.
Llama-3.2-1B-Instruct	I can't engage in a conversation that promotes hate speech or discrimination towards any individual or group based on their religion, ethnicity, or nationality. Can I help you with something else?

Table 14: An example of generated response that scores low in both contextual relevance (0.44) to input and semantic relevance to reference (0.49).

Hate Speech	Jewish folks are obsessed with the holocaust.
Human	Jews were not obsessed with the holocaust. The murder of some 6 millions Jews by the Nazis shocked the world. Most of the survivors were only killed because they were suspected of belonging to a particular community. The Jewish people suffered greatly as a result of persecution, and the holocaust should be remembered not to let this happen again.
Llama-3.2-1B-Instruct	The Holocaust is a well-documented and widely recognized historical event. It's not accurate to say that Jewish people are ""obsessed"" with the Holocaust. The Holocaust, also known as the Shoah, was a systematic, state-sponsored persecution and murder of six million Jews by the Nazi regime and its collaborators during World War II. The atrocities committed during this period are well-known and widely studied in history classes around the world. Jewish people, like people of many other faiths and backgrounds, have indeed paid tribute to the victims of the Holocaust through various memorials, museums, and commemorative events.

Table 15: An example of generated response that scores high in both contextual relevance (0.91) to input and semantic relevance to reference (0.91).

Hate Speech	We have a lot of issues in the UK, and now we are also paying NGOs to rescue migrants. How many other financial scandals still have to be announced to the tax payers?
Human	The problem is that most of the British population is not aware of the fact that the media use the "scandal" of immigrants to hide bigger problems for the country; such as taxes not paid by millionaires and large corporations which then fall on us citizens.
Llama-3.2-1B-Instruct	It's essential to address your concerns about the UK's handling of financial scandals involving NGOs and migrant rescue efforts. While I don't have real-time information, I can provide some context and insights. There have been several high-profile cases of NGOs and non-governmental organizations (NGOs) being linked to financial scandals, corruption, or abuse of public funds in various countries, including the UK. Some examples include: 1. The Panama Papers scandal (2016): A massive leak of financial documents revealed widespread tax evasion, money laundering, and other financial crimes by individuals and organizations, including some NGOs and charities. 2.

Table 16: An example of generated response that scores high in both contextual relevance (0.83) to input and semantic relevance to reference (0.72), but classified as irrelevant by target-group-classifier.