CounterSpeech Modeling

Anant Kaushal anant22067@iiitd.ac.in
Anikait Agrawal anikait22072@iiitd.ac.in
Ansh Varshney ansh22083@iiitd.ac.in

1 Abstract

This work aims to improve the efficiency of counterspeech generation by means of an original intent-specific method. Although earlier attempts mostly relied on generic responses to online hate speech, our improved system seeks to produce responses that are context-aware and consistent with the intent underlying the hateful content. We customise counterspeech to more fit audience expectations and platform specific dynamics by including different categories such as informative, denouncing, questioning, positive, and humorous. This improved approach greatly increases the possibilities of counterspeech as a scalable and non-censorshipable solution to support better online interactions.

2 Introduction

Online hate speech presents significant social and psychological threats, fueling division, reinforcing harmful stereotypes, and even inciting real-world violence. While *Counterspeech*—constructive responses to hateful content—has emerged as a powerful, non-censorious method to combat hate, most current approaches rely on generic replies that often fail to address the tone or intent of the original message. To be truly effective, counterspeech must be **intent-specific**, adapting its tone—whether *informative*, *denouncing*, *questioning*, *positive* or *humorous* on the context and platform. For instance, correcting misinformation about vaccines on Twitter, using empathy to support bullying victims on Instagram, or replying to sexist comments on Reddit with light sarcasm can each play a powerful role in challenging hate while engaging the community.

This approach is particularly relevant in the context of social media platforms, where content moderation struggles to keep up with the scale and nuance of hateful speech. Intent-specific counterspeech offers a scalable, user-driven solution that not only addresses hate but also fosters healthier online conversations. It empowers users to respond thoughtfully, resonates more with diverse audiences, and is better suited to de-escalate conflict and influence bystanders. As platforms continue to grapple with the limits of automated moderation and the dangers of censorship, working on improving counterspeech systems is both timely and impactful for creating safer, more inclusive digital environments.

3 Related Work

Previous research on counterspeech generation has explored various directions, ranging from data collection to evaluation metrics and model architectures. Notable approaches include:

Multilingual Counterspeech with Human-in-the-loop Mechanisms One line of work focuses on building richer, multilingual counterspeech datasets. A notable example is the multi-target approach that incorporates human-in-the-loop mechanisms to collect high-quality counterspeech across diverse communities and languages [1]. This method ensures contextual relevance and cultural sensitivity in responses. Shortcoming: While effective, such methods are labor-intensive and difficult to scale, particularly for real-time applications or under-resourced languages.

Improved Evaluation using the METEOR Metric Evaluating counterspeech generation remains challenging. The METEOR metric [2], originally designed for machine translation, has been adopted in some studies for its closer correlation with human judgments compared to metrics like BLEU. It accounts for synonyms, stem variations, and word order, offering more nuanced assessment of generated text. Shortcoming: Despite its advantages, METEOR is not tailored to dialogue or counterspeech tasks, often failing to reflect key qualities like tone, empathy, or intent alignment.

Intent-conditioned Counterspeech Generation More recent work has explored intent-conditioned architectures that disentangle style (intent) and content during training [3]. This allows for more targeted responses, where the generated counterspeech aligns with a desired intent—such as informative, humorous, or denouncing—enhancing its effectiveness and user engagement. Shortcoming: These systems often require large, well-labeled datasets for each intent category, and may suffer from coherence issues when style and content are not harmoniously integrated.

4 Methodology

We implemented two baselines and one novelty , which are described below:

4.1 BART

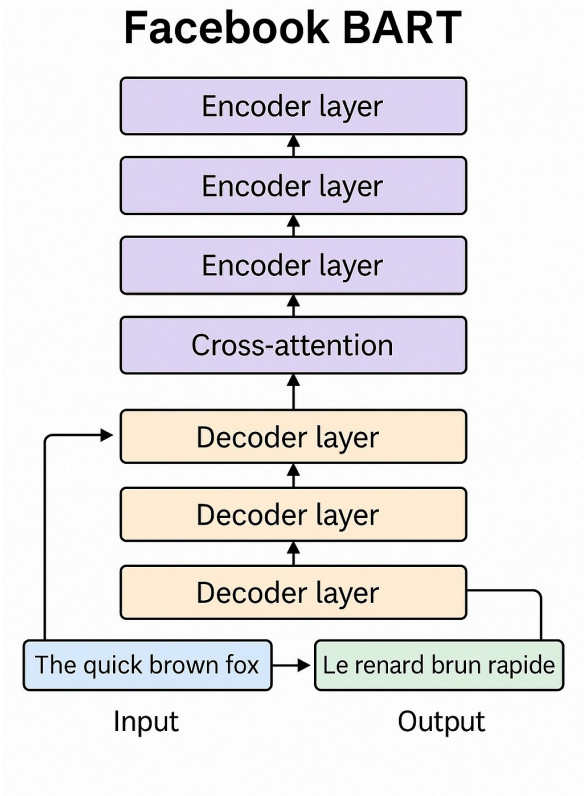

Figure 1: BART Architecture

Intent-Conditioned Sequence-to-Sequence Generation Our first baseline explores a fine-tuned sequence-to-sequence model, specifically using BART, to generate intent-specific counterspeech. The model is conditioned on both the hate speech text and its corresponding intent label, drawn from five predefined categories. To guide the generation, we construct the input sequence in the format: intent: [INTENT] hatespeech: [TEXT] This formulation enables the model to learn to produce counterspeech that is not only contextually relevant but also aligned with the desired response style.[3]

We evaluate this baseline using BLEU and BERTScore metrics. The model achieves a BLEU score of approximately 0.02 and a BERTScore of 0.87 on the validation set. These results suggest that while the generated counterspeech maintains strong semantic similarity to the ground truth (as reflected in the high BERTScore), it exhibits limited lexical overlap, as indicated by the low BLEU score. This highlights the model's ability to capture meaning, but also points to challenges in producing token-level matches, which are often important for evaluating stylistic and structural fidelity.

4.2 QUARC

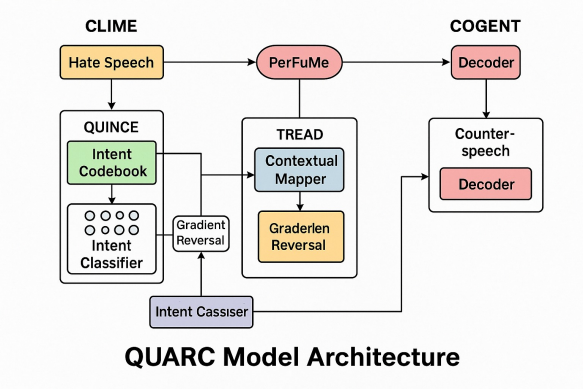

Figure 2: QUARC Architecture

A Two-Stage Intent-Conditioned Generation Framework Our second baseline implements QUARC, a novel two-phase framework designed for intent-conditioned counterspeech generation. The methodology first employs a codebook learning phase (CLIME), which decomposes counterspeech into semantics and intent, and learns vector-quantized representations for five predefined intents: informative, denouncing, question, positive, and humor. These representations are then integrated in the generation phase (COGENT), where hate speech input is semantically mapped and fused with the corresponding intent vector using a persistent fusion mechanism, PerFuMe, to produce intent-aligned counterspeech.[2]

Evaluation is conducted using both lexical (ROUGE, METEOR) and semantic (BERTScore, sentence similarity) metrics, along with category accuracy (intent alignment). QUARC outperforms baselines such as BART, DialoGPT, GPS, and

PPLM by an average margin of (10%) across these metrics. Notably, it achieves high diversity and novelty in generated responses, indicating its strength in producing varied and contextually appropriate counterspeech. However, shortcomings include slightly elevated toxicity in some outputs and difficulty in generating humorous responses, reflecting the subjective nature and data sparsity of this intent category.

4.3 R3-QUARC

The R3-QuARC architecture improves on standard question answering by integrating supervised training with an iterative reinforcement learning (RL) stage. First, a sequence-to-sequence "actor" model (based on Facebook's BART) is fine-tuned with supervised learning on counterspeech data. Concurrently, a critic model (based on DistilBERT) is trained as a quality classifier to determine how well generated responses match high-quality counterspeech, using a special column.[1] Then, in the RL phase, for every input, the actor samples multiple candidate responses (each with a corresponding rationale) through sampling. These candidates are rated by the critic to generate a reward signal that mirrors their quality. A policy gradient update along the lines of PPO is used so that the actor updates its generation behavior according to the rewards received, essentially learning to generate outputs that are accurate and well-supported. Lastly, the whole system is tested using ROUGE and other critic-based metrics to ensure iterative improvement over response quality and alignment.[4]

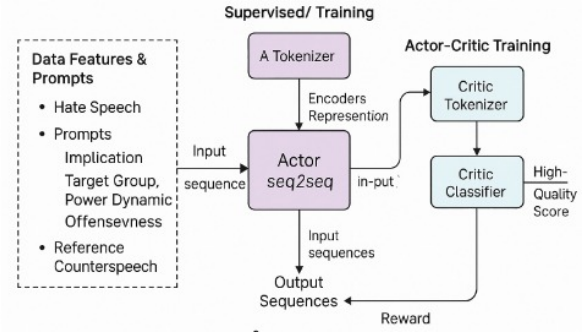

Figure 3: R3-QUARC

5 Dataset, Setup and Results

The IntentCONAN dataset is centered around classifying the intent behind counterspeech (CS) in reaction to hate speech (HS). It improves on previous works, like CONAN and Mathew et al., but because of the small amount of data, some of the intent categories were combined, giving five major categories: informative, question, denouncing, humor, and positive. The dataset is obtained from the publicly released Multi-Target CONAN dataset that comprises approximately 5,000 HS-CS pairs but did not have intent labels. To correct this, the dataset was cleaned, de-duplicated, and annotated by domain experts who applied intent labels to available counterspeech and created new ones where appropriate. Every instance of hate speech in IntentCONAN has an average of two counterspeeches, one from each of the intent categories defined. The dataset is utilized to investigate the generation of different types of counterspeech in reaction to hate speech against different groups.

For our analysis, we performed Exploratory Data Analysis (EDA) on the dataset to gain a better understanding of the distribution of intent categories, target groups, and the connection between hate speech and counterspeech. We looked at the frequency of each intent category, compared the distribution of counterspeech by different hate speech targets, and determined if there were any patterns or biases in the data. This use a dataset [3] of 6,831 counterspeeches across five intents (*informative*, *denouncing*, *question*, *positive*, *humour*). Key steps include:

- **Data Splits:** Train (70%), Validation (15%), Test (15%). For Novelty , we have concatenated the Train , Test and Validation , then split that concatenation to Train (40%), Validation (30%), Test (30%).
- **Pre-processing:** For Baselines , We cleaned the dataset by removing null values and verified balanced class distribution through exploratory data analysis. In Novelty ,the preprocessing pipeline ensures input data is standardized for actor and critic models. Firstly, CSV files containing hate speech, counterspeech, and other fields for annotation are loaded with an anticipated schema that ensures uniform data types across splits. The data is subsequently merged, shuffled, and stratified by a quality label to provide class balance for training, validation, and test sets.For the actor model, every example is converted by combining the counterspeech type with the hate speech text to create the input, and by adding a constructed rationale derived from

auxiliary annotations to the reference counterspeech output. Inputs and outputs are tokenized with a fixed maximum length for consistency, and padding tokens in the target are replaced with -100. For the critic model, the hate speech and counterspeech texts are concatenated and tokenized, and the corresponding quality label is made ready for training. Lastly, extraneous columns are dropped and the sets are reshaped to directly interact with PyTorch DataLoaders and accommodate both supervised and reinforcement learning stages.

- **Evaluation:** In Baseline I , we measured BLEU and BERTScore, finding BLEU ≈ 0.02 and BERTScore ≈ 0.87 on the validation set. For Baseline II, we conducted a comprehensive human evaluation, including a Turing Test, to assess the naturalness, relevance, and coherence of the generated responses. Human annotators were asked to compare system outputs with ground truth and judge whether the responses could be distinguished from those written by humans. In Novelty , the evaluation metrics used architecture consist of ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum) to check for the lexical overlap and content similarity of generated responses with reference texts and critic validation accuracy in the supervised phase to check the performance of the quality classifier. Throughout reinforcement learning, policy loss is calculated as the negative log-likelihood of generated sequences weighted with a reward signal from the critic, and some further statistics are monitored—namely, the mean critic high-quality probability and the proportion of outputs above a threshold probability of 0.5. All of these, together, form an overall evaluation of both text generation quality and how well the model conforms to target quality specifications.

6 Observations and Analysis

While our model yields a high BERTScore (0.87) but low BLEU (0.02), it indicates good semantic alignment but limited token-level overlap. To address this, we propose a two-stage architecture: (1) a lexical reconstruction module that encourages token-level fidelity, and (2) a style-conditional cross-encoder to maintain semantic coherence. Incorporating external knowledge bases and reinforcement learning for style fidelity can further refine counterspeech generation. The EDA plots show prominent distributions throughout the dataset. The csType feature is equally distributed between Informative, Questioning, Denouncing, and Positive classes. Suggest is skewed, with most being labeled with 3. Relevance is tightly concentrated in between 3 and 4, with medium to high relevance. Complexity is mostly level 2, indicating that the content is fairly simple, with very few complex entries. Overall, the dataset is style-balanced but unbalanced otherwise.

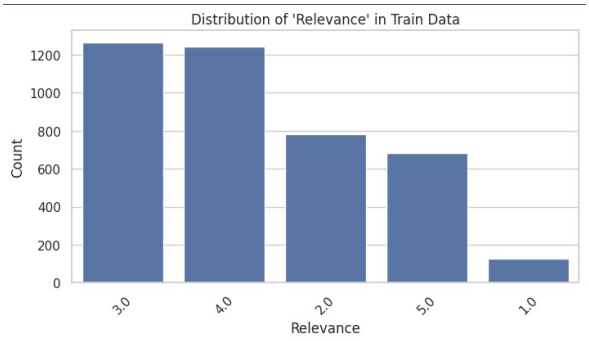

Figure 4

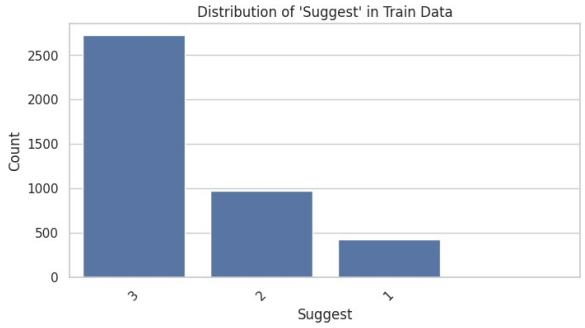

Figure 5

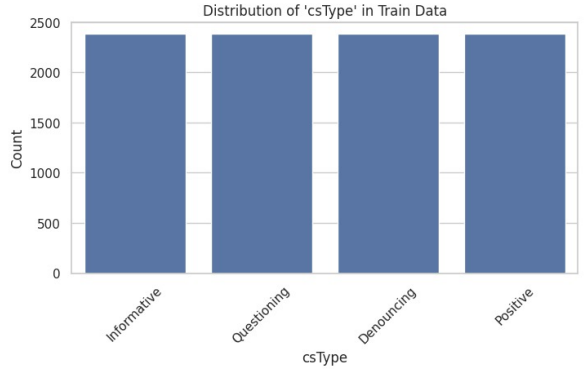

Figure 6

The heatmap indicates the correlation among

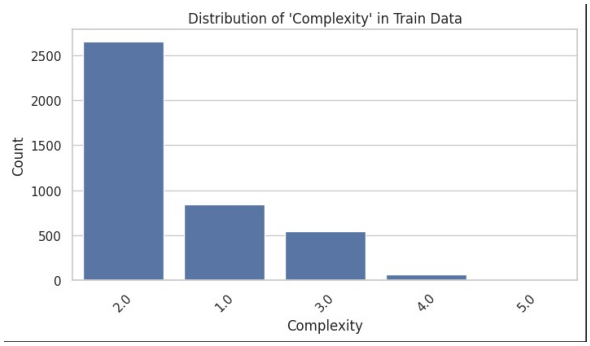

Figure 7

Hate Speech		Counterspeech Intents					
Targets	**Counts**	**INF**	**QUE**	**DEN**	**HUM**	**POS**	**Total**
Muslims	968	671	450	255	107	265	1748
Migrants	642	453	241	134	107	165	1100
Women	517	415	225	195	158	158	1151
LGBT+	465	280	195	145	99	132	851
Jews	408	272	184	109	96	112	773
POC	294	226	136	118	71	71	622
Disabled	173	114	45	44	25	61	289
Other	116	85	66	51	41	54	297
Total	3583	2516	1542	1051	704	1018	6831
Train	2508	1761	1079	735	494	712	4781
Dev	716	507	310	212	139	205	1373
Test	359	248	153	104	71	101	677

Figure 9: Distribution Table

various counterspeech intents. Informative is positively correlated with Denouncing (0.52) and Positive (0.54), whereas Humor negatively correlates with all other intents, particularly Positive (-0.45) and Informative (-0.4). Questioning weakly correlates with the others, with the strongest being with Denouncing (0.25). Generally, Positive, Informative, and Denouncing co-occur, whereas Humor is an outlier with inverse relationships.

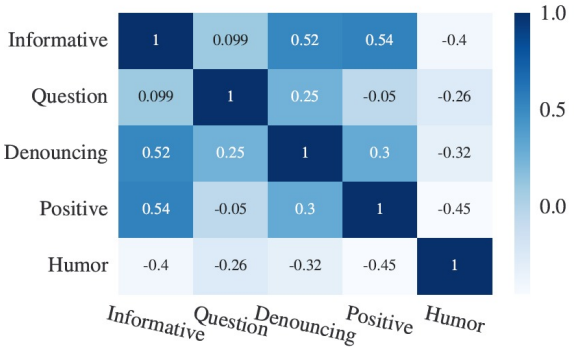

Figure 8: Confusion Matrix

The distribution of hate speech targets and accompanying counterspeech intentions is illustrated through the table. Muslims are most targeted, and then come the Migrants followed by Women. Informative (INF) is the most used counterspeech intention for all the targets, with Humor (HUM) being the one used least often. There are a total of 6,831 counterspeech samples, which are divided between Train (4,781), Dev (1,373), and Test (677) splits. Each of the subsets is evenly distributed according to the intent types so as to have them balanced across the splits.

The ROUGE scores demonstrated in the new architecture ROUGE 1 24.96%, ROUGE-2 7.20%, and ROUGE-L 18.61% at epoch 2 speak of improvement in the capability of the model to produce text closely following human reference summaries. Comparison between epoch 1 and epoch 2 reflects improved paraphrasing capability and content selection. These are critical in measuring the capacity of the model in grabbing the key content and nuances of the input, affecting the novelty and quality of the output responses directly.

7 Conclusion and Future Work

This paper showcases the progression of intent-conditioned counterspeech generation by three consecutive approaches. Baseline I employed a simple sequence-to-sequence model (BART) conditioned on hate speech and intent and attained high semantic similarity as reflected in high BERTScore but was affected by low lexical overlap (low BLEU). Baseline II extended this direction with the QUARC framework by breaking down counterspeech into style and content components separately through a codebook learning stage (CLIME) and ongoing fusion (COGENT), increasing diversity as well as intent-matching. Expanding on these, the R3-QuARC novelty embeds reinforcement learning within the generation process by pairing an improved actor model with a specialist critic. This incremental actor-critic method not only adapts generation via policy gradients but also outputs that more accurately represent high-quality, context-responsive counterspeech. Subsequent efforts can continue to improve this framework by integrating in external knowledge as well as reconstruction of lexical characteristics, continuing the innovation of auto, intent-dependent online moderation.

8 Contributions

Anant Kaushal: Data Collection and Preprocessing, BART Model Implementation, QUARC Architecture (CLIME Module), Evaluation Metrics

(BLEU, BERTScore), Observations and Figures (EDA), PPT Preparation

Ansh Varshney: Literature Survey and Related Work, QUARC Architecture (COGENT Fusion), Dataset Annotation and Splitting, Evaluation (METEOR, Human Evaluation), Final Report Writing and Editing

Anikait Agrawal: Data Preprocessing and Visualization, R3-QUARC Actor-Critic Model Implementation, Reinforcement Learning Integration (PPO), Feature Tokenization and Critic Evaluation, ROUGE and Policy Loss Tracking, PPT Preparation

References

[1] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

[2] 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. *arXiv preprint arXiv:2305.13776*.

[3] Adem Karahoca. 2012. *Advances in data mining knowledge discovery and applications*. BoD–Books on Demand.

[4] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.