Project Proposal for CounterSpeech Modeling

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

1 Introduction

Online hate speech poses serious social and psychological threats, often escalating tensions within communities. *Counterspeech*—a positive, corrective response to hateful content—has been shown to help mitigate harm but is challenging to generate at scale. While many approaches produce *generic* replies, the need for **intent-specific counterspeech** (e.g., *informative*, *denouncing*, *questioning*, *positive*, *humorous*) is crucial for aligning responses with context and audience expectations.

2 Related Work

Research on hateful content has examined counterspeech generation from various angles. For example, efforts to collect richer data for multilingual counterspeech were described in a multi-target approach that leverages *human-in-the-loop* mechanisms [1]. Additionally, the *METEOR* metric was introduced to improve correlation with human judgments in machine translation tasks, influencing how we might evaluate text generation [2]. Recent developments in *intent-conditioned counterspeech* introduce novel architectures for learning separate representations of style and content; this helps achieve better performance and more precise control in responses [3].

3 Methodology

We propose to fine-tune a **sequence-to-sequence model** (e.g., BART), conditioning on both the **hate speech** and its desired **intent label** (from five categories). Specifically, we concatenate:

$$\text{“}intent : [INTENT] hatespeech : [TEXT]''$$

as input, prompting the model to generate the corresponding *counterspeech*.

4 Dataset, Setup and Observations

We 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%).
- **Pre-processing:** We cleaned the dataset by removing null values and verified balanced class distribution through exploratory data analysis.
- **Evaluation:** We measure BLEU and BERTScore, finding BLEU $\approx$ 0.02 and BERTScore $\approx$ 0.87 on the validation set.

5 Conclusion and Future Work

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.

6 References

[1] Multi-target approach, multilingual counterspeech dataset, and human-in-the-loop data curation.

[2] Introduction of METEOR for improved correlation with human judgments.

[3] Novel intent-conditioned counterspeech architecture using separate style and content representations.