Intent-Aware Counterspeech Generation

Mehul Agarwal
IIIT Delhi
New Delhi, India
mehul222294@iiitd.ac.in

Noel Abraham Tiju
IIIT Delhi
New Delhi, India
noel22338@iiitd.ac.in

Rahul Ramesh Omalur
IIIT Delhi
New Delhi, India
rahul22392@iiitd.ac.in

Abstract

We explore intent-specific counterspeech generation to tackle hate speech online. Using the IntentCONAN dataset—6,831 examples across five intents—we aim to generate targeted, effective, and relevant counterspeech responses.

1 Introduction

Hate speech remains a persistent issue on online platforms, harming communities and spreading hostility. While content removal is common, it often treats the symptoms rather than the cause.

This project focuses on **counterspeech**—civil, reasoned replies to hate. We explore **intent-aware** generation, creating responses that align with one of five specific intents:

- **Informative:** Offers factual corrections.
- **Denouncing:** Condemns the message.
- **Question:** Challenges assumptions.
- **Positive:** Shows empathy or kindness.
- **Humour:** Uses wit to defuse tension.

2 Dataset Description

We use the IntentCONAN dataset, which contains 6,831 counterspeech samples categorized into the five aforementioned intents. Each sample is paired with a corresponding hate speech instance, providing a rich context for intent-specific counterspeech generation.

3 Baseline

As a baseline, we fine-tuned DialoGPT on the IntentCONAN dataset to generate intent-conditioned responses.

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 1: Counterspeech intent distribution in the IntentCONAN dataset.

Metric	Score
ROUGE-1	0.1297
ROUGE-2	0.0030
ROUGE-L	0.1045
METEOR	0.0670
BERTScore (Precision)	0.7914
BERTScore (Recall)	0.8076
BERTScore (F1)	0.7993

Table 1: Evaluation results for the baseline DialoGPT model.

4 Future Work

We plan to expand on the baseline **GPS (Generate–Prune–Select)** approach by generating multiple candidate responses, pruning weaker ones, and selecting the most intent-aligned reply.

We also aim to:

- Integrate **intent distribution learning** to better fuse intent signals during generation.
- Use **multi-task instruction tuning with RLAIF** to boost diversity and reduce toxicity.
- Evaluate via both automatic metrics (e.g., BLEU, ROUGE) and human assessments for fluency, relevance, and intent alignment.
- Optimize model performance and document findings with clear analysis and visuals.

References

[1] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhkavi, 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.

[2] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. *DialoGPT: Large-Scale Generative Pretraining for Conversational Response Generation*. arXiv preprint arXiv:1911.00536.

[3] Wanzheng Zhu and Suma Bhat. 2021. *Generate, Prune, Select: A Pipeline for Counterspeech Generation*. arXiv preprint arXiv:2106.01625.

[4] Amey Hengle, Aswini Kumar, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakroborty. *Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF*. arXiv preprint arXiv:2403.10088, 2024.