RspectNLP@LT-EDI 2026:Rubric-Driven Prompting for Safe Multilingual Counter Narrative Generation

S.B.Priya
St. Joseph's Institute
of Technology
OMR, India
priyait0843@gmail.com

B.Bharathi
Sri Sivasubramaniya Nadar
College of Engineering
Kalavakkam, India
bharathib@ssn.edu.in

Abstract

The problem of harmful online discourse against the LGBTQ+ community is still a concern on social media platforms. Although hate speech detection is a well-explored area, the task of constructive counter-narrative generation is still an emerging field of research, especially in the multilingual and low-resource settings. Counter-narratives are designed to counter harmful discourse with respectful and empathetic responses, as opposed to mere content deletion. In this paper, the model proposes a zero-shot multilingual system for counter-narrative generation in English and Tamil. The proposed system employs the pretrained `google/flan-t5-base` transformer model guided by rubric-aligned prompts to encourage politeness, contextual relevance, and non-toxic response generation. The system operates in a zero-shot setting without task-specific fine-tuning and uses beam search decoding for controlled response generation. On the English test data, the system scored an overall score of 70.33 per cent with a contextual coherence score of 81.82 per cent. On the Tamil test data, the system scored an overall score of 33.57 per cent with significantly lower scores on coherence and quality. These findings indicate that structured prompting can facilitate safe and coherent generation in English, but also underscore the challenges of zero-shot multilingual models in low-resource language scenarios.

Keywords: Counter Narrative Generation Homophobia and Transphobia Detection Multilingual Natural Language Processing Zero-Shot Learning Low-Resource Languages.

1 Introduction

Online forums have grown to be important venues for social interaction and public debate. They do, however, also post damaging information that targets LGBTQ+ populations and other vulnerable groups. Speech that is homophobic or transphobic can exacerbate animosity, promote prejudice, and have a detrimental impact on mental health. As a result, creating responsible language technology to counteract such damaging talk has emerged as a top research priority. The majority of current research in this field focuses on identifying and eliminating hate speech. Detection systems are crucial, but they don't necessarily foster comprehension or lessen prejudice. Eliminating offensive material might make it less visible, but it doesn't always promote healthy discussion. An alternative strategy is provided by counter-narrative generation. It produces courteous and educational reactions that dispel bias and foster empathy rather than stifling offensive speech. Research in multilingual and low-resource languages is still scarce, despite the fact that counter-narrative production has drawn attention in English. The lack of labelled data and linguistic variation in Tamil creates more difficulties. It is so challenging to design systems that produce logical, culturally sensitive, and contextually appropriate answers in such circumstances. In this work, we propose a zero-shot multilingual counter-narrative creation system in Tamil and English. We use a pretrained transformer-based sequence-to-sequence model guided by rubric-aligned prompting. The model is specifically instructed by the prompts to refrain from using harmful words, maintain civility, and guarantee contextual relevance. The results of the experiment show that while performance in Tamil is much poorer, it is moderate in English, especially in contextual coherence. These results demonstrate the limitations of zero-shot multilingual production in low-resource settings as well as the promise of rubric-guided prompting. This effort advances the creation of language-generating systems that are more secure and inclusive. Unlike fine-tuning-based approaches, the proposed system relies entirely on structured prompting in a zero-shot setting. This enables lightweight

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 212–216
July 4, 2026 ©2026 Association for Computational Linguistics

deployment without additional training cost while allowing evaluation of multilingual generation behaviour in both high-resource and low-resource languages.

2 Related Work

Recent advances in natural language processing have expanded research beyond hate speech detection toward constructive response generation. While traditional systems primarily focused on classification, newer approaches aim to generate counter-narratives that challenge harmful content in a respectful manner. Several recent studies explore alignment-based generation methods.(Wadhwa et al., 2025) introduced a multilingual counter speech system that applies Direct Preference Optimisation to better align model outputs with human judgments. Their findings suggest that preference alignment improves contextual appropriateness and tone control. Similarly, (Jiang et al., 2025) proposed a retrieval augmented zero-shot framework that enhances relevance by incorporating external contextual information during generation. While retrieval improves grounding, it increases system complexity. Fact-grounded counter-narrative generation has also gained attention.(Wilk et al., 2025) demonstrated that grounding responses in verifiable information improves credibility and persuasiveness. In low-resource settings,(Prasannan et al., 2025) investigated counter speech generation for Malayalam, revealing the difficulty of generating stable and culturally appropriate responses without language-specific adaptation. Parallel to generation research, multilingual hate speech detection has progressed significantly.(Mnassri et al., 2024) proposed a semi-supervised adversarial framework to improve cross-language classification.(Ghosh and Senapati, 2025) analyzed transformer based monolingual and multilingual models for low-resourced Indian languages and observed notable performance gaps. Findings from shared tasks (Ghosh et al., 2025) confirm that low-resource languages continue to face challenges in both detection and generalisation. Beyond detection, research has emphasised safety and responsible generation. Comprehensive reviews (Albladi et al., 2025) highlight the need for fairness and toxicity control in large language models. Multimodal approaches (Saddozai et al., 2025; Raza Ur Rehman et al., 2025) extend hate speech analysis beyond text, incorporating images and code-mixed data. Prompt-based safety frameworks such as PromptGuard (Vu et al., 2025) demonstrate how structured prompting can improve controllability and reduce harmful outputs. The detection and mitigation of homophobic and transphobic content on internet platforms has been the subject of recent research. Chakravarthi (Chakravarthi, 2024) investigated the automated identification of such detrimental remarks in YouTube data. Further research on span identification techniques for low-resource languages was conducted by Kumaresan et al. (Kumaresan et al., 2025). The shared task presented by Kumaresan et al. (Kumaresan et al., 2026) offers a baseline for multilingual counter-narrative production, whereas Prasannan et al. (Prasannan et al., 2025) suggested counter-speech generation for homophobic and transphobic content. Recent studies, such as Saha et al., explored zero-shot counter-speech generation using prompt-based strategies, demonstrating the growing importance of prompt engineering for safe language generation. Despite these developments, limited research has examined zero-shot multilingual counter-narrative generation without fine-tuning or retrieval augmentation. The present study addresses this gap by evaluating a zero-shot multilingual transformer model guided solely by rubric-aligned prompting in English and Tamil.

Table 1: Dataset statistics for counter narrative generation.

Lang	Split	Homo	Trans	Total
Tamil	Train	342	458	800
Tamil	Test	73	36	109
English	Train	1044	756	1800
English	Test	49	17	66

3 Dataset Description

The collection is made up of Tamil and English social media comments that contain homophobic or transphobic content, as shown in Table 2. A brief comment is represented by each instance. The goal of Task 2 is to produce one positive counter-narrative for every negative remark. Compared to Tamil, the English dataset is bigger and more evenly distributed. Tamil is a low-resource environment, which makes generating more difficult. Evaluation of multilingual counter-narrative production in both high-resource and low-resource settings is made possible by this distribution.

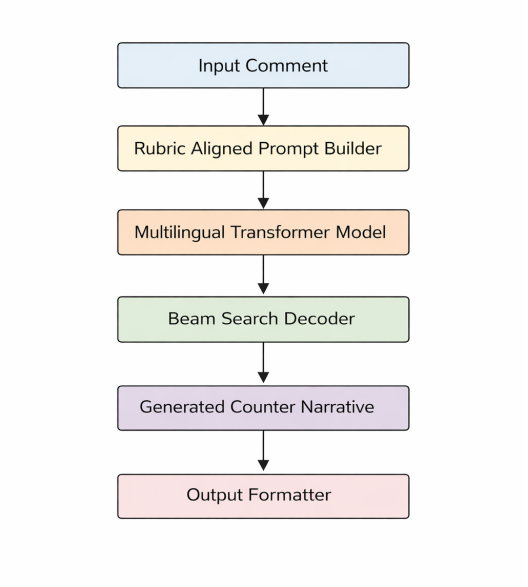

Figure 1: Architecture of the proposed multilingual counter-narrative generation system.

4 Proposed Methodology

This section outlines the design of the counter-narrative generation system. The system is a structured pipeline that turns a toxic comment into a respectful and constructive response.

System Overview

As shown in Figure 1, the system takes a comment as input that is homophobic or transphobic in nature. The comment is then put through a rubric-aligned prompt builder. The structured prompt is then fed into a multilingual transformer model, which produces a counter-narrative through controlled decoding. Finally, the response is formatted in accordance with the required submission format. The main objective of the system is to produce responses that are polite, relevant, and non-toxic.

Rubric Aligned Prompt Design

4.1 Prompt Template

The model was guided using a rubric-aligned prompt designed to encourage politeness, contextual relevance, and non-toxic language generation.

Example prompt:

> Generate a respectful and constructive counter-narrative for the following harmful comment. The response should discourage hate, promote empathy, and avoid aggressive or toxic language.
>
> Comment: [INPUT COMMENT]

Prompt design is a key aspect of our approach. Rather than allowing the model to respond freely, we design the prompts to be aligned with the rubric criteria. For the generating task, a multilingual sequence-to-sequence transformer model is used. The machine can produce responses in Tamil and English and has been trained on a sizable multilingual dataset. The system operates in a zero-shot environment. There is no task-specific fine-tuning. Rather, the model generates responses based solely on the specified request.

4.2 Decoding Strategy

Beam search decoding was used to improve response stability and coherence. Let y denote the generated response and x denote the input prompt. The model aims to maximise the conditional probability:

$$y^* = \arg\max_y P(y \mid x) \tag{1}$$

Beam search maintains the top candidate sequences at each decoding step and selects the most probable response.

Label Assignment and Output Formatting

As the test data does not contain labels, we use a simple keyword-based rule to assign a label. Comments about gender identity are assigned the label transphobia, and others are assigned the label homophobia. This enables us to format the output in the required submission format. For every input comment, the system generates one counter-narrative. The output contains the fields Id, text, span, counter-narrative, and label. The span field is assigned the value NA, as span detection is not required in this task.

Design Rationale

The system design emphasises safety and organisation. Instead of fine-tuning, the research emphasises prompt engineering that is aligned with the evaluation criteria. This enables us to assess the pros and cons of zero-shot multilingual generation in a controlled setting.

5 Evaluation

The system is assessed using both reference-based metrics and rubric-based human evaluation scores. The reference-based metrics calculate similarity between the produced counter-narrative and the gold reference response. The rubric-based metrics calculate politeness, quality, and contextual coherence.

Table 2: Dataset statistics for counter narrative generation.

Lang	Split	Homo	Trans	Total
Tamil	Train	342	458	800
Tamil	Test	73	36	109
English	Train	1044	756	1800
English	Test	49	17	66

The final ranking is obtained by aggregating the scores.

5.1 Distinct-2

Distinct-2 measures response diversity using the ratio of unique bigrams to total bigrams:

$$\text{Distinct-2} = \frac{U_2}{T_2} \tag{2}$$

5.2 BERTScore

BERTScore evaluates semantic similarity between generated and reference responses using contextual embeddings.

$$P = \frac{1}{|G|} \sum_{g \in G} \max_{r \in R} \text{sim}(g, r) \tag{3}$$

$$R = \frac{1}{|R|} \sum_{r \in R} \max_{g \in G} \text{sim}(r, g) \tag{4}$$

$$\text{BERTScore-F1} = \frac{2PR}{P + R} \tag{5}$$

5.3 Results

The performance of the proposed system is evaluated using both reference-based and rubric-based metrics. Table **??** summarises the results for the English and Tamil test sets. To examine the effect of rubric-aligned prompting, preliminary comparisons were conducted using a generic zero-shot prompt. The rubric-guided prompt produced more contextually relevant and less repetitive responses, particularly in English. A detailed quantitative baseline comparison is left for future work.

The model obtains a good answer variety with a Distinct-2 score of 78.56 and strong contextual coherence with a CCNC score of 81.82 for the English test set. The politeness and quality scores, however, are still mediocre, suggesting the need for more thorough explanations and better tone consistency. The Tamil findings, on the other hand, indicate a marked drop in performance. Weak coherence and quality are shown by the QS score falling to 11.47 and the CCNC score to 7.80 The overall score of 33.57 emphasises the difficulties of zero-shot multilingual creation in low-resource contexts, even though the BERTScore-F1 is still comparatively high at 80.23. Code for the proposed system is available in the link ([1])

6 Limitations

The proposed system relies entirely on zero-shot prompting without task-specific fine-tuning. As a result, performance decreases significantly in low-resource settings such as Tamil. The system also depends on keyword-based label assignment, which may not generalise well to complex linguistic expressions. Future work can explore fine-tuning, retrieval augmentation, and culturally adaptive prompting strategies.

7 Conclusion

The research introduced a zero-shot multilingual method for Tamil and English counter-narrative generation in this study. The system directs a pretrained transformer model toward courteous, pertinent, and non-toxic responses using rubric-aligned prompts. It makes use of controlled decoding and organised prompting rather than task-specific fine-tuning. The results of the experiment indicate that English performance is mediocre, especially when it comes to semantic similarity and contextual coherence. However, quality and etiquette still need to be improved. Tamil performance was noticeably worse, particularly in terms of coherence and general quality. Overall, the study demonstrates that rubric-aligned prompting can support safer multilingual counter-narrative generation in zero-shot settings. The findings also highlight the limitations of multilingual generation in low-resource languages such as Tamil. Future work can explore fine-tuning, retrieval augmentation, and language-specific adaptation to improve response quality and coherence.

References

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *In-*

[1] https://github.com/Prisur2013/RESPECTNLP_1991

ternational Journal of Data Science and Analytics, 18(1):49–68.

Koyel Ghosh, Saptarshi Saha, Thomas Mandl, and Sandip Modha. 2025. Findings from shared tasks on hate speech detection: Performance patterns for low-resource languages. *Pattern Recognition Letters*.

Koyel Ghosh and Apurbalal Senapati. 2025. Hate speech detection in low-resourced indian languages: An analysis of transformer-based monolingual and multilingual models with cross-lingual experiments. *Natural language processing*, 31(2):393–414.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, and Wenxian Wang. 2025. Rezg: Retrieval-augmented zero-shot counter narrative generation for hate speech. *Neurocomputing*, 620:129140.

Prasanna Kumar Kumaresan, Devendra Deepak Kayande, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2025. Homophobia and transphobia span identification in low-resource languages. *Natural Language Processing Journal*, page 100169.

Prasanna Kumar Kumaresan, Praveen Prasannan, Tanay Singh, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Saranya Rajiakodi, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2026. Findings of the Shared Task on Counter-Narrative Generation on Homophobic and Transphobic Comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual hate speech detection: a semi-supervised generative adversarial approach. *Entropy*, 26(4):344.

Praveen Prasannan, Prasanna Kumar Kumaresan, Saranya Rajiakodi, CN Subalalitha, and Bharathi Raja Chakravarthi. 2025. Counter-speech generation for homophobic and transphobic social media content in malayalam. *Social Network Analysis and Mining*, 15(1):87.

Hafiz Muhammad Raza Ur Rehman, Mahpara Saleem, Muhammad Zeeshan Jhandir, Eduardo Silva Alvarado, Helena Garay, and Imran Ashraf. 2025. Detecting hate in diversity: a survey of multilingual code-mixed image and video analysis. *Journal of Big Data*, 12(1):109.

Furqan Khan Saddozai, Sahar K Badri, Daniyal Alghazzawi, Asad Khattak, and Muhammad Zubair Asghar. 2025. Multimodal hate speech detection: a novel deep learning framework for multilingual text and images. *PeerJ Computer Science*, 11:e2801.

Tung Vu, Lam Nguyen, and Quynh Dao. 2025. Promptguard: An orchestrated prompting framework for principled synthetic text generation for vulnerable populations using llms with enhanced safety, fairness, and controllability. *arXiv preprint arXiv:2509.08910*.

Sahil Wadhwa, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, and Divya Chaudhary. 2025. Northeastern uni at multilingual counterspeech generation: Enhancing counter speech generation with llm alignment through direct preference optimization. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 19–28.

Brian Wilk, Homaira Huda Shomee, Suman Kalyan Maity, and Sourav Medya. 2025. Fact-based counter narrative generation to combat hate speech. In *Proceedings of the ACM on Web Conference 2025*, pages 3354–3365.