DLRG@LT-EDI 2026: Automating Counter-Narratives for Homophobic and Transphobic Comments

Ramesh Kannan R and Ratnavel Rajalakshmi
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, India.
Corresponding Author: rajalakshmi.r@vit.ac.in

Abstract

Online hate speech is spreading rapidly, creating significant challenge, particularly in low-resource language such as Tamil. Lack of developed automated content moderation systems makes it difficult to control harmful content effectively. In this study, we propose a computational framework for generating Counter Narratives (CNs) using classical NLP techniques. With this, we leverage TF-IDF features with n-grams to identify the labels as Homophobic or Transphobic. Span detection is performed with TF-IDF features with n-grams and Machine learning models. Counter narratives are then retrieved by computing cosine similarity, ensuring semantic alignment and contextual relevance. Evaluation on the expanded human curated dataset demonstrates that our approach produces contextually appropriate and semantically coherent counter narratives. Notably, the proposed system is submitted at Task 2 shown a overall average score of 80.40 % for Tamil and 77.29 % for English and secured **first** and fourth rank respectively. GitHub: https://github.com/kannanrrk/Span-Counter-Feature-Based

1 Introduction

In today's globalized world, issues surrounding homophobia and transphobia remain pervasive and they are often amplified in societies where linguistic resources, are scarce. Low-resource languages spoken by smaller populations with limited access to technological or academic tools present unique challenges in addressing these issues (Manukonda and Kodali, 2024). In many of these communities, negative stereotypes and discriminatory practices against individuals based on their sexual orientation and gender identity are not only culturally ingrained but are also reinforced by the lack of linguistic tools to challenge these social norms (Soled et al., 2022). A counter narrative approach seeks to disrupt these harmful narratives by offering alternative perspectives that empower marginalized groups. It involves challenging the stereotypical portrayals of LGBTQ+ individuals and providing the linguistic and social tools to address issues of homophobia and transphobia (Prasannan et al., 2025).

In low-resource language contexts, the development and use of inclusive language, community driven dialogue are crucial to overcome the barriers posed by limited vocabulary and societal prejudice (Hedderich et al., 2021). The creation of counter narratives not only fosters inclusivity. It also encourages the reshaping of cultural and societal norms, demonstrating that language can be a powerful agent for social change (Zhu and Bhat, 2021). This speech aims to explore the importance of counter narratives in advocating for LGBTQ+ rights in low-resource languages. Such approaches are pivotal in challenging harmful stereotypes and offering alternative perspectives that empower marginalized groups (Schrading et al., 2015). In particular, language revitalization and linguistic inclusivity play essential roles in combating homophobia and transphobia, as these strategies enable more effective counterspeech and foster a sense of empowerment among underserved communities (Chhaya et al., 2024; May, 2012). Furthermore, the work on linguistic human rights highlights how access to language resources is crucial for the survival and well being of minority groups, supporting the idea that promoting linguistic inclusivity can help address issues of discrimination and exclusion (May, 2012). The organisers conducted shared task on Counter Narrative Generation for Homophobia and Transphobia in Tamil and English as part of DravidianLangTech@LT-EDI 2026. This paper is structured as follows: We first present related works, then describe the dataset analysis, followed by evaluation metrics. Next, we have our proposed methodology, results & discussion, conclusion and limitations.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 161–166

2 Related Works

The issue of combating homophobia and transphobia in low-resource language communities remains a complex challenge. Many languages in these contexts lack sufficient resources for natural language processing (NLP) tasks, leading to difficulties in developing effective models for hate speech detection or counter narratives. While advances in NLP and machine learning (ML) have been made, low-resource languages often lack sufficient labeled data and existing models frequently fail to capture the cultural and social nuances of homophobic and transphobic rhetoric. To address this, several studies have proposed novel methods and techniques to create more inclusive, accurate and culturally sensitive models for low-resource languages.

(Chung et al., 2021) proposed a counter narrative generation framework based on Generative Pre-trained Transformers (GPT) for mitigating abusive content. Their approach formulates counter narrative generation as a conditional text generation task, where the model is trained to produce responses that counteract hateful input while preserving semantic coherence. The authors utilized the CONAN dataset, a benchmark corpus comprising 6,645 English Hate Speech–Counter Narrative (HS–CN) pairs, to fine-tune the model. This dataset enables supervised learning for generating contextually relevant counter narratives aimed at shifting user perspectives and mitigating harmful stereotypes through controlled language generation.

Developing an AI model to detect homophobic and transphobic speech involves addressing ethical concerns like fairness and bias(Wang et al., 2025). The challenge is to build a system that works across multiple languages, including low-resource ones, while ensuring cultural sensitivity. This model aims to be inclusive and accurate in identifying harmful language worldwide with different languages(Mnassri et al., 2024). In another approach, (Singh et al., 2023) developed a multilingual hate speech detection model designed specifically to detect hate speeches across different languages, including low-resource languages. Their methodology involved transfer learning, applying a pre-trained BERT model and fine-tuning it with a small dataset of hate speech collected from various social media platforms. The model obtained 87.7 %, with an F1-score on multilingual setting. With this the limitations includes irrelevant content, slang and ambiguous phrases potentially leading to misclassifications.

Similarly, (Usman et al., 2025) proposed an LLM based hate speech model for detecting and mitigating offensive speech in low-resource language urdu. For urdu, GPT 3.0 performs better than XLM-R model on three different languages. (Chakraborty et al., 2025) introduced a graph-based approach for identifying homophobic and transphobic narratives in low-resource languages. Their model employed neural graph attention networks (NGAN) to map relationships between words and phrases commonly used in discriminatory discourse. The limitation of this model is its computational intensity. Neural graph attention networks require substantial processing power, which can be a limitation in low-resource settings where access to high-performance computing may be limited.

(Chakraborty et al., 2025) integrated various transformer based embeddings with Relational Graph Convolution Networks to enhance the performance ensemble based voting classifier is utilised. The proposed method shown 0.98 F1 score on hindi dataset. For the coarse grain categories achieved relatively better performance on Hate, Offensive and Defamation categories. (Hashmi et al., 2025) explored a cross-lingual approach for hate speech detection in low-resource languages using intra-lingual and cross-lingual. They employed a meta learning approach with attention mechanism to boost the performance. Combination of transformer model and sequence model with few shot method achieved a F1 score of 79 % and 90 % in Norwegian and English. A key limitation is that translation errors or misinterpretations may propagate into the model, leading to less accurate predictions, especially in cases where idiomatic expressions or cultural references (Nozza, 2021) are lost in translation (Firmino et al., 2024).

In summary, existing studies demonstrate that machine learning and transformer-based approaches have achieved promising results in detecting homophobic and transphobic speech in low-resource languages. However, limitations such as data sparsity, model bias, scalability constraints, span-level detection issues, and translation errors remain unresolved. These challenges highlight the need for more robust and culturally aware models capable of handling linguistic diversity effectively.

3 Dataset

The organisers (Prasannan et al., 2025; Kumaresan et al., 2025; Chakravarthi, 2024) released dataset on Counter Narrative Generation for Homophobia and Transphobia in Tamil and English as part of DravidianLangTech@LT-EDI 2026.The dataset is split into training and testing sets for each language. From Table 1, the Tamil dataset contains training set of 800 (88.00 %) instances with 342 labeled as homophobic (42.75 %) and 458 labeled as transphobic (57.25 %). The testing set contains 109 (12.00 %) instances, where 73 are homophobic (67.89 %) and 36 are transphobic (32.11 %). From Table 1, English dataset includes 1800 (96.5%) instances with 1044 labeled as homophobic (58.00 %) and 756 labeled as transphobic (42 %). The test set consists of 66 (3.50 %) instances, where 49 are homophobic (74.24 %) and 17 are transphobic (25.76 %). Dataset analysis reveals a strong imbalance in the English split, with 96.5 % of the data in the training set and only 3.5 % in the test set. A similar imbalance is observed in the Tamil dataset, where 88 % of the data is allocated to the training set and only 12 % to the test set.

Category	Tamil		English	
	Train	Test	Train	Test
Homophobia	342	73	1044	49
Transphobia	458	36	756	17
Total	**800**	**109**	**1800**	**66**

Table 1: Dataset Statistics

3.1 Evaluation Metrics

For Counter Narrative Generation, the evaluation of generated counter narratives was based on both reference-based metrics and rubric-based evaluation scores. BERTScore was used to measure the semantic similarity between the generated and reference counter narratives, where higher scores indicates better alignment. Distinct-2 (diversity metric) measures only the diversity of the generated responses. In addition, rubric-based evaluation included the Politeness and Respectful Score (PRS), which evaluated the tone of the counter narrative towards individuals expressing homophobia and transphobia. Contextual Counter Narrative Coherence (CCNC) evaluated the relevance and coherence of responses to harmful speech.The Quality Score (QS) measured grammatical correctness and linguistic richness. These scores were converted into percentages, averaged to determine the final score. The final score reflects both the semantic accuracy and contextual relevance of their counter narratives, as well as their overall quality and respectfulness.

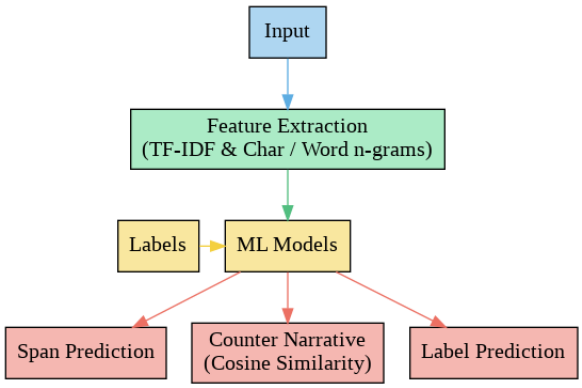

Figure 1: Overall Architecture of the proposed work

4 Methodology

In this study, we addressed the classification of homophobic and transphobic content, span detection and counter narrative generation using machine learning techniques. For the Tamil dataset, features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF) with character n-grams (3-5 grams) and Multinomial Naive Bayes (MNB) was applied for classification as shown in the architecture in Figure 1. For span detection, the same n-gram features were utilized with Logistic Regression to identify harmful spans in the text. For counter narrative generation, cosine similarity was employed to retrieve the top K (K=3) relevant features and generated appropriate counter narratives. Similarly, for the English dataset, character n-grams were replaced with word n-grams (1-2 grams) along with stop word removal during feature extraction, while Linear Support Vector Classifier (LSVC) was used instead of MNB for classification. Overall, the proposed methodology effectively integrates feature extraction, classification and counter narrative generation to identify and mitigate harmful speech while encouraging respectful discourse.

5 Results and Discussion

We constructed two different models, Tamil MNB (Multinomial Naive Bayes) and English L-

Team	Reference Based score		Rubric-Based Score				
	Diversity (Distinct-2)	Semantic Similarity (BERTScore)	PRS	QS	CCNC	Overall Avg.	Rank
DLRG	27.30	85.73	100	97.71	91.28	**80.40**	1
Amritha	20.89	85.27	100	100	89.45	79.12	2
NEUNI	19.16	85.09	95.41	86.24	92.66	75.71	3

Table 2: Task 2: Counter narrative Generation Results (in %) - Tamil

Team	Reference Based score		Rubric-Based Score				
	Diversity (Distinct-2)	Semantic Similarity (BERTScore)	PRS	QS	CCNC	Overall Avg.	Rank
Team_V	73.56	88.78	90.91	90.15	93.94	87.47	1
SigJBS	69.32	86.66	93.18	90.91	91.67	86.35	2
NEUNI	64.50	86.29	91.67	86.36	86.36	83.04	3
DLRG	74.36	85.55	72.73	69.70	84.09	**77.29**	4

Table 3: Task 2: Counter narrative Generation Results (in %) - English

SVC (Linear Support Vector Classifier) on the tasks of counter narrative generation, homophobic/transphobic classification and span detection. From Table 2, Tamil MNB performs well with semantic similarity, producing higher BERTScore of 85.73 %, but it was highly diverse and produced more lexically varied responses. However, From Table 3, English L-SVC showed much higher semantic alignment with reference counter narratives, achieved a higher BertScore of 85.55 %, indicating that it is better matched with the reference responses. Despite greater lexical diversity in Tamil MNB, the semantic alignment was satisfactory. On the other hand, English L-SVC performed better in matching reference counter narratives semantically but was less diverse in its responses. Distinct-2 score of English is performing better than Tamil Distinct-2 score.

Regarding rubric-based evaluation, the Tamil MNB model slightly outperformed the English L-SVC model in terms of politeness, coherence, and quality. The Tamil MNB consistently generated polite, coherent, and contextually relevant responses, resulting in higher overall rubric scores. However, the model exhibits semantic limitations due to its probabilistic nature, as it primarily relies on character-level frequency patterns and does not effectively capture broader contextual dependencies. A key observation from this comparison is that Tamil homophobia/transphobia detection with counter narrative generation produces polite and coherent outputs by leveraging frequent character patterns. However, Tamil requires sub-word-level feature representations to better address data sparsity and morphological complexity. In contrast, word-level features perform effectively for English in both hate speech detection and counter narrative generation, as large corpora sufficiently capture lexical variations. While the Tamil MNB model produces safe and human-like responses, its semantic understanding remains shallow, making it more suitable for low-resource settings where fluency and politeness are prioritized. On the other hand, the English L-SVC model demonstrates stronger semantic accuracy, making it more effective for high-resource languages where precise alignment with reference content is essential.

6 Conclusion

This paper presents a computational framework for generating Counter Narratives (CNs) to combat online hate speech in low-resource language like Tamil. By leveraging classical NLP techniques such as TF-IDF features and character n-grams for label identification and word n-grams for span detection, the proposed approach effectively detects hate speech and identifies the most relevant portions of text for targeted intervention. The utilization of cosine similarity ensures the generated narratives are both semantically aligned and contextually appropriate. These findings underscore the potential of classical feature-based models, combined with similarity driven retrieval for addressing the challenge of hate speech and generating effective counter narratives in low-resource language environments.

Limitations

Despite the strong performance of the proposed framework, several limitations exists. The use of TF-IDF n-grams may not fully capture deeper semantic nuances, particularly in Tamil, where morphologically complex and idiomatic expressions can pose challenges. Additionally, while cosine similarity ensures alignment with reference counter narratives, it may limit the diversity and creativity of the generated responses. Nevertheless, the approach offers a solid foundation for future improvements with more advanced NLP techniques.

References

Angana Chakraborty, Subhankar Joardar, Dilip K. Prasad, and Arif Ahmed Sekh. 2025. Graph-based hostile content detection in hindi language. *Discover Computing*, 28:264. Proposes a graph neural network approach combining contextual and semantic features for hostile/hate content detection in a low-resource language (Hindi), highlighting computational considerations of graph-based methods.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Bhargav Chhaya, Prasanna Kumar Kumaresan, Rahul Ponnusamy, and Bharathi Raja Chakravarthi. 2024. Homophobia and transphobia span identification in low-resource languages. *Research on Language and Computation*.

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*, pages 3846–3858. Association for Computational Linguistics. Introduces counter-narrative generation with transformers using curated contextual knowledge.

Anderson Almeida Firmino, Cláudio de Souza Baptista, and Anselmo Cardoso de Paiva. 2024. Improving hate speech detection using cross-lingual learning. *Expert Systems with Applications*, 235:121115.

Ehtesham Hashmi, Sule Yildirim Yayilgan, and Mohamed Abomhara. 2025. Metalinguist: Enhancing hate speech detection with cross-lingual meta-learning. *Complex & Intelligent Systems*, 11:179. Introduces a meta-learning based cross-lingual framework that combines attention mechanisms and multilingual representation learning for robust hate speech detection across languages.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2545–2568.

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.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Stephen May. 2012. *Linguistic Human Rights: Overcoming Linguistic Discrimination*. Routledge, New York, NY.

Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual hate speech detection: A semisupervised generative adversarial approach. *Entropy*, 26(4):344. Multilingual semisupervised approach using pretrained transformers to handle data scarcity and cross-lingual detection.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Nathan Schrading, Hemant Purohit, and Amit Sheth. 2015. Analysis of persuasion techniques in tweets for public policy debates. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1206–1213. IEEE.

Pardeep Singh, Nitin Kumar Singh, Monika, and Satish Chand. 2023. mbert-gru multilingual deep learning framework for hate speech detection in social media. *Journal of Intelligent & Fuzzy Systems*, 45(5):8177–8192. Proposes a multilingual BERT (mBERT) + GRU model fine-tuned for hate speech classification across multiple languages.

Kodiak R. S. Soled, Kristen D. Clark, Molly R. Altman, Jordon D. Bosse, Roy A. Thompson, Allison Squires, and Athena D. F. Sherman. 2022. Changing language, changes lives: Learning the lexicon of lgbtq+ health equity. *Research in Nursing & Health*, 45(6):621–632.

Muhammad Usman, Muhammad Ahmad, Irina Gelbukh, Grigori Sidorov, and Rolando Quintero Tellez. 2025. A large language model-based approach for multilingual hate speech detection on social media. *Computers*, 14(7):279. Hybrid framework using transformer embeddings and statistical features for multilingual hate speech detection across English, Spanish, and Urdu.

Yifan Wang, Mayank Jobanputra, Ji-Ung Lee, Soyoung Oh, Isabel Valera, and Vera Demberg. 2025. Bridging fairness and explainability: Can input-based explanations promote fairness in hate speech detection? *arXiv preprint arXiv:2509.22291*. Systematic study of bias, fairness, and explainability in NLP hate speech detection models.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.