Susmitha@LT-EDI 2026: Detecting LGBTQ+ Phobia in Multilingual Memes via Joint Representation

Susmitha Jaishri[1] **, Kogilavani Shanmugavadivel**[2]**,**
Malliga Subramaniyan[3]**, Mouleeshuwarapprabu R**[4]
[1]Department of CSE, NITTTR Chennai
[2] Professor, Department of CSE, NITTTR Chennai
[3] Professor,Department of CSE,Kongu Engineering College, Erode
[4]Assistant Professor(SLG),Department of EIE, Kongu Engineering College, Erode
susmithajaishri@gmail.com

Abstract

Automated detection of LGBTQ+ phobia is critical for digital safety. We participated in the LT-EDI@ACL 2026 shared task for multimodal meme classification across English, Hindi, and Chinese. Our methodology leverages a late-fusion multimodal architecture combining XLM-RoBERTa for textual features and ResNet-50 for visual representation. By utilizing weighted cross-entropy to address extreme class imbalance, we achieved Rank 3 in Chinese (F1: 0.7371), Rank 4 in English (F1: 0.6121), and Rank 7 in Hindi (F1: 0.1616). This paper details our system description and provides a critical analysis of the "Hindi Failure," where class imbalance led to a lack of model convergence.

1 Introduction

The proliferation of memes presents a sophisticated multimodal challenge for moderation. Memes frequently utilize Benign Confounders, where harmless textual and visual components convey phobic intent only when combined (Ponnusamy et al., 2026). Detecting LGBTQ+ phobia in this format requires deep cultural nuance and cross-modal context beyond standard text-based methods.

This work addresses homophobia and transphobia detection in English, Hindi, and Chinese to protect digital communities across diverse linguistic landscapes. We propose a gated late-fusion architecture integrating XLM-RoBERTa textual representations with ResNet-50 visual features. By treating this as a joint representation task, we capture the critical interplay between imagery and text. Furthermore, we evaluate weighted loss functions to mitigate extreme class imbalances. Our findings contribute to Dravidian and Indo-Aryan language research by identifying specific low-resource bottlenecks, categorized as the "Hindi Failure" in our analysis.

2 Problem Description

The objective of this task is to categorize anti-LGBTQ+ content in multimodal memes, specifically addressing "benign confounders" where modalities appear neutral in isolation but convey phobic intent when synthesized. Classification involves three mutually exclusive categories: Homophobic (H), expressing prejudice toward sexual orientation; Transphobic (T), targeting gender identity; and Non-anti-LGBTQ+ (N), for benign content. The primary challenge is developing a joint representation to distinguish these classes when intent is implicit and context-dependent.

3 Literature Review

Hate speech detection has evolved toward multimodal reasoning to incorporate visual context in memes. Ponnusamy et al. (2026) established the current benchmarks for identifying homophobia and transphobia, emphasizing Dravidian language challenges. Foundational work by Chakravarthi (2024) defined the initial parameters for phobia detection in social media.

Modern methodologies emphasize joint representations to capture nuanced hate. Hande et al. (2021) demonstrated that cross-lingual transformers are essential for code-mixed text, while Kiela et al. (2020) highlighted how visual extractors like ResNet-50 identify "benign confounders." This synergy is critical for robust generalization in the LT-EDI @ ACL 2026 tasks.

4 Methodology

Our framework utilizes a dual-stream late-fusion architecture designed to target "benign confounders," where phobic context is emergent only through multimodal synthesis.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 222–225

4.1 Feature Extraction and Encoding

The textual stream (T) employs **XLM-RoBERTa** to handle the morphologically rich scripts of the Hindi and Chinese datasets via sub-word tokenization. This produces a semantic vector $h_t \in \mathbb{R}^{768}$. Simultaneously, the visual stream (I) utilizes a **ResNet-50** backbone to extract spatial hierarchies and symbolic imagery from meme images, generating a feature vector h_v. Both are mapped into a joint 512-dimensional space via linear projections:

$$t_{feat} = W_{text}h_t + b_t, \quad v_{feat} = W_{vis}h_v + b_v \quad (1)$$

where $t_{feat}, v_{feat} \in \mathbb{R}^{512}$.

4.2 Gated Multimodal Fusion

To capture cross-modal dependencies, we implement a gating mechanism. A gate vector g is computed to weight the importance of each modality based on their combined features:

$$g = \sigma(W_g[t_{feat} \oplus v_{feat}] + b_g) \quad (2)$$

The final representation h_{fused} is calculated via element-wise multiplication ($\odot$) and processed by a Multi-Layer Perceptron (MLP) with Softmax activation for classification:

$$h_{fused} = g \odot t_{feat} + (1 - g) \odot v_{feat} \quad (3)$$

4.3 Training and Optimization

We address extreme class imbalance using **Weighted Cross-Entropy (WCE)** loss. This penalizes minority class errors more heavily using weights α_i inversely proportional to class frequency:

$$\mathcal{L}_{WCE} = -\sum_{i=1}^{M} \alpha_i y_i \log(\hat{y}_i) \quad (4)$$

The model is optimized using **AdamW** for 3 epochs with a batch size of 8 and a learning rate of 10^{-5}. To prioritize phobic classes, loss weights were set to [1.5, 2.0, 1.0] for the Homophobic, Transphobic, and Non-anti-LGBTQ+ categories, respectively.

5 Proposed Workflow

The operational pipeline of the proposed framework handles the unique challenges of multimodal sentiment analysis in low-resource languages. The process, illustrated in Figure 1, begins with the parallel feature extraction of paired meme inputs through specialized encoding mechanisms.

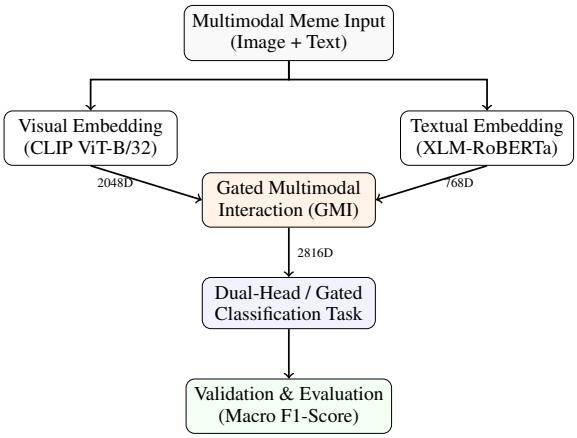

Figure 1: Proposed Gated Multimodal Framework for Meme Classification.

Following the encoding phase, the independent feature sets are integrated into a joint representation space. This unified vector is processed by the classification module, where target labels are predicted and validated using the Macro F1-score as the primary performance metric. The study utilizes the LT-EDI 2026 dataset, targeting the "benign confounder" problem across English, Hindi, and Chinese tracks. To address the inherent class imbalances typical of social media hate speech, the system is optimized via weighted cross-entropy, prioritizing the detection of minority phobic classes (H and T) over the majority non-anti-LGBTQ+ (N) class.

6 Dataset Statistics

The LT-EDI@ACL 2026 dataset provides a challenging multimodal environment with significant class imbalances across linguistic tracks. Table 1 details the distribution of the Homophobic (H), Transphobic (T), and Non-Anti-LGBTQ+ (N) classes for both training and testing phases.

Class	**English**		**Hindi**		**Chinese**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Homophobic	160	40	366	64	705	176
Transphobic	160	41	274	96	55	14
Non-Anti-LGBT	240	60	158	40	196	49
Total	**560**	**141**	**798**	**200**	**956**	**239**

Table 1: Class distribution across language tracks for training and testing sets.

The distribution highlights that while the En-

glish track is relatively balanced, the Hindi and Chinese tracks exhibit a sharp majority in the Homophobic category. This data sparsity for Transphobic instances in the Chinese track (only 55 training samples) and the overall skew in Hindi provides essential context for the weighted cross-entropy approach utilized in our methodology.

7 Experimental Analysis

The experimental results across the three linguistic tracks demonstrate varying levels of cross-modal alignment. Performance is evaluated using Macro F1-score, as summarized in Table 2.

Track	Rank	Macro F1	Primary Challenge
Chinese	3	0.7371	Multimodal Synergy
English	4	0.6121	Category Overlap
Hindi	7	0.1616	Class Collapse

Table 2: Official performance metrics and rankings.

7.1 English Track Analysis

The system secured the fourth rank in the English track. The confusion matrix in Figure 2 reveals that while the model effectively identifies "Non-anti-LGBTQ+" memes, it struggles to distinguish between Homophobia and Transphobia. A significant portion of Transphobia instances was misclassified as Homophobia, suggesting that the textual encoder captures general aggressive sentiment but lacks the granular linguistic markers needed for category-specific identification.

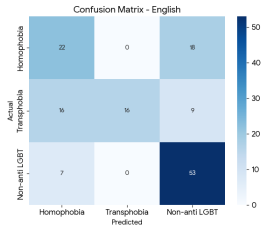

Figure 2: Confusion Matrix for the English Track.

7.2 Hindi Track: The Convergence Failure

Our submission achieved the seventh rank but exhibited a 100% majority class bias towards "Homophobic," as visualized in Figure 3. The model failed to identify any instances of Transphobia or Non-anti-LGBTQ+ classes. This collapse indicates a failure to learn features from script-mixed data or an inability of the weighted loss to overcome the extreme data imbalance. This "Hindi Failure" highlights a critical limitation in late-fusion architectures when applied to low-resource, high-imbalance scenarios.

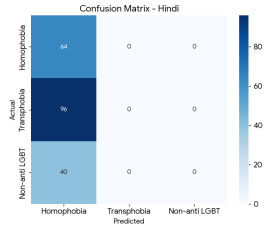

Figure 3: Confusion Matrix for the Hindi Track showing class collapse.

7.3 Chinese Track Analysis

The system was most successful in the Chinese track, securing the third rank. The model effectively aligned Hanzi script features with visual cues. As shown in the confusion matrix in Figure 4, the Chinese track exhibited the highest diagonal density. The synergy between ResNet features and the Hanzi textual features provided the most reliable classification, specifically in isolating the Non-anti-LGBTQ+ class with minimal false positives.

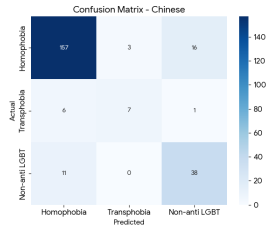

Figure 4: Confusion Matrix for the Chinese Track.

8 Conclusion

Our gated multimodal framework demonstrates that while late-fusion is effective for English and Chinese (Rank 3), it suffers from a "Hindi Failure" caused by extreme class imbalance and script-mixing. The system currently lacks cross-modal attention, limiting its ability to resolve culturally specific metaphors or masked sarcasm.

Future work will replace the fixed weighted loss with dynamic sampling and investigate pre-trained Vision-Language Models (VLMs) to improve zero-shot generalization in low-resource, imbalanced multilingual environments.

Ethics Statement

Data was handled per LT-EDI @ ACL 2026 guidelines. While our model assists in mitigating online harm, users should remain aware of potential algorithmic biases; therefore, this system is intended to support, not replace, human moderation.

AI Disclosure

The authors declare that no generative Artificial Intelligence (AI) or AI-assisted technologies were utilized in the writing, data analysis, or development of the methodologies presented in this paper. All content is the original work of the authors, produced in accordance with the ethical guidelines of LT-EDI @ ACL 2026.We thank the organizers of the DravidianLangTech workshop for providing the datasets and the platform for this shared task.

Code Availability

The source code is publicly available at: https://github.com/susmithajaishri/homophobia_transphobia_meme_classification

References

Judith Jeyafreeda Andrew. 2023. Judithjeyafreeda@lt-edi-2023: Using GPT model for recognition of Homophobia/Transphobia detection from social media. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI 2023)*, pages 88–93. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in YouTube comments. In *Social Network Analysis and Mining*, volume 12, pages 82–95. Springer Science and Business Media LLC.

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

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv pre-print*, abs/2108.12177.

M. Jaganath, J. Ramya, K. Sangeetha, P. Nithya, and V. A. Subashini. 2026. Memecheck: Automated meme analysis for identifying offensive text and visuals. In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25)*, volume 4, pages 676–682. SCITEPRESS – Science and Technology Publications, Lda.

S. F. Karim, M. Rahman, and S. Islam. 2025. CUET_Blitz_Aces@lt-edi-2025: Leveraging transformer ensembles and majority voting for hate speech detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2025)*, pages 142–148. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 1–12.

C. Jerin Mahibha. 2026. A few-shot approach to classify hate speech based on severity from texts in Dravidian languages. *PeerJ Computer Science*, 12:e3711.

Kishore Kumar Ponnusamy, Bharathi Raja Chakravarthi, Mahesh Susaladi, Junru Ren, Prasanna Kumar Kumaresan, Premjith B, Durairaj Thenmozhi, Ruba Priyadharshini, and Subalalitha Chinnaudayar Navaneethakrishnan. 2026. Overview of Homophobia and Transphobia Meme Classification Shared Task. In *Proceedings of the Workshop on Language Technology for Equality, Diversity, and Inclusion*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Naveenram C. E, Vishal Rs, and Srinesh S. 2025. KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 232–236. Association for Computational Linguistics.