Team_One@DravidianLangTech 2026: A Gated Multimodal Architecture for Multi-Level Stance and Target Detection in Malayalam Political Memes

Ashmi S N[1], **Nimisha M Iyer**[1], **Balasubramanian Palani**[1]
Jobin Jose[1], **Siranjeevi Rajamanickam**[2]
[1]Indian Institute of Information Technology Kottayam, Kerala, India
[2]Govt. Polytechnic College-Thuvakudi, Tiruchirappalli, Tamil Nadu, India
ashmi23bcs196@iiitkottayam.ac.in, nimisha23bcd4@iiitkottayam.ac.in
pbala@iiitkottayam.ac.in, jobin@iiitkottayam.ac.in, rajasiranjeevi@gmail.com

Abstract

Stance and target detection in multimodal political memes presents notable challenges in low-resource and highly imbalanced settings. This task is based on the Malayalam dataset from the DravidianLangTech 2026 Shared Task (500 samples with a 95.4:4.6 stance imbalance). The primary challenges stem from linguistic variability and visually complex meme formats, which hinder accurate text extraction and effective multimodal alignment. A lightweight yet high-performing multimodal framework is proposed that integrates bilingual OCR, a Vision Transformer (ViT), and IndicBERT to learn complementary visual and textual representations. A gated fusion mechanism effectively combines multimodal features, while asymmetric loss weighting and post-training threshold optimization address extreme class imbalance. The methodology achieves a Weighted F1-score of 0.9535 for stance detection and 0.5283 for target identification, demonstrating strong robustness and generalization under realistic multimodal constraints. The training pipeline is publicly available for reproducibility.[1]

1 Introduction

Political internet memes generally depend on the intricate combination of visual signals and accompanying text that are often laden with local sarcasm, code-mixed expressions, and slang. The DravidianLangTech 2026 Shared Task (Rajiakodi et al., 2026) challenges systems to automate this understanding across two hierarchical levels: Level 1 (Stance Detection), a binary classification of Support vs. Oppose; and Level 2 (Target Identification), a 5-class fine-grained classification identifying the specific target.

This task is severely hampered by an ultra-low-resource training environment and an extreme lack of minority class examples. Standard multilingual models and simplistic late-fusion techniques frequently fail to capture the semantic interplay between modalities under such scarcity. In this paper, a comprehensive pipeline tailored for Malayalam is detailed. By combining bilingual Tesseract Optical Character Recogonition (OCR) (Smith, 2007), along with the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2021), leveraging IndicBERT (Kakwani et al., 2020) for multilingual contextual encoding, and introduce a mathematically robust Gated Attention fusion module inspired by attention gating mechanisms (Ilse et al., 2018). This architecture performs simultaneous multi-level classification, effectively isolating rare minority classes without sacrificing majority-class precision.

2 Related Work

Multimodal Meme Analysis While models like CLIP (Radford et al., 2021) revolutionized multimodal alignment, their English-centric training limits regional applicability. Processing localized memes requires robust OCR. Because OCR extractions from stylized memes are inherently noisy, specialized fusion techniques are required to dynamically process garbled text based on visual context (Suryawanshi et al., 2020). Traditional architectures relied on CNNs (He et al., 2016), but Vision Transformers (Dosovitskiy et al., 2021) now allow networks to capture global spatial relationships via self-attention from the first layer.

Dravidian NLP & Encoders Code-mixing on social media complicates standard NLP pipelines (Chakravarthi et al., 2020). Three pre-trained architectures were benchmarked: XLM-RoBERTa ($\approx$ 270M parameters) (Conneau et al., 2020), MuRIL ($\approx$ 236M parameters) (Khanuja et al., 2021), and IndicBERT ($\approx$ 33M parameters) (Kakwani et al., 2020). Built on ALBERT (Lan et al., 2020), In-

[1]Code available at: https://github.com/ashmisn/dravidlangtech

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 409–413

dicBERT employs cross-layer parameter sharing that acts as a structural regularizer against overfitting, making it well-suited for agglutinative Malayalam text.

3 Proposed Methodology

3.1 OCR and Data Augmentation

Text extracted via bilingual Tesseract OCR (`mal+eng`) is regex-filtered to retain only Malayalam Unicode and English alphanumerics, minimizing out-of-vocabulary tokens to yield sequence S. To combat data scarcity, standard image augmentations (flipping, rotation, brightness) are applied. The complete architecture is illustrated in Figure 1.

3.2 Textual Encoding via IndicBERT

Sequence S is encoded using IndicBERT. By utilizing cross-layer parameter sharing, it reduces its footprint to $\approx$ 33M parameters, acting as a natural regularizer against overfitting. It employs standard multi-head self-attention, projecting inputs into Query (Q), Key (K), and Value (V) matrices:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the key dimensionality. This yields a dense contextualized representation $H_{text} \in \mathbb{R}^{768}$. For multimodal fusion, H_{text} is projected into a shared latent space ($d_{proj} = 512$) via linear mapping, Layer Normalization (LN), and ReLU:

$$z_{text} = \text{ReLU}(\text{LN}(W_t H_{text} + b_t)) \quad (2)$$

where $W_t \in \mathbb{R}^{512\times768}$ and $b_t \in \mathbb{R}^{512}$ are learnable projection parameters.

3.3 Visual Encoding via ViT

Image I is processed by a pre-trained Vision Transformer (ViT-Base). ViT divides the image into 16×16 patches and applies the identical self-attention mechanism (Equation 1). This captures global spatial dependencies directly from the `[CLS]` token, outputting visual representation $H_{vis} \in \mathbb{R}^{768}$. Similar to the textual pathway, H_{vis} is identically projected into the shared latent space:

$$z_{vis} = \text{ReLU}(\text{LN}(W_v H_{vis} + b_v)) \quad (3)$$

where $W_v \in \mathbb{R}^{512\times768}$ and $b_v \in \mathbb{R}^{512}$ are the visual projection parameters.

3.4 Gated Attention Fusion

Simple concatenation propagates OCR noise directly into the visual context (Tsai et al., 2019). To enable adaptive filtering, our architecture leverages the structurally aligned latent embeddings z_{vis} and z_{text}. A Sigmoid (σ) gate evaluates their concatenated features to learn a dimension-wise weight vector $\alpha \in (0, 1)$:

$$\alpha = \sigma(W_g[z_{vis} \oplus z_{text}] + b_g) \quad (4)$$

The final joint representation dynamically balances the modalities via element-wise multiplication ($\odot$):

$$F = \alpha \odot z_{text} + (1 - \alpha) \odot z_{vis} \quad (5)$$

If unstructured OCR noise is detected, the network learns to drive $\alpha \to 0$, filtering the corrupted text and forcing reliance on robust visual cues.

3.5 Multi-Task and Asymmetric Learning

The fused vector F passes through a shared Fully Connected block with Dropout ($p = 0.4$) to yield a joint representation $R \in \mathbb{R}^{256}$. This shared latent space facilitates multi-task learning by transferring beneficial inductive biases between the stance and target classification heads. To address extreme class imbalance without relying on generative upsampling (which can distort compact embedding spaces), class-specific asymmetric weights are applied to the Cross-Entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} \lambda_i y_i \log(P(\hat{y}_i|I, S)) \quad (6)$$

Setting $\lambda_{minority} = 3.0$ and $\lambda_{majority} = 1.0$ artificially inflates the gradient magnitude for minority errors, heavily penalizing misclassifications on rare classes.

3.6 Post-Training Threshold Optimization

Standard binary classifiers default to a static decision boundary of $\tau = 0.50$. Under a severe 95.4:4.6 skew, continuous probability outputs inherently bias toward the prior distribution, often leading to majority-class collapse. Instead of modifying the data distribution, the decision logic is optimized. Post-training, validation thresholds $\tau \in [0.05, 0.50]$ are systematically swept to identify the optimal boundary τ^* that maximizes the primary evaluation metric (Macro F1):

$$\tau^* = \arg\max_{\tau} F1_{Macro}(\tau) \quad \text{s.t.} \quad TP > 0 \quad (7)$$

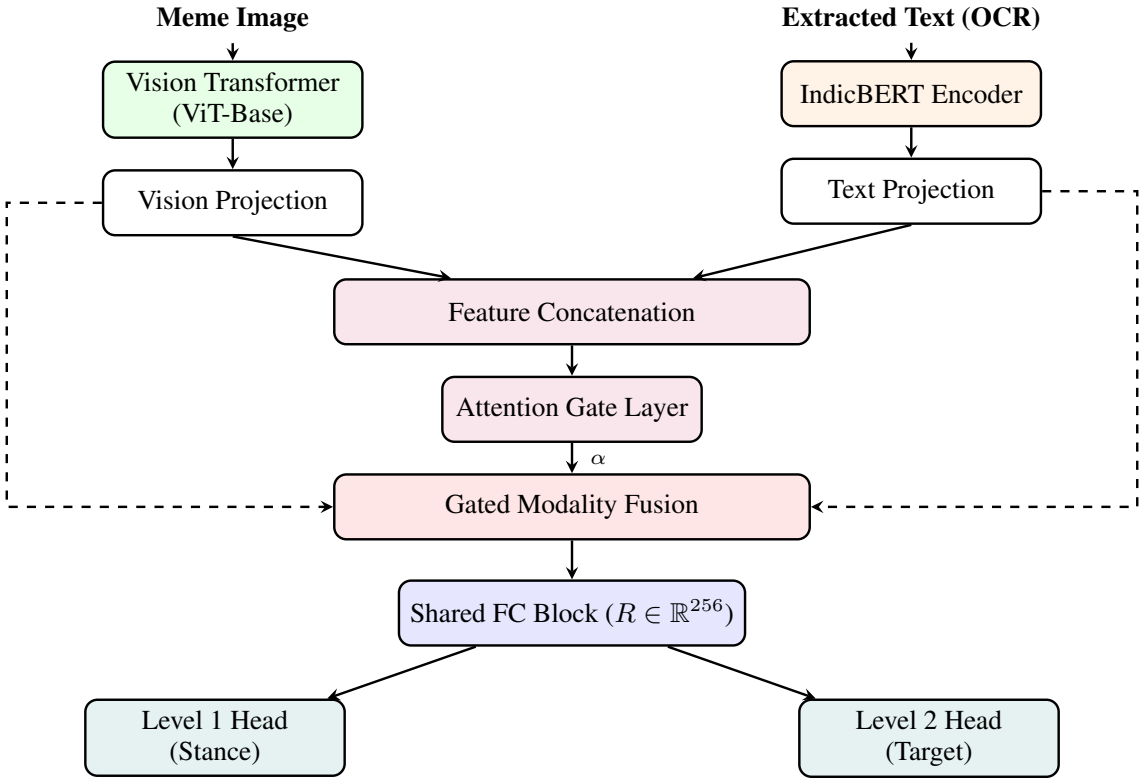

Figure 1: Gated Multimodal Architecture. The dynamic attention gate calculates optimal fusion weights before passing the unified representation to multi-level classification heads.

This calibration explicitly aligns the model's predictive boundary with the task's minority-sensitive objectives.

4 Experimental Setting

4.1 Implementation Setup

Experiments were conducted on an NVIDIA Tesla T4 GPU using PyTorch and HuggingFace Transformers. Training was performed for 12 epochs using the AdamW optimizer. A batch size of 8 was used due to memory constraints, which also introduced mild stochastic regularization.

To prevent catastrophic forgetting, a conservative learning rate of $1e^{-5}$ was applied to the pretrained ViT and IndicBERT encoders, while a larger rate of $1e^{-4}$ was assigned to the randomly initialized fusion layers to enable faster convergence. A decoupled weight decay of $1e^{-2}$ was used to further control overfitting.

4.2 Dataset Description

The dataset distribution is presented in Table 1.

4.3 Performance Metrics

To evaluate performance under severe class imbalance, Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) is used, where

Table 1: Class Distribution (Train vs. Test Set)

Level 1 (Stance)	**Train (n=500)**	**Test (n=100)**
TROLL / OPPOSE	477 (95.4%)	96 (96.0%)
Support/Praise	23 (4.6%)	4 (4.0%)
Level 2 (Target)	**Train (n=500)**	**Test (n=100)**
Against individual person	315 (63.0%)	50 (50.0%)
Against party	110 (22.0%)	31 (31.0%)
Intersection	53 (10.6%)	15 (15.0%)
Support for individual	12 (2.4%)	4 (4.0%)
Support for party	10 (2.0%)	0 (0.0%)

TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively. Macro F1 ($\frac{1}{C}\sum_{i=1}^{C} F1_i$) is prioritized, as it computes the unweighted mean of the C class-wise F1-scores and assigns equal importance to minority classes. Additionally, Weighted F1 ($\sum_{i=1}^{C} w_i F1_i$) scales the scores by the true proportion (w_i) of samples in each class.

5 Results and Discussion

5.1 Level 1 (Binary Classification): Stance Detection Results

Table 2 presents the ablation results. Although MuRIL + ViT achieves high accuracy, its low

Table 2: Performance of models on Binary Classification (Level 1)

Model	Macro F1	Weighted F1	Accuracy
XLM-R + ViT	0.5105	0.9210	93.5%
MuRIL + ViT	0.4898	0.9404	96.0%
IndicBERT + ResNet	0.5400	0.8700	86.2%
IndicBERT + ViT	**0.6564**	**0.9535**	**96.0%**

Table 3: Performance of models on Multiclass Classification (Level 2)

Model	Precision	Recall	F1-score
XLM-R + ViT	0.5010	0.4680	0.4805
MuRIL + ViT	0.4920	0.4520	0.4708
IndicBERT + ResNet	0.5180	0.4890	0.5012
IndicBERT + ViT	**0.5471**	**0.5500**	**0.5283**

Table 4: Class-wise Performance of IndicBERT+ViT (Level 2)

Target Class	Precision	Recall	F1-score
Against indv. person	0.5441	0.7400	0.6271
Against party	0.5417	0.4194	0.4727
Intersection	**0.7143**	0.3333	0.4545
Support for individual	0.0000	0.0000	0.0000
Weighted Average	**0.5471**	**0.5500**	**0.5283**

Macro F1 indicates inadequate minority class detection, a byproduct of the extreme data imbalance. IndicBERT + ResNet improves minority sensitivity slightly but suffers a drop in accuracy, suggesting weaker visual feature extraction.

The proposed IndicBERT + ViT model achieves the best balance. Reinforced by the asymmetric loss weighting and threshold calibration ($\tau^* = 0.10$), the model achieves a Macro F1-score of 0.6564. It successfully detects scarce "Support" memes without compromising the 96.0% majority class stability.

5.2 Level 2 (Multiclass classification): Target Identification Results

Table 3 compares the overall multiclass classification performance across different architectural combinations. The proposed IndicBERT + ViT fusion achieves the highest precision, recall, and F1-score, demonstrating the effectiveness of combining strong linguistic representations with robust visual feature extraction for fine-grained target identification.

The detailed breakdown in Table 4 reveals class-specific behavior. The model achieves its highest F1-score for the dominant "Against individual person" category, indicating stable learning for majority targets. The "Intersection" class records the highest precision (0.7143), suggesting that the gated fusion mechanism effectively captures overlapping target semantics. However, its comparatively lower recall reflects limited exposure during training. The "Support for individual" category records zero performance. This outcome primarily stems from extreme data sparsity, with only a small number of training instances and the absence of sufficient representative samples in the evaluation split. Such scarcity prevents the formation of reliable and generalizable decision boundaries.

Heavily stylized and code-mixed typography significantly degrades standard OCR extraction quality. Naive late-fusion approaches tend to propagate this textual noise into the final representation. In contrast, the gated attention mechanism adaptively filters unreliable signals. When high-entropy OCR outputs are detected, the sigmoid gate suppresses the corrupted textual modality and shifts reliance toward visual embeddings (z_{vis}), thereby preserving classification robustness despite severe OCR degradation.

6 Conclusion and Future Work

Processing code-mixed slang in ultra low-resource meme datasets remains a challenging task due to noisy text extraction and severe class imbalance. To address these challenges, this paper introduces a robust, gated multimodal architecture specifically tailored for multi-level stance and target detection in Malayalam political memes. Unlike existing approaches that predominantly cater to high-resource languages or rely on standard feature concatenation, our framework uniquely integrates advanced multimodal gating with imbalance-aware training. This ensures the model effectively captures the nuanced synergy between stylized code-mixed text and visual cues, providing a significant improvement over traditional baselines in low-resource settings.

Future work will focus on improving OCR quality for code-mixed and stylized meme text, particularly by fine-tuning OCR models on Malayalam-English slang and low-resolution meme images. Additionally, expanding annotated data for minority classes by collecting more diverse and representative samples will be crucial to strengthening model robustness and improving its applicability in real-world content moderation scenarios.

7 Limitations

OCR extraction remains a major limitation in Malayalam political memes due to highly stylized fonts, low-resolution images, and heavy Malayalam-English code-mixing. Although the proposed gated fusion mechanism reduces the impact of noisy OCR outputs, overall performance still depends significantly on the quality and completeness of extracted text.

8 Ethical Considerations

Automated moderation of subjective Malayalam sarcasm risks inadvertently flagging benign political discourse. Hence, this model is intended strictly for human-in-the-loop moderation, not autonomous censorship. Additionally, AI tools were used solely for proofreading; all methodologies and analyses remain entirely human-authored.

References

B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages*, pages 202–210.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

M. Ilse, J. M. Tomczak, and M. Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2127–2136. PMLR.

D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., P. Bhattacharyya, M. M. Khapra, and P. Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.

S. Khanuja, D. Dikshit, P. Talukdar, M. Hasan, A. Agrawal, and C. Carvallo. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

S. Rajiakodi, S. P. M. Chinnan, B. Premjith, C. Subalalitha, P. Rahul, K. Anshid, S. Bhuvaneshwari, V. Jananayagan, N. Ragavan, P. Santhini, and B. R. Chakravarthi. 2026. Shared task on multi-level political meme classification for dravidian languages. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, and P. Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 14–22.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. 2019. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*, 30.