CUET-2567@DravidianLangTech-ACL 2026: Multimodal Stance and Target Identification in Dravidian Political Memes

Arka Dutta, Anindya Majumder, Adnan Faisal, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2204025, u2204067, u2004002}@student.cuet.ac.bd,hasanmurad@cuet.ac.bd

Abstract

In Dravidian languages, political memes progressively shape public opinion and political discourse, influencing digital conversations and public narratives. Our paper proposes a multilevel multimodal framework for political meme classification in Tamil and Malayalam as part of the Multi Level Political Meme Classification-DravidianLangTech@ACL 2026 shared task. The task has involved two levels: Level 1 has identified whether a meme expresses Troll/Oppose or Support/Praise, while Level 2 has determined the specific target category (Individual, Party, or Intersection). We have evaluated unimodal and multimodal architectures to analyze the impact of textual and visual representation. Experimental results have highlighted the importance of a multimodal approach over unimodal approaches. This work confirms the effectiveness of combining image and text features in meme understanding. Among the evaluated models, the mBERT+ViT architecture has achieved the best overall performance across both languages and classification levels. According to the evaluation of shared task we achieved average F1 score of 0.72 securing the 2nd rank in Malayalam task and F1 score of 0.76 in Tamil task securing the 6th rank. However after our experimental evaluation we got best average F1 score of 0.62 for Tamil and 0.49 for Malayalam. Despite moderate results, challenges have remained mainly due to the dataset size, class imbalance, and noisy text extraction from images.

1 Introduction

As internet accessibility has increased, memes have become influential units of cultural transmission (Dawkins, 1976), often combining text and images to shape political discourse (Afridi et al., 2021). However, political meme classification—especially for Tamil and Malayalam—remains underexplored, highlighting the need for effective multimodal approaches (Galipeau, 2023).

The Multi-Level Political Meme Classification - DravidianLangTech@ACL 2026 Shared Task addresses the challenge of classifying political memes in Dravidian languages, requiring models to perform analysis across multiple levels, including sentiment identification (Support/Praise vs. Troll/Oppose) and fine-grained target categorization (Individual, Party, or Intersection

In our task on the Tamil meme dataset, we have explored multiple multimodal and unimodal combinations for multilevel classification. Among them, multilingual BERT (mBERT) for textual features combined with Vision Transformer (ViT) for visual features achieved the best performance. Specifically, Level 1 has achieved an F1-score of 0.68, and Level 2 has achieved an F1-score of 0.53. This combination has outperformed alternative setups, including CLIP + XLM-RoBERTa and IndicBERT + ResNet, indicating that mBERT captures textual nuances and ViT extracts complementary visual features, with the multimodal combination yielding more balanced performance than single-modality models. Similarly, on the Malayalam dataset, the fusion of BERT and ViT has outperformed combinations such as XLM-RoBERTa + CLIP and IndicBERT + ResNet, achieving an F1-score of 0.53 for Level 1 and 0.42 for Level 2. These results have demonstrated the effectiveness of multimodal fusion for multilevel political meme classification across languages. The core contributions of this work are summarized as follows:

- We have proposed a multilevel multimodal framework for Tamil and Malayalam political meme classification (Level 1 and Level 2).
- We have demonstrated that mBERT + ViT consistently have outperformed other multimodal combinations through systematic comparison.
- We have implemented a preprocessing pipeline that includes text transliteration and

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 196–200

image enhancement techniques to improve overall data quality

For a comprehensive guide on the implementation process and to access the complete codebase, please visit the GitHub repository: CUET-2567 Multi-Level Political Meme Classification

2 Related Work

Political meme classification has emerged as a significant research area, as memes inherently integrate textual and visual cues that require multimodal deep learning frameworks for comprehensive semantic interpretation.

Prior work highlights the superiority of multimodal approaches, including Chinese BERT+VGG16 for misogynistic memes (Faisal et al., 2025), the Hateful Memes benchmark exposing unimodal limitations (Kiela et al., 2020), ViT+BERT with self-attentive fusion for multitask classification (Hossain et al., 2026), and BERT+GGNN for effective sarcasm detection (Qin et al., 2025).

Vision–language models such as OSCAR+RF (Chen and Pan, 2022), transformer-based Tamil troll detection (Hariprasad et al., 2022), and Bangla BERT+ViT on the MAC dataset (Alam et al., 2024) further demonstrate the effectiveness of multimodal fusion over unimodal approaches in meme classification.Motivated by prior multimodal meme analysis studies, our system adopts a late-fusion transformer-based framework (mBERT+ViT) tailored to the multi-level political meme classification task in Tamil and Malayalam.

3 Data Description

The datasets consist of Tamil and Malayalam political memes, each containing an image and corresponding textual labels. The Tamil dataset includes 1003 memes (Train: 802, Validation: 100, Test: 101, Total: 1003), while the Malayalam dataset includes 600 memes (Train: 500, Validation: 50, Test: 50, Total: 600). The training dataset has been used to learn model parameters, the validation dataset has been used for model selection and hyperparameter tuning during training, and the test dataset has been used only for final evaluation. The data distribution is shown in Table 1.

Level	Category	Malayalam	Tamil	Total
Level 1	Troll/Oppose	477	691	1168
	Support/Praise	23	112	135
Level 2	Individual Person	327	633	960
	Party	120	170	290
	Intersection	53	0	53
Total		1000	1606	2606

Table 1: Dataset Statistics for Tamil and Malayalam Political Memes

4 Methodology

4.1 Problem Formulation

The task has been formulated as a multimodal multi-task classification problem. Given a political meme m that has been composed of a textual component t and a visual component i, the objective has been to perform hierarchical classification at two levels:

- **Level 1:** Support vs. Troll/Oppose
- **Level 2:** Person vs. Party vs. Intersection

Let $t \in R^m$ denote the textual representation of features that have been extracted from the meme text, and $i \in R^m$ denote the visual representation of features that have been extracted from the meme image. The goal has been to learn a mapping function:

$$f(t, i) \rightarrow \{y_1, y_2\} \quad (1)$$

As defined in Eq. 1, where $y_1 \in \{0, 1\}$ has represented the Level 1 label (0 = Support, 1 = Troll/Oppose), and $y_2 \in \{0, 1, 2\}$ has represented the Level 2 label (0 = Person, 1 = Party, 2 = Intersection). The model has leveraged multimodal fusion of textual and visual features to maximize multilevel classification performance.

4.2 Data Preprocessing

4.2.1 Text Preprocessing

For the textual modality, pretrained multilingual tokenizers have been employed with a maximum sequence length of 128 tokens. Unwanted symbols, punctuation, numbers, URLs, and emojis have been removed. Language-specific stopwords for Tamil and Malayalam have been eliminated to preserve semantically meaningful content. Contextual embeddings have been extracted using transformer-based encoder models.

4.2.2 Image Preprocessing

For the visual modality, images have been resized to 224×224 pixels and have been normalized using ImageNet mean and standard deviation statistics. The processed images are converted into tensor representations. Global visual features are extracted using pretrained backbone networks.

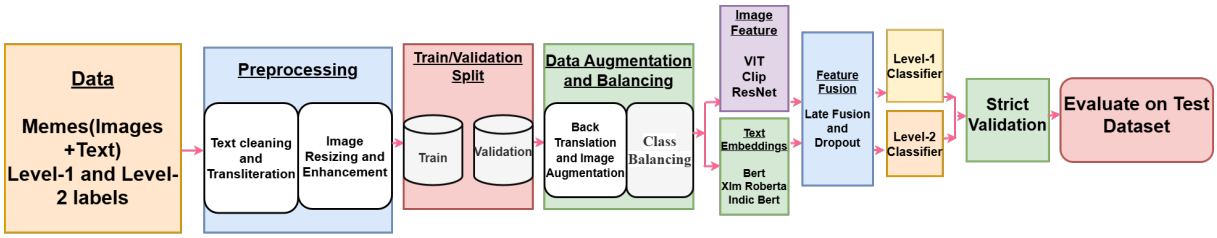


Figure 1: Multimodal Process Flow Framework for Detecting Political Memes

4.3 Data Augmentation

To improve generalization and address class imbalance, data augmentation has been applied to the image modality. Minority classes have been augmented in both Level 1 and Level 2 for Tamil and Malayalam datasets to reduce class disparity.

For the textual modality, a back-translation strategy has been employed using the *deep-translator* framework to generate semantically equivalent paraphrased samples.For the visual modality, a data augmentation strategy has been applied using controlled geometric and color transformation to generate visually varied yet semantically preserved image samples.

4.4 Multimodal Architecture

We implement a late-fusion multimodal architecture with separate pretrained text encoders (mBERT, XLM-RoBERTa, IndicBERT) and vision backbones (ViT, CLIP, ResNet-50), whose representations are concatenated and fed into a dropout layer followed by two parallel fully connected heads for Level-1 and Level-2 classification. As illustrated in Figure 1, the framework integrates preprocessing, augmentation, feature extraction, and fusion within a unified pipeline. Among all configurations, mBERT+ViT achieves the best performance on both Tamil and Malayalam datasets, using a maximum sequence length of 128 tokens and image resolution of 224×224.

4.5 Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-score. Due to potential class imbalance, Macro-F1 was considered the primary evaluation metric. Performance was assessed separately for Level 1 and Level 2 classification tasks on both the Tamil and Malayalam datasets.

4.6 Parameter Setting

We have used AdamW as the optimizer and cross-entropy loss for training. ViT + mBERT uses a higher learning rate because both components are more stable during fine-tuning and require faster adaptation, whereas XLM-R and CLIP are more sensitive and need smaller learning rates to avoid overfitting or instability. Table 2 lists the hyperparameters used for XLM-RoBERTa + CLIP, mBERT + ViT, and IndicBERT + ResNet.

Model	Learning Rate	Batch Size	Epochs
ViT + mBERT	1e-4	16	10
XLM-R + CLIP	2e-5	16	10
IndicBERT + ResNet	2e-5	16	10

Table 2: Key Hyperparameters for Model Training

4.7 Result Analysis

This section presents the experimental results and comparative analysis of the proposed multimodal approaches. The models were analyzed based on their overall classification effectiveness across different tasks and datasets, with particular emphasis on Macro-F1 performance.

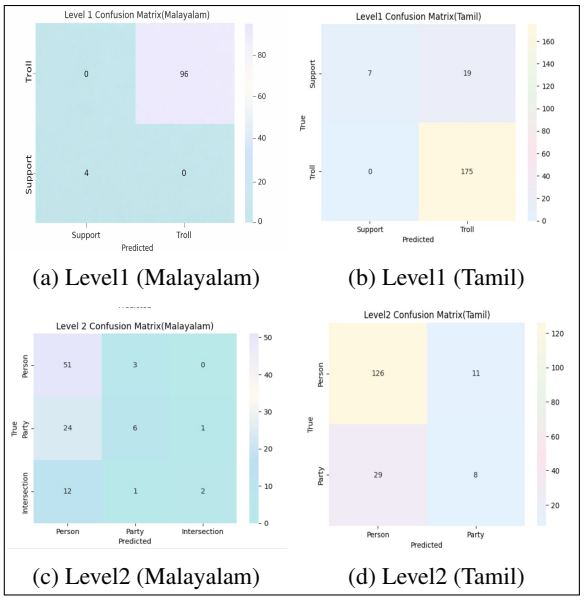
4.8 Comparative Analysis

We evaluate three multimodal configurations—ViT+mBERT, XLM-R+CLIP, and IndicBERT+ResNet—on Tamil and Malayalam datasets. As shown in Table 3, ViT+mBERT achieves the best performance across both levels, obtaining weighted F1 scores of 0.68 (Tamil) and 0.49 (Malayalam) at Level-1, and 0.53 (Tamil) and 0.42 (Malayalam) at Level-2. While Level-1 consistently yields higher scores, Level-2 remains more challenging due to fine-grained target distinctions. The results indicate that ViT+mBERT effectively captures complementary textual and visual patterns with stable generalization under strict validation (dropout = 0.3). The relatively lower performance in Malayalam and Level-2 primarily reflects dataset limitations, class imbalance, noisy OCR text, and intrinsic task complexity rather than model overfitting.

Figure 2 indicates strong but imbalanced Level-1 performance. For Malayalam, the model correctly classified 96 Troll and 4 Support instances, with

Lang	Model	P	R	F1
	Level-1			
Tamil	Text (mBERT)	0.68	0.53	0.50
Tamil	Image (ViT)	0.83	0.71	0.66
Tamil	Multi (mBERT+ViT)	**0.95**	**0.63**	**0.68**
Tamil	Multi (CLIP+XLM-R)	0.91	0.60	0.65
Tamil	Multi (IndicBERT+ResNet)	0.88	0.58	0.61
Malayalam	Text (mBERT)	0.47	0.46	0.48
Malayalam	Image (ViT)	0.46	0.49	0.47
Malayalam	Multi (mBERT+ViT)	**0.48**	**0.50**	**0.49**
Malayalam	Multi (CLIP+XLM-R)	0.47	0.49	0.48
Malayalam	Multi (IndicBERT+ResNet)	0.46	0.48	0.47
	Level-2			
Tamil	Text (mBERT)	0.39	0.50	0.44
Tamil	Image (ViT)	0.59	0.56	0.53
Tamil	Multi (mBERT+ViT)	**0.52**	**0.59**	**0.57**
Tamil	Multi (CLIP+XLM-R)	0.50	0.55	0.52
Tamil	Multi (IndicBERT+ResNet)	0.48	0.53	0.50
Malayalam	Text (mBERT)	0.35	0.37	0.35
Malayalam	Image (ViT)	0.52	0.46	0.42
Malayalam	Multi (mBERT+ViT)	**0.61**	**0.41**	**0.49**
Malayalam	Multi (CLIP+XLM-R)	0.49	0.39	0.36
Malayalam	Multi (IndicBERT+ResNet)	0.41	0.36	0.32

Table 3: Performance Comparison of models in Tamil and Malayalam Datasets



(a) Level1 (Malayalam) (b) Level1 (Tamil)

(c) Level2 (Malayalam) (d) Level2 (Tamil)

Figure 2: Confusion matrices for Level-1 and Level-2 classification in Tamil and Malayalam using mBERT and ViT.

no misclassification of Troll but a clear bias toward predicting Troll. For Tamil, it correctly identified 175 Troll and 7 Support, with 19 Support samples misclassified as Troll, again showing similar bias.

At Level-2 (Malayalam), the model correctly classified 51 Person, 6 Party, and 2 Intersection instances, but showed notable confusion, with many Party and Intersection samples misclassified as Person. At Level-2 (Tamil), the model correctly classified 126 Person and 8 Party instances, while misclassifying 29 Party as Person and 11 Person as Party, indicating a similar bias.

Overall, Level-1 demonstrates high accuracy, especially for the Troll class, whereas Level-2 remains more challenging due to class imbalance and fine-grained distinctions.

5 Error Analysis

The Level-1 results indicate that the models are quite effective at distinguishing between Troll and Support content in both Malayalam and Tamil, with relatively few misclassifications in each language. At Level-2, however, errors increase due to the more fine-grained class definitions. For Malayalam, most confusion occurs between posts about individuals, political parties, and intersectional content, while in Tamil the errors are mainly between individual and party-related posts. This pattern suggests that subtle contextual cues and overlapping semantic signals make fine-grained classification more difficult. Across these results, multimodal fusion shows limited additional benefit when one modality already captures the dominant signal, whereas improvements are more likely when both modalities contribute complementary contextual information. For future work, improvements in multimodal fusion and external knowledge integration can help reduce confusion arising from fine-grained class granularity and overlapping contextual cues.

6 Conclusion

In this work, we have presented a multi-level meme classification framework for Tamil and Malayalam political memes. We have conducted a comparative analysis of various unimodal and multimodal architectures for detecting political content in memes. The experimental results have demonstrated that multimodal fusion of textual and visual features have significantly improved classification performance, highlighting the importance of integrating both modalities. Among the evaluated models, the mBERT+ViT architecture has achieved the best overall performance across both Level 1 and Level 2 tasks. Future work should ensure better generalization by exploring more extensive and diverse datasets. Additionally, experimenting with larger vision language pre-trained models may enhance the performance.

7 Limitations

Despite moderate results, limitations persist due to small and imbalanced datasets, noisy OCR-based text extraction, and reliance on general-purpose pretrained encoders without task-specific linguistic adaptation, which may restrict the capture of nuanced cultural and contextual cues in regional political memes.

Acknowledgment

The authors would like to thank the organizers of DravidianLangTech@ACL 2026 for providing the competition platform and resources that facilitated this research. We also acknowledge the dataset introduced in the shared task overview paper (Rajiakodi et al., 2026), which formed the basis of our experiments.

Ethical Statement

All data processing and modeling followed ethical guidelines for handling sensitive, political content. The study aims to improve political memes detection while protecting users' rights and privacy. The goal is to improve automated moderation of political memes on online platforms. We have addressed any biases or limitations in the dataset to the best of our ability.

References

Tariq Afridi, Aftab Alam, Numan Khan, and Jawad Khan. 2021. *A Multimodal Memes Classification: A Survey and Open Research Issues*, pages 1451–1466.

Md Ashraful Alam, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Multimodal aggressive meme classification using bidirectional encoder representations from transformers. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3542–3547.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9):e0274300.

Richard Dawkins. 1976. *The Selfish Gene*. Oxford University Press, Oxford, U.K.

Adnan Faisal, Shiti Chowdhury, Momtazul Arefin Labib, and Hasan Murad. 2025. Team_Luminaries_0227@LT-EDI-2025: A transformer-based fusion approach to misogyny detection in Chinese memes. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 116–120, Naples, Italy. Unior Press.

Thomas Galipeau. 2023. The impact of political memes: A longitudinal field experiment. *Journal of Information Technology & Politics*, 20(4):437–453.

Shruthi Hariprasad, Sarika Esackimuthu, Saritha Madhavan, Rajalakshmi Sivanaiah, and Angel Deborah Suseelan. 2022. Ssn_mlrg1@dravidianlangtech-acl2022: Troll meme classification in tamil using transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 132–137, Dublin, Ireland. Association for Computational Linguistics.

Md Mithun Hossain, Md Shakil Hossain, M. F. Mridha, and Nilanjan Dey. 2026. A vision-language model for multitask classification of memes. *Neural Networks*, 194:108089.

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. *arXiv preprint arXiv:2005.04790*.

Zhenkai Qin, Qining Luo, Zhidong Zang, and Hongpeng Fu. 2025. Detecting sarcasm in user-generated content integrating transformers and gated graph neural networks. *PeerJ Computer Science*, 11:e2817.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Premjith B, Subalalitha CN, Rahul Ponnusamy, Anshid K A, Bhuvaneshwari Sivagnanam, Jananayagan V, Bharathi Raja Chakravarthi, Ragavan N, and Santhini P. 2026. Overview of the shared task on multilevel political meme classification in tamil and malayalam. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.