Semantica@DravidianLangTech 2026: Vision-Language Models for Hierarchical Political Meme Classification in Tamil and Malayalam

Junain Uddin, Rahul Datta, Taha Ibne Abdullah, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104040, u2104048, u2104052}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd
*

Abstract

Political memes are widely used to express opinions, sarcasm, and ideological narratives on social media platforms. However, detecting political trolling in low-resource languages such as Tamil and Malayalam remains challenging due to limited datasets and tools. To address this problem, DravidianLangTech@ACL 2026 organized a shared task on hierarchical political meme classification.

This work explores text-only models, classical multimodal fusion, and Vision-Language Models (VLMs) for Tamil and Malayalam political meme classification. Our experiments include IndicBERTv2, XLM-RoBERTa, EfficientNet-based multimodal fusion, and Qwen-VL models. Among the submitted systems, Qwen2.5-VL-7B-Instruct with 4-bit QLoRA fine-tuning achieved competitive performance, ranking 3rd in the Malayalam track and 4th in the Tamil track based on weighted-F1 score. Additional post-evaluation experiments with Qwen3-VL-8B further improved macro-F1 performance, highlighting the effectiveness of VLMs for low-resource multilingual political meme classification.

1 Introduction

Memes are widely used to express humor, sarcasm, and opinions through combined visual and textual content (Hegde et al., 2021), and have become a common form of communication on social media platforms. While many memes are created for entertainment, political memes are increasingly used to target individuals, parties, or ideologies through trolling and subtle forms of online harassment (Das et al., 2022). Such content can amplify polarized narratives and influence societal perceptions and democratic discussions, making NLP and AI techniques important for analyzing multimodal political content at scale.

Most existing studies mainly focus on high-resource languages such as English (Liu et al., 2025; Huertas-Tato et al., 2024), while comparatively less attention has been given to low-resource languages like Tamil, Malayalam, and Bangla. In addition, prior works primarily focus on hateful or harmful meme detection (Hossain et al., 2024; Pramanick et al., 2021), with limited emphasis on fine-grained political stance classification in multilingual settings.

To address this challenge, DravidianLangTech@ACL 2026 organized a shared task on "Multi-Level Political Meme Classification" (Rajiakodi et al., 2026) for Tamil and Malayalam political memes. The task aims to encourage the development of culturally aware multimodal systems capable of identifying political stance and target information from meme discourse in low-resource Dravidian languages. The main contributions of this work are as follows:

- We compare text-only, classical multimodal fusion, and Vision-Language Models for hierarchical political meme classification in Tamil and Malayalam.
- We analyze the impact of severe class imbalance on multimodal political meme classification across Level-1 and Level-2 tasks.
- We demonstrate the effectiveness of Vision-Language Models for low-resource multilingual political meme understanding.

2 Related Work

Multimodal meme analysis gained prominence with the Hateful Memes Challenge (Kiela et al., 2020), which highlighted the importance of jointly modeling visual and textual signals. Early cross-modal transformers such as ViLBERT (Lu et al., 2019) introduced attention-based alignment mech-

*https://github.com/Junainn/Shared-Task-2026

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 348–353

Language	Train	Test
Tamil	803	201
Malayalam	500	100
Total	1303	301

Table 1: Language-wise train/test distribution.

anisms for vision-language reasoning, forming the foundation of modern multimodal architectures.

In Dravidian languages, prior shared-task systems explored troll and misogyny meme detection using multimodal fusion techniques (Hegde et al., 2021; Das et al., 2022; Hossan et al., 2025). However, fine-grained political meme classification with stance and target prediction remains relatively underexplored.

Large vision-language models such as Qwen-VL (Bai et al., 2023) extend transformer-based multimodal reasoning to diverse downstream tasks. Parameter-efficient methods like LoRA (Hu et al., 2022), along with memory-efficient optimization techniques including 8-bit quantization (Dettmers et al., 2022) and AdamW optimization (Loshchilov and Hutter, 2019), enable practical adaptation of large multimodal models in low-resource settings.

Despite strong general-domain performance, large multimodal models remain underexplored for political meme classification involving cultural nuance, multilingual text, and severe class imbalance in low-resource languages.

3 Data Description

The dataset consists of political memes in two Dravidian languages: Tamil and Malayalam. Each sample contains a meme image, associated textual content, and hierarchical annotations. The task is formulated as a two-level classification problem: Level-1 represents coarse-grained stance (Troll/Oppose vs. Support), while Level-2 provides fine-grained target-specific labels.

Dataset Split: The dataset is divided into training and test sets for both languages, with a validation split created from the training data. Table 1 summarizes the overall distribution.

Level-1 Distribution: Tamil contains 691 Troll/Oppose and 112 Support/Praise samples, while Malayalam contains 477 Troll/Oppose and 23 Support/Praise samples, indicating severe imbalance in both datasets.

Level-2 Distribution: Tamil contains four fine-grained classes, while Malayalam contains five classes including an additional *Intersection* category. Tables 2 and 3 show the Level-2 training distribution.

Tamil Level-2 Class	Train
Troll/Oppose Against Person	547
Troll/Oppose Against Party	146
Support for person	86
Support for party	24

Table 2: Tamil Level-2 class distribution.

Malayalam Level-2 Class	Train
Against individual person	315
Against party	110
Intersection	53
Support for individual person	12
Support for party	10

Table 3: Malayalam Level-2 class distribution.

4 Methodology

We explored three modeling paradigms: (i) text-only models, (ii) classical multimodal fusion, and (iii) large Vision-Language Models (VLMs). All approaches were evaluated under identical train-test splits for both Level-1 and Level-2 classification tasks.

4.1 System Overview

The objective of exploring multiple paradigms is to systematically analyze the contribution of textual information, visual features, and large-scale cross-modal pretraining in political meme classification. Figure 1 shows the overall experimental pipeline used in this study.

4.2 Text-Only Approach

The text-only baseline evaluates classification using OCR-extracted text without visual features, employing a hierarchical two-stage strategy.

OCR Extraction and Hierarchical Classification: Text was extracted from meme images using Qwen2-VL-2B-Instruct with the prompt *"Extract all text from this image. Return only the text."* Outputs were cleaned using regex to remove conversational artifacts and cached to avoid recomputation. For hierarchical classification, we first predicted Level-2 (fine-grained) labels from OCR text, then prepended the prediction to the original text as `[Level2: <prediction>] <ocr_text>` for Level-1 classification. This allows coarse-level predictions to benefit from fine-grained contextual signals.

Models and Training: We fine-tuned two transformer encoders: ai4bharat/IndicBERTv2-MLM-

Figure 1: Overall experimental pipeline showing text-only, multimodal fusion, and VLM-based approaches.

only and xlm-roberta-base, with separate models for Level-1 and Level-2. Training used 15 epochs on an 85/15 stratified split with batch size 16, learning rate 1×10^{-5}, weight decay 0.01, warmup ratio 0.1, and max sequence length 128. Class imbalance was handled via weighted cross-entropy loss with FP16 mixed precision. Best checkpoints were selected by validation macro F1-score, which was also the primary evaluation metric alongside accuracy and per-class metrics.

4.3 Classical Multimodal Fusion

The classical multimodal baseline models visual and textual information jointly using an early-fusion architecture without relying on OCR.

Visual and Text Encoders: Images were encoded using pretrained EfficientNet-B3 (Tan and Le, 2019) (ImageNet pretrained). The final classification layer was removed, and the 1536-dimensional pooled feature was projected to a 512-dimensional representation via a linear layer with ReLU. For text, *bert-base-multilingual-cased* (mBERT) (Devlin et al., 2019) was used with bilingual political vocabulary prompts containing commonly used political terms, party references, and stance-related keywords in both English and Tamil/Malayalam. These prompts were intended to encourage politically relevant multilingual representations within mBERT. The 768-dimensional [CLS] representation served as the textual feature.

Fusion and Training: The 512-dimensional visual and 768-dimensional textual embeddings were concatenated into a 1280-dimensional joint representation and passed through a small MLP (Dropout $\rightarrow$ Linear $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear) for classification. Training followed a three-phase schedule: phases 1–2 (LR=2×10^{-4}, 5×10^{-5}) updated all layers, while phase 3 (LR=1×10^{-5}) froze mBERT and continued training the visual and fusion layers. Optimization used AdamW with class-weighted cross-entropy loss and a OneCycle scheduler, with strong image augmentation and class-balanced oversampling to address data imbalance.

4.4 Vision-Language Models

We fine-tuned instruction-tuned Vision-Language Models from the Qwen-VL family (Bai et al., 2023). While Qwen3-VL-8B achieved stronger post-evaluation performance, Qwen2.5-VL-7B was used for the official shared-task submission, as the Qwen3 experiments were conducted later during additional evaluation and analysis.

Architecture and Optimization: Qwen-VL processes images and text jointly within a unified transformer, enabling cross-modal reasoning without explicit feature fusion. The models can directly interpret Tamil and Malayalam text within images through multimodal pretraining, eliminating the need for OCR preprocessing. Models were loaded in 4-bit quantized format (Dettmers et al., 2022) with gradient checkpointing to enable efficient fine-tuning on limited GPU resources. Table 4 shows the training parameters used in this study.

Parameter-Efficient Fine-Tuning: We adopted Low-Rank Adaptation (LoRA) (Hu et al., 2022), inserting adapters into vision layers, language layers, attention modules, and MLP blocks.

Data Balancing and Training: To address severe class imbalance (23:1 for Tamil and 32:1 for Malayalam), minority classes were oversampled to approximately 200 samples per class. Training followed a supervised instruction-tuning setup, where each meme image was paired with a natural language classification prompt. The prompts explicitly defined each class label (e.g., *troll_person*, *troll_party*, and *support_person*) and instructed the model to generate only the final category name. For Malayalam, additional prompt clarification was added for the *intersection* class to better distinguish memes targeting both a political individual and their party, particularly in sarcastic or ambiguous cases. Tamil experiments converged in one epoch, while Malayalam required two epochs with a reduced learning rate in the second stage to refine

Parameter	Tamil	Malayalam
Quantization	4-bit	4-bit
LoRA rank (r)	16	16
LoRA alpha (α)	16	16
Batch Size	2	2
Learning Rate	2e-5	2e-5 (epoch 1) 5e-6 (epoch 2)
Epochs	1	2
Optimizer	AdamW (8-bit)	AdamW (8-bit)
LR Schedule	Cosine	Cosine

Table 4: VLM fine-tuning configuration

Model	Accuracy	Macro-F1	Weighted-F1
Tamil (201 samples)			
IndicBERTv2 (Text-only)	38.31	28.68	43.00
XLM-RoBERTa (Text-only)	44.28	33.87	47.00
mBERT + EfficientNet (Multimodal)	63.18	34.59	58.39
Qwen2.5-VL-7B	66.67	45.85	67.85
Qwen3-VL-8B	**70.15**	**49.70**	**69.71**
Malayalam (100 samples)			
IndicBERTv2 (Text-only)	48.00	27.12	42.00
XLM-RoBERTa (Text-only)	42.00	30.54	41.00
mBERT + EfficientNet (Multimodal)	36.00	22.16	34.11
Qwen2.5-VL-7B	49.00	35.94	49.78
Qwen3-VL-8B	**57.00**	**36.40**	**58.29**

Table 5: Level-2 classification performance (%) across text-only, multimodal, and VLM approaches.

class boundaries. Inference used greedy decoding for deterministic predictions.

5 Results and Analysis

5.1 Level-2 Classification Results

Level-2 classification focuses on fine-grained stance categories in political memes. Table 5 compares text-only models, classical multimodal fusion, and vision-language models.

Across both languages, models that incorporate visual information perform better than text-only baselines. The best results are obtained by Qwen3-VL-8B, which achieves a macro-F1 of 49.70% for Tamil and 36.40% for Malayalam. This improvement suggests that jointly modeling visual and textual cues is important for capturing the nuanced meaning of political memes.

5.2 Level-1 Classification Results

Level-1 classification focuses on binary stance detection (Support/Praise vs. Troll/Oppose). Table 6 reports results across Tamil and Malayalam.

For Tamil, models that incorporate visual information perform better than text-only baselines, with Qwen3-VL-8B achieving the best macro-F1 score of 82.12%.

Model	Accuracy	Macro-F1	Weighted-F1
Tamil (201 samples)			
IndicBERTv2 (Text-only)	67.16	51.18	72.00
XLM-RoBERTa (Text-only)	78.11	56.90	79.00
mBERT + EfficientNet (Multimodal)	81.59	67.63	83.39
Qwen2.5-VL-7B	87.06	75.84	88.05
Qwen3-VL-8B	**90.55**	**82.12**	**91.22**
Malayalam (100 samples)			
IndicBERTv2 (Text-only)	96.00	48.98	94.00
XLM-RoBERTa (Text-only)	96.00	48.98	94.00
mBERT + EfficientNet (Multimodal)	92.00	31.94	92.00
Qwen2.5-VL-7B	88.00	31.54	90.84
Qwen3-VL-8B	**85.00**	**30.80**	**88.70**

Table 6: Level-1 binary classification performance (%) across text-only, multimodal, and VLM approaches.

Malayalam results are strongly affected by severe class imbalance, with only four Support/Praise instances compared to ninety-six Troll/Oppose samples in the test set. As a result, text-only models achieve high accuracy (96%) by predicting the majority class, while macro-F1 reveals poor minority-class performance.

Overall, Level-1 classification is substantially easier than the fine-grained Level-2 task, leading to higher scores across all models.

6 Conclusion

This work investigated hierarchical political meme classification for Tamil and Malayalam using text-only models, classical multimodal fusion, and Vision-Language Models (VLMs). The best-performing results were achieved by Qwen3-VL-8B, which obtained the highest macro-F1 and weighted-F1 scores across both Tamil and Malayalam datasets in the Level-1 and Level-2 classification tasks.

Our analysis also highlights severe class imbalance in the Malayalam dataset, where accuracy can be misleading despite poor minority-class performance. This emphasizes the importance of macro-averaged metrics for reliable evaluation.

Overall, jointly modeling visual and textual information improves political meme classification in low-resource languages.

Limitations

The provided dataset was relatively small, especially for the malayalam dataset. Both datasets were extremely imbalanced which impacted the predictions for minority classes. The performance was constrained by limited access to GPU resources, which restricted the use of more computationally intensive and powerful models.

Ethical Statement

All sources referenced in this study have been appropriately cited, and care was taken to ensure the originality and integrity of the work. Through this research, we aim to contribute to improved detection of harmful political memes, with the broader objective of helping mitigate online trolling and bullying. The study was conducted in accordance with ethical research standards, and the dataset was used strictly for academic purposes without promoting or endorsing any political ideology.

Acknowledgement

We express our gratitude to the DravidianLangTech 2026 shared task organizers for facilitating this evaluation campaign.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. hate-alert@dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification. *Preprint*, arXiv:2204.12587.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Uvce-iiitt@dravidianlangtech-eacl2021: Tamil troll meme classification: You need to pay more attention. *Preprint*, arXiv:2104.09081.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M. Preum. 2024. Deciphering hate: Identifying hateful memes and their targets. *Preprint*, arXiv:2403.10829.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A multimodal fusion-based approach for identifying misogyny memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 427–434, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Javier Huertas-Tato, Christos Koutlis, Symeon Papadopoulos, David Camacho, and Ioannis Kompatsiaris. 2024. A clip-based siamese approach for meme classification. *Preprint*, arXiv:2409.05772.

Douwe Kiela et al. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.

Jiaqi Liu, Ran Tong, Aowei Shen, Shuzheng Li, Changlin Yang, and Lisha Xu. 2025. Memeblip2: A novel lightweight multimodal system to detect harmful memes. *Preprint*, arXiv:2504.21226.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu et al. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *Preprint*, arXiv:2109.05184.

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.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

A Error Analysis

Analysis of misclassifications reveals several consistent patterns across both tasks and languages. For Level-2 classification, models often confuse "Against individual person" and "Against party", as these categories share similar rhetorical expressions and visual cues in political memes. Performance also drops for minority classes, particularly "Support for individual person", which consistently receives much lower F1 scores than the dominant opposition categories.

This issue is especially evident in the Malayalam dataset, where the Level-1 test set contains only four Support/Praise instances compared to ninety-six Troll/Oppose samples. As a result, text-only models achieve high accuracy (96%) by predicting the majority class, while failing to detect the minority class, leading to zero F1 score for Support/Praise.

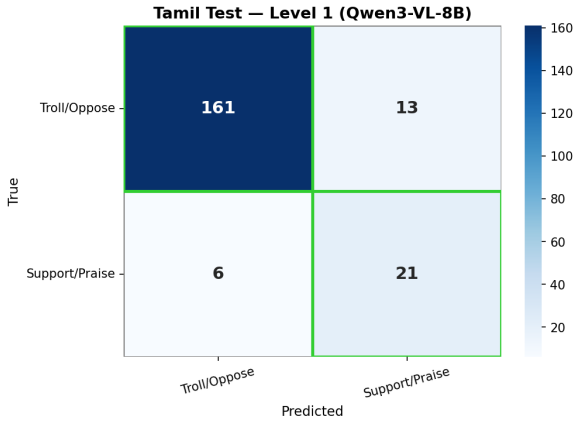

Figure 2: Tamil — Level 1 (Binary)

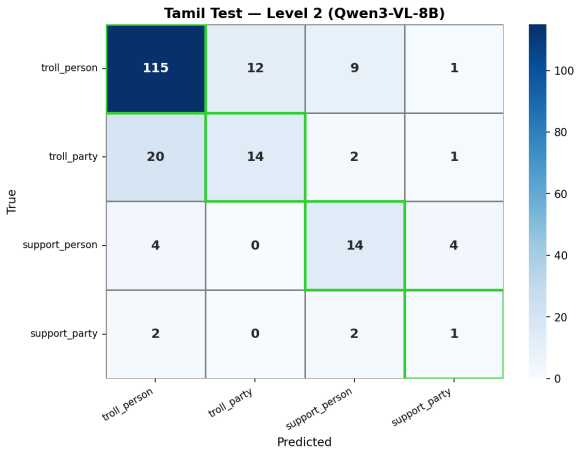

Figure 3: Tamil — Level 2 (4-class)

Visual information improves performance on the Tamil dataset by reducing confusion between fine-grained Level-2 categories. Vision-Language Models show better ability to distinguish person-oriented and party-oriented attacks. In the Tamil Level-1 matrix (Figure 2), the model correctly identifies the dominant Troll/Oppose class, while Level-2 (Figure 3) shows confusion between *troll_person* and *troll_party*. However, in the Malayalam dataset, severe class imbalance limits the ability of all models to learn reliable patterns for minority classes. The Malayalam Level-1 matrix (Figure 4) confirms near-total collapse onto the Troll/Oppose class, while Level-2 (Figure 5) shows that the *intersection* class is frequently misclassified as *troll_person*. These findings highlight the importance of balanced evaluation datasets and macro-averaged metrics.

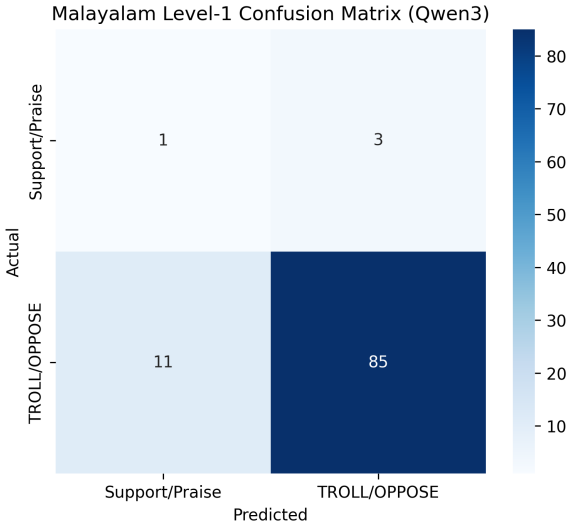

Figure 4: Malayalam — Level 1 (Binary)

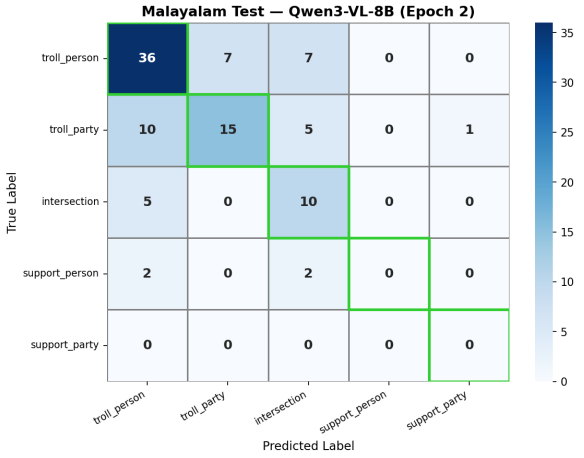

Figure 5: Malayalam — Level 2 (5-class)