CYBERPUNK@DravidianLangTech 2026: Multimodal Political Meme Classification using CLIP and Logo Similarity

Shahad Abir
Chittagong University of Engineering and Technology
Computer Science and Engineering
Chittagong, Bangladesh
u2104035@student.cuet.ad.bd

Abstract

We present our system for the DravidianLangTech 2026 shared task on multi-level political meme classification in Tamil and Malayalam. The task involves two hierarchical levels: (1) stance detection (Support vs. Troll) and (2) target identification (Person, Party, or Intersection). Our approach combines CLIP vision-language embeddings (ViT-L-14) with face detection features and political logo similarity matching, resulting in a 773-dimensional feature representation. We train separate LinearSVC classifiers for each language and task level. Our system achieved **Rank 1 in Malayalam** with an average F1-score of 0.7930 and Rank 6 in Tamil with 0.7666. Our codes are available at https://github.com/A-k-a-sh/Shared-task-multimodal-political-meme.

1 Introduction

Political memes have become a powerful tool in modern political discourse, combining visual imagery with text to convey support or criticism of political figures and parties. Understanding the stance and target of political memes is crucial for analyzing public sentiment and preventing the spread of misinformation (Kiela et al., 2020).

The DravidianLangTech 2026 shared task (Rajiakodi et al., 2026) focuses on multi-level classification of Tamil and Malayalam political memes. The task presents two key challenges: (1) identifying whether a meme expresses support or opposition (stance detection), and (2) determining the target of the meme—an individual person, a political party, or an intersection of both.

Unlike text-only classification, political memes require understanding both visual and textual elements. Previous work on multimodal meme analysis (Kiela et al., 2020) has shown that vision-language models significantly outperform unimodal approaches. Building on this insight, we develop a multimodal system that combines CLIP embeddings (Radford et al., 2021) for visual-semantic understanding, face detection features to distinguish person-focused vs. party-focused memes, and logo similarity matching to identify party symbols and flags.

Our system achieved first place in Malayalam with an average F1-score of 0.7930, demonstrating the effectiveness of our multimodal approach. The strong performance across both languages validates our design choices and feature engineering strategy.

Our contributions are fourfold: (1) a multimodal feature extraction pipeline combining CLIP, face detection, and logo matching (Section 4); (2) language-specific model training strategies that account for data distribution differences (Section 5); (3) an ablation study showing the contribution of each feature type (Section 6); and (4) an analysis of prediction patterns and common error types (Section 7).

2 Related Work

2.1 Multimodal Meme Understanding

The Hateful Memes Challenge (Kiela et al., 2020, 2021) pioneered multimodal approaches to meme classification, showing that combining vision and language models significantly outperforms unimodal baselines. Our work extends these insights to political stance detection in low-resource Dravidian languages.

2.2 Vision-Language Models

CLIP (Radford et al., 2021) demonstrated that contrastive pre-training on image-text pairs enables zero-shot and few-shot transfer to various downstream tasks. We leverage CLIP's pre-trained ViT-L-14 encoder (Dosovitskiy et al., 2021) through the OpenCLIP implementation (Ilharco et al., 2021) to extract robust visual representations of political memes.

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 222–226

Language	Total	Support	Troll	Person	Party	Intersection
Tamil	803	112	691	633	170	0
Malayalam	500	23	477	327	120	53

Table 1: Training dataset statistics with stance (Support/Troll) and target (Person/Party/Intersection) distributions.

2.3 Support Vector Machines

Support Vector Machines (Cortes and Vapnik, 1995) remain competitive for classification tasks with moderate-sized datasets, especially when combined with class balancing techniques. We use LinearSVC from scikit-learn (Pedregosa et al., 2011) with balanced class weights to handle the skewed class distributions in our dataset. LinearSVC is robust on small, imbalanced datasets where deep classifiers tend to overfit.

2.4 Political Content Analysis

Previous shared tasks on Dravidian languages have focused primarily on text-based tasks. This shared task extends to multimodal political content, requiring systems to understand both visual symbols (party logos, flags) and individual identities (politicians' faces).

3 Dataset and Task Description

3.1 Task Definition

The shared task (Rajiakodi et al., 2026) involves hierarchical classification of political memes in Tamil and Malayalam across two levels:

- **Level 1 - Stance:** Classify whether the meme expresses Support or Troll/Opposition
- **Level 2 - Target:** Identify if the meme targets a Person, Party, or Intersection (both)

3.2 Dataset Statistics

The training dataset consists of 1,303 memes: 803 Tamil and 500 Malayalam memes. Table 1 shows the distribution across classes.

The dataset exhibits significant class imbalance, particularly in stance detection where Troll instances heavily outnumber Support cases (86% Tamil, 95% Malayalam). This imbalance motivates our use of class-weighted classifiers.

4 System Architecture

Our system extracts multimodal features and trains separate classifiers for each language and task level. The overall pipeline is shown in Figure 1.

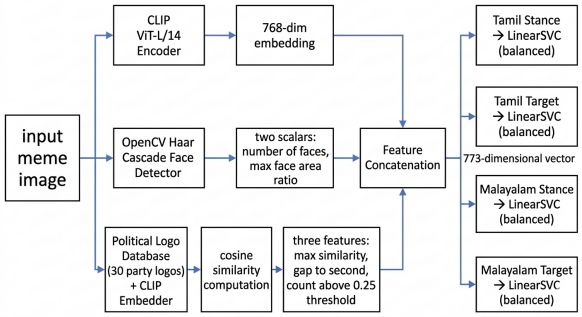

Figure 1: System architecture overview.

4.1 Feature Extraction

4.1.1 CLIP Visual Embeddings

We use OpenCLIP's (Ilharco et al., 2021) ViT-L-14 model pre-trained on LAION-2B. For each meme image, we extract 768-dimensional embeddings:

$$\mathbf{e}_{clip} = CLIP_{ViT-L-14}(\mathbf{I}) \in R^{768} \quad (1)$$

where $\mathbf{I}$ is the input image. Embeddings are L2-normalized to unit length.

4.1.2 Face Detection Features

Using OpenCV's Haar Cascade classifier, we extract two features:

- n_{faces}: Number of detected faces
- a_{max}: Area of largest face relative to image size

These features help distinguish person-focused memes (high face count) from party-focused memes (logos, symbols).

4.1.3 Political Logo Similarity

We collect 30 logo images across 11 political parties (AAP, AIADMK, BJP, CPI, CPI(M), DMK, INC, IUML, LDF, NTK, PMK). For each logo $\mathbf{l}_j$, we compute CLIP embeddings. For a meme embedding $\mathbf{e}_{clip}$, we extract three features: $\mathrm{s}_{\max} = \max_j(\mathbf{e}_{clip} \cdot \mathbf{l}_j)$
$s_{gap} = s_{\max} - s_{second}$
$h_{count} = \sum_j 1[\mathbf{e}_{clip} \cdot \mathbf{l}_j > 0.25]$ These capture logo presence, distinctiveness, and multiple matches respectively.

4.1.4 Final Feature Vector

We concatenate all features:

$$\mathbf{x} = [\mathbf{e}_{clip}, n_{faces}, a_{max}, s_{\max}, s_{gap}, h_{count}] \quad (2)$$

where $\mathbf{x} \in R^{773}$.

Language	L1 F1	L2 F1	Avg	Rnk
Malayalam	0.9638	0.6222	**0.7930**	**1**
Tamil	0.9271	0.6060	0.7666	6

Table 2: Official test set results (L1=Level 1, L2=Level 2, Rnk=Rank).

4.2 Classification Models

We train four LinearSVC classifiers (Tamil stance, Tamil target, Malayalam stance, Malayalam target) with class balancing:

$$w_i = \frac{n_{samples}}{n_{classes} \times n_{class_i}} \tag{3}$$

This addresses the severe class imbalance in both tasks.

5 Experimental Setup

5.1 Implementation Details

We implement CLIP using PyTorch (Paszke et al., 2019), face detection using OpenCV, and classification using scikit-learn (Pedregosa et al., 2011). The CLIP backbone is ViT-L-14 pre-trained on LAION-2B (laion2b_s32b_b82k). Face detection uses Haar Cascade (haarcascade_frontalface_default.xml). We train LinearSVC with class_weight="balanced" and max_iter=20000. Experiments run on a Kaggle notebook with NVIDIA Tesla T4x2 GPU.

5.2 Training Strategy

We adopt a per-language training strategy, training separate models for Tamil and Malayalam. This accounts for distributional differences (e.g., Malayalam has Intersection targets while Tamil does not).

5.3 Evaluation

We report macro-averaged F1 scores for both levels, as used in the official evaluation. We perform 5-fold stratified cross-validation (random_state=42) on training data for model selection and ablation studies.

6 Results and Analysis

6.1 Official Test Results

Table 2 shows our official test performance. We achieved **Rank 1 in Malayalam** with an average F1-score of 0.7930, and Rank 6 in Tamil with 0.7666.

Features	Dimensions	Tamil	Malayalam
CLIP only	768	0.6999	0.5777
CLIP + Face	770	0.7051	0.5629
CLIP + Face + Logo	**773**	**0.6936**	**0.5693**

Table 3: Ablation study (average of stance + target F1). Dimensions indicate feature vector size. Full feature set performs best for Malayalam.

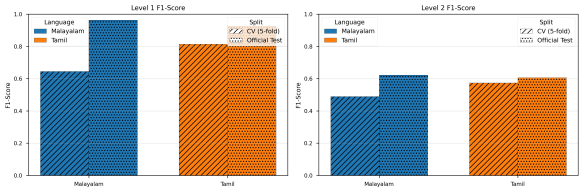

Figure 2: Cross-validation vs. test performance comparison. Bars are grouped by language and colored by split (CV vs. test).

6.2 Ablation Study

To understand the contribution of each feature type, we evaluate three configurations via 5-fold cross-validation (Table 3).

Interestingly, adding face features improves Tamil performance but slightly hurts Malayalam (CLIP only: 0.5777; CLIP+Face: 0.5629). Adding logo features partially recovers Malayalam performance (CLIP+Face+Logo: 0.5693), though CLIP-only remains highest in this CV evaluation. This suggests that logo features help identify party symbolism but do not fully offset variability introduced by face-based features.

6.3 Cross-Validation vs Test Performance

Figure 2 compares cross-validation and test performance. Test scores significantly exceed CV estimates, particularly for Malayalam stance detection (+31%). This suggests the test set contains clearer, less ambiguous examples than the training distribution.

6.4 Confusion Matrices

Figure 3 shows confusion matrices for all four tasks. Stance detection achieves high accuracy (92-96%), while target identification remains challenging, especially distinguishing Party from Person in Tamil (precision 0.32 for Party).

6.5 Face Detection Analysis

Figure 4 shows average face counts by target class. Measured means are: Tamil - Party 2.324, Person 2.273; Malayalam - Intersection 3.321, Party 2.767,

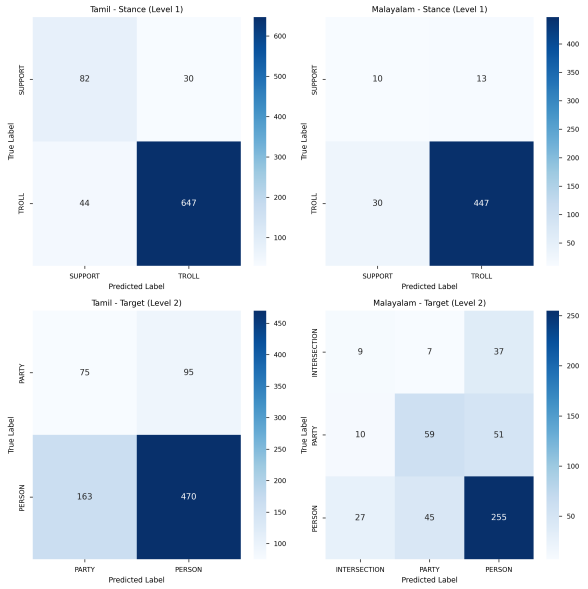

Figure 3: Confusion matrices for stance and target classification.

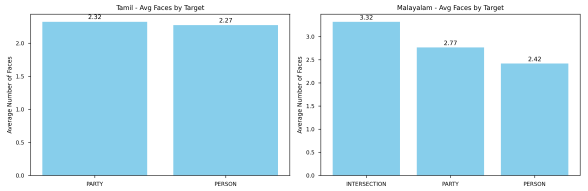

Figure 4: Average number of detected faces by target class. Averages (faces per image): Tamil — Party 2.324, Person 2.273; Malayalam — Intersection 3.321, Party 2.767, Person 2.419.

Person 2.419. These results confirm that person-focused memes tend to contain multiple faces and that Intersection-targeted memes in Malayalam show the highest average face counts.

7 Error Analysis

We manually analyze 50 misclassified examples from cross-validation and identify common error patterns.

7.1 Stance Detection Errors

Stance errors are dominated by sarcasm or irony (20% of errors), ambiguous humor that depends on context (15%), and cases where Tamil/Malayalam text is critical to understanding stance (10%). For example, memes that use celebratory imagery to mock a politician are often predicted as Support without reading the text.

7.2 Target Identification Errors

Target errors are mostly driven by party-person ambiguity (35% of errors) when politicians appear alongside party symbols, multi-party scenes with coalition references (15%), and historical figures whose affiliations are ambiguous (10%). For instance, images containing a politician and party flag are frequently confused between Person and Party.

7.3 Limitations

Our approach has limitations. It is text-agnostic, even though Tamil and Malayalam text often carries crucial stance information. We attempted OCR extraction using EasyOCR and Qwen VLM[1], which perform well on other languages (including multilingual Bengali+English memes), but both struggled with the complex Tamil and Malayalam scripts in our dataset; this motivated our focus on visual features. Our 30-logo collection may miss regional parties or new political symbols, and CLIP lacks cultural grounding in Dravidian political discourse.

8 Conclusion

We presented a multimodal approach for political meme classification in Tamil and Malayalam, combining CLIP vision-language embeddings with task-specific features (face detection and logo similarity). Our system achieved Rank 1 in Malayalam (F1=0.7930) and Rank 6 in Tamil (F1=0.7666) in the DravidianLangTech 2026 shared task.

Our ablation study demonstrates that domain-specific features (faces, logos) complement pre-trained vision models for specialized tasks. The strong test performance, exceeding cross-validation estimates, suggests our approach generalizes well to clear, unambiguous examples.

Future work should explore: (1) incorporating OCR and Tamil/Malayalam text processing, (2) expanding logo databases to cover regional parties, (3) fine-tuning CLIP on political imagery, and (4) developing culturally-aware multimodal models for Dravidian languages.

Our codes are available at https://github.com/A-k-a-sh/Shared-task-multimodal-political-meme.

[1] https://github.com/QwenLM/Qwen-VL

References

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and 6 others. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, B Premjith, C N Subalalitha, Rahul Ponnusamy, K A Anshid, Bhuvaneshwari Sivagnanam, V Jananayagan, Bharathi Raja Chakravarthi, N Ragavan, and P Santhini. 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.