CPSLab2@MMLow 2026: A Hierarchical Multimodal Framework with Stratified K-Fold Training for Political Meme Classification

Anika Srivastava[1,*], Vishal Kumar[1,†], Harshita Sharma[1,‡], Nitisha Aggarwal[1,§], Geetika Jain Saxena[2,¶], Amit Pundir[2,‖], Sanjeev Singh[1,]**

[1]Institute of Informatics and Communication, University of Delhi
[2]Department of Electronics, Maharaja Agrasen College, University of Delhi

[*]anikasrivastava151@gmail.com, [†]researchwithvishal@gmail.com, [‡]harshitasharma@south.du.ac.in, [§]nitisha@south.du.ac.in, [¶]gsaxena@mac.du.ac.in, [‖]amitpundir@mac.du.ac.in, [**]sanjeev@south.du.ac.in

Abstract

Political memes communicate opinions using the interplay between visual and textual information, posing a challenging multimodal classification task. We present a hierarchical multimodal framework for political meme classification, which decomposes the task into stance and target entity prediction. Our framework was evaluated on the data from the Multi-Level Political Meme Classification Shared Task at MMLow @ AACL-IJCNLP 2026. The task comprised two levels of classification: (i) predicting the political stance of a meme as either Support/Praise or Troll/Oppose, and (ii) identifying the target of the meme combined with the stance. We approached the task by first predicting the Level-1 labels followed by predicting the target entity (Person/Party) and then combining them in a deterministic manner to predict the final Level-2 labels. Our method used a SigLIP vision transformer for stance classification and a hybrid SigLIP-XLM-RoBERTa architecture for target classification with late fusion. We trained our models using 5-fold stratified cross-validation with checkpoint ensembling. We also performed ablation experiments to analyze the contribution of individual modalities, observing that visual features are more informative for stance prediction while textual features contain more information for target identification, which supports the multimodal task at hand. Our system achieved a macro F1 score of 0.9027 and 0.6694 for Level-1 and Level-2 classification, respectively, achieving a final score of 0.7860, ranking first in the shared task.

1 Introduction

Political memes have proved to be one of the most prominent online phenomena that shape political opinions through their multimodality, including visual, textual, and symbolic content, as well as their ability to convey complex semantic messages through humor and satire (Anderau and Barbarrusa, 2024). These memes pose unique challenges to computational methods due to their multimodal nature, which requires joint analysis of multiple signals. Political memes, especially those found in low-resource languages such as Tamil, Malayalam, Nepali (Thapa et al., 2025) etc. also pose additional challenges due to code-switching between languages, as well as the need to infer specific political context.

The MMLow Shared Task on Multi-Level Political Meme Classification approached the problem of classifying political memes through a multi-level classification paradigm. In particular, the task is split into two hierarchical levels, with Level 1 asking to classify the input meme into either Support/Praise or Troll/Oppose classes. Level 2 then uses the output of the first classifier to determine the finer classification of the meme, which is one of Support for Person, Support for Party, Troll/Oppose against Person, or Troll/Oppose against Party (Rajiakodi et al., 2026).

Rather than predicting each of the four Level-2 classes directly, we decomposed the problem into two binary classification subtasks: stance prediction (Support/Praise vs. Troll/Oppose) and target entity prediction (Party vs. Person). The two predicted labels were then combined using a deterministic function to obtain the final Level-2 label.

The first stage of our pipeline is based on the Vision Transformer model SigLIP that classifies the image part of the meme into either Support or Troll category. The second pipeline classified the meme into Party or Person by using a combination of multimodal representations created by SigLIP and textual representations obtained with the help of the XLM-RoBERTa language model. Furthermore, we performed an ablation study to measure the contribution of visual and textual modalities to the stance, target, and Level-2 classification tasks. Our proposed method achieved the best overall score (0.7860) for the shared task, ranking first.

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 91–96
November 9, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Advances in vision-language pretraining have enabled better understanding of the associations between images and text through the joint learning of representations in a large-scale dataset. Specifically, CLIP (Contrastive Language–Image Pretraining) learns image and text encoders in a joint manner using natural language supervision. This allowed for achieving strong zero-shot performance across a variety of vision tasks (Radford et al., 2021). SigLIP, an updated version of CLIP, uses a sigmoid-based pairwise loss, allowing for optimizing individual image-text pairs, as opposed to large batches of pairs in the original CLIP. This modification allowed a significant reduction in the amount of calculation needed to obtain results comparable to CLIP in terms of image-text matching accuracy and achieving higher performance on downstream vision tasks (Zhai et al., 2023). XLM-RoBERTa (XLM-R) is a Transformer-based model which was pretrained using masked language modeling on one hundred different languages (Conneau et al., 2020). It has shown strong performance on various downstream tasks in different languages, especially for tasks that may contain code-switching between two or more languages.

Low-resource languages pose additional challenges in achieving multimodal understanding due to a lack of annotated data, code-mixing, transliteration, and culturally specific references that are not present in high-resource languages. For Dravidian languages, recent works have aimed to construct benchmark data sets and initiate shared tasks for automatic social media content understanding. Such works include the tasks of detecting offensive language, identifying hate speech, and classifying memes (Chakravarthi et al., 2021); (B et al., 2022). These works showed that pretrained multilingual language models can benefit from the joint modeling of vision and language with limited parallel data. However, understanding political memes requires common sense, political, and regional cultural knowledge, necessitating a tailored multimodal approach.

Recent works have addressed the task of multimodal political meme analysis in the low-resource Dravidian languages context in the framework of the DravidianLangTech–ACL 2026 Shared Task on Multi-Level Political Meme Classification (Rajiakodi et al., 2026). The shared task, similar to the task at hand, proposed a hierarchical approach to classifying political memes in Tamil and Malayalam languages requiring first to detect the stance of a meme and then to identify the political target. The participants attempted to address the challenge by employing various multimodal approaches including transformer-based text encoders, CLIP-based vision encoders, OCR-assisted pipelines, cross-modal fusion techniques, and Vision Language Models. The results revealed that fusion of modalities and hierarchical modeling contribute significantly to improving the performance of stance detection while the identification of targets remains a challenging task due to sarcasm, cultural contexts, OCR errors, and label imbalance. The shared task provided a benchmark for exploring effective multimodal approaches to classifying political memes in the low-resource language setup.

3 Methodology

Figure 1 shows the overall architecture of our proposed framework. Our approach broke the problem into two hierarchical subtasks. First, a visual classifier was used to predict the coarse-grained stance (Support/Praise or Troll/Oppose). Second, a multimodal classifier was utilized to classify if the target of the meme is a political person or a political party by combining visual and textual information. Finally, a deterministic mapping was applied to the outputs of the first two classifiers to achieve the four fine-grained Level-2 categories (Support for Person, Support for Party, Troll/Oppose Against Person, and Troll/Oppose Against Party).

3.1 Data Preprocessing

The meme dataset consisted of images and their corresponding text i.e. the text present in the meme. The images were processed using the SigLIP AutoImageProcessor. Then the images were resized to 512×512 pixels. Normalization of the images was done according to the pretrained SigLIP Vision Transformer requirements (Zhai et al., 2023). To ensure robustness and reduce overfitting risks, each image underwent data augmentation, including RandomResizedCrop, 5-degree rotation, and ColorJitter. The meme text was tokenized using the XLM-RoBERTa tokenizer with a maximum sequence length of 128. Notably, any caption exceeding the limit was truncated, while short texts were padded. Moreover, to avoid blank captions, any missing text was replaced with “no caption”.

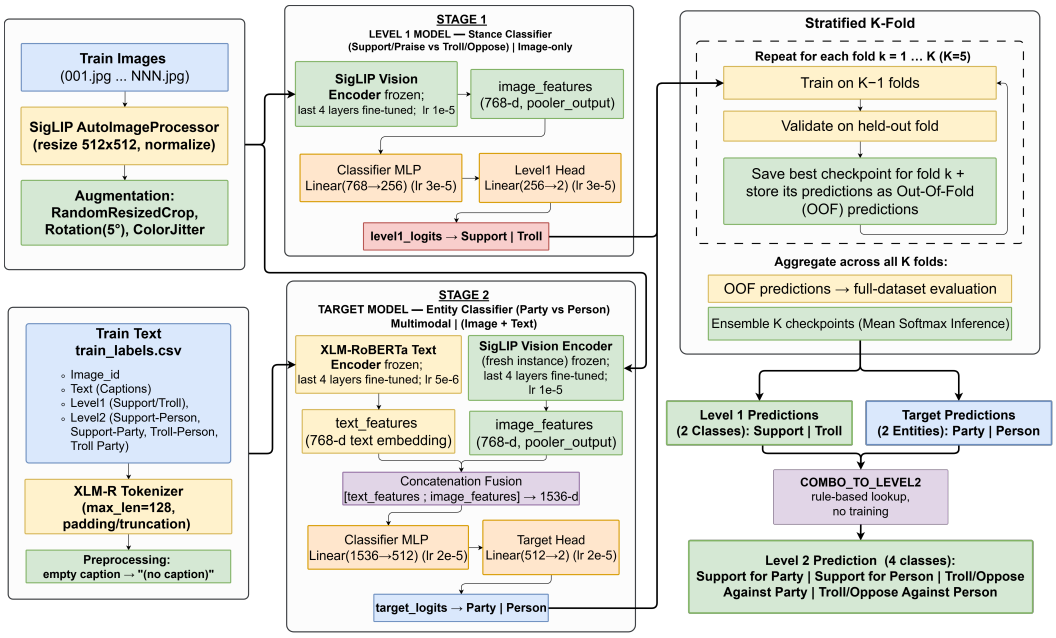


Figure 1: Overview of the training pipeline for the proposed hierarchical multimodal framework.

3.2 Stage 1: Visual Stance Classification for Level 1

The first stage was an image-only classifier to differentiate between the Support/Praise and Troll/Oppose stances. A pretrained SigLIP Vision Transformer was utilized as the image encoder. Given an input image (I), the encoder (E) produced a pooled visual representation

$$v = E_v(I), \quad \text{where } v \in R^{768} \tag{1}$$

To retain the pretrained visual knowledge while adapting to the downstream task, we partially fine-tuned the encoder by freezing all transformer layers except the last four. The resulting visual embedding was then fed into a lightweight multilayer perceptron, described in Table A.4. The model was trained with a class-weighted cross-entropy loss to address the imbalance between the two classes.

3.3 Stage 2: Multimodal Target Classification

The second level used both visual and textual information to predict whether the target of the meme was a party or a person. The input image was encoded using the pretrained SigLIP Vision Transformer from Section 3.2 (Eq. 1) resulting in a 768-dimensional visual embedding. The text from the accompanying meme was encoded using XLM-RoBERTa-base to obtain a 768-dimensional text embedding using attention-mask-aware mean pooling,

$$t = E_t(T), \quad \text{where } t \in R^{768} \tag{2}$$

where E_t denotes the text encoder and T represents the input meme text. For both models, only the last four transformer layers were fine-tuned, while the rest were frozen. Finally, visual and textual embeddings were concatenated: h=[v;t] to get a 1536-dimensional fused representation. This feature-level concatenation is termed as "late fusion" which allows the classifier to exploit complementary information from the two modalities for target prediction. The fused embedding, h, was then used as input to the classification network, which is described in Table A.4. The classifier then predicted whether the meme targets a political party or a political person.

3.4 Five-Fold Stratified Cross-Validation

Both the Stage-1 and Stage-2 models were trained using 5-fold stratified cross-validation (Kohavi, 1995) to ensure maximum utilization of the data, due to the low resource nature of the dataset, while retaining the distribution of the different classes in each fold. For each fold, one of them was taken as a hold-out validation set, and the other four were used for training. For each fold, a new model was trained from the pretrained encoder weights and early-stopped based on the performance of the validation set. The best model checkpoint and OOF predictions on the hold-out set were saved for ensemble and performance estimation purposes.

Once the five folds had been processed, the OOF predictions of all folds were pooled together, giv-

ing an unbiased estimate of performance on the training set. This allowed the models to generalize properly on unseen data later. When making predictions on the test set, all five of the best performing checkpoints were used to make the final prediction using mean softmax ensembling.

3.5 Level-2 Prediction

For a meme, the Stage 1 classifier predicts the stance label, i.e., the Level 1 for the task, while the target classifier predicts the target entity. The final Level-2 label is then obtained through a deterministic Cartesian-product mapping of the predicted stance and target labels, as shown in Table A.3, which produced the output of one of the four official shared-task labels for Level 2 (Support for Person, Support for Party, Troll Against Person, and Troll Against Party).

4 Results

Table 1 presents the official evaluation results of our submitted system on the test set which achieved Rank 1. The system is evaluated using the macro F1 Score, as described in Appendix A.2.

Category	Level 1	Level 2	Final
Score	0.9027	0.6694	0.7860

Table 1: Evaluation results as reported by the organizers.

5 Discussion

5.1 Stratified K-Fold Cross Validation

To obtain an unbiased assessment of model performance while using the training data most effectively, we performed a 5-fold stratified cross-validation. The validation was stratified on the Level-2 labels as it preserved the labels of Level 1. We performed stratified sampling to ensure that each class is represented with the same distribution in all folds. Training and validating on different splits reduces the model's dependence on a particular train-validation split and gives a better estimation of its generalization ability.

5.2 Ablation Study

The results of the ablation study comparing modalities, as shown in Table A.1, demonstrated that the majority modality depended on the task. The image-only model achieved a comparatively high Macro F1 of 0.8472 ± 0.0520 on the Level-1 stance classification, suggesting the effectiveness of visual cues such as political symbols, facial expressions, and meme templates, in differentiating the Support/Praise from Troll/Oppose stances. Seeing these results, we independently fine-tuned a Level-1 image-only model which attained an even better result of 0.8952 ± 0.0179, as compared to the Text+Image model at Level 1 (0.8838 ± 0.0263). Thus, an image-only classifier was ultimately used for the Level-1 prediction. We concluded that the addition of text modality did not consistently improve the performance and may have introduced additional noise, possibly due to the non-uniform and highly variable nature of meme accompanying texts.

Target prediction benefited more from textual information, as the text-only model (0.6618 ± 0.0562) significantly outperformed its image-only counterpart (0.2732 ± 0.0872). The multimodal model achieved a relatively higher performance than the unimodal models of 0.7288 ± 0.0433, implying the complementarity of vision and language cues for target identification. These insights directly informed the design of our hierarchical framework.

6 Conclusion

Our paper proposed a hierarchical classification framework for the Shared Task on Multi-Level Political Meme Classification. It decomposed the Level-2 classification task into two separate tasks: stance prediction and target entity discovery. The method predicted the stance label using a SigLIP-based image classifier while detecting target entities relies on a multimodal SigLIP–XLM-RoBERTa classifier with late fusion. The results of the two classifiers are merged into the final solution using a lookup table. Our method demonstrated its effectiveness as the validation score achieved using the 5-fold stratified cross-validation procedure almost equals the official test result, which suggests that the model is not overfitting. Our final shared-task score was 0.7860, which placed us in the first position. Some limitations faced were the dataset used for training is small and imbalanced, which impacts the quality of representation. Secondly, the target entity classifier used only learned representations for classification; incorporating additional knowledge, like a list of political parties and people, would allow improving the classification, especially for the Party/Person targets distinction.

References

Glenn Anderau and Daniel Barbarrusa. 2024. The function of memes in political discourse. *Topoi*, 43(5):1529–1546.

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman KP, Dhanalakshmi Vadivel, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John Philip McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

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. *Preprint*, arXiv:2103.00020.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Premjith B, Subalalitha CN, Rahul Ponnusamy, Anshid K A, Bhuvaneswari Sivagnanam, Jananayagan V, Ragavan N, Santhini P, and Bharathi Raja Chakravarthi. 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*, pages 101–113, Underline (Virtual). Association for Computational Linguistics.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45(4):427–437.

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025. A multimodal prompt-based framework for analyzing Code-Mixed and Low-Resource memes. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1913–1923.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

A Appendix

A.1 Ablation Study Results

Model	Level 1	Target	Level 2
Text-only	0.2618 ± 0.0726	0.6618 ± 0.0562	0.2154 ± 0.0147
Image-only	0.8472 ± 0.0520	0.2732 ± 0.0872	0.3510 ± 0.0350
Text + Image	$\mathbf{0.8838 \pm 0.0263}$	$\mathbf{0.7288 \pm 0.0433}$	$\mathbf{0.6668 \pm 0.0357}$

Table A.1: Macro F1 score results from the ablation study.

A.2 Evaluation Metrics

Following the shared task evaluation protocol, the model's performance was measured in terms of Precision, Recall, and F1-score. Precision represents the ratio of instances of a given class predicted by the model as positive that are actually positive. In turn, the Recall is the ratio of instances of a given class predicted as positive by the model to the total number of instances of this class. Finally, with F1-score being the harmonic mean of Recall and Precision, it summarizes the model's performance in predicting positive instances.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}, \\ \text{F1-score} &= 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{3}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively.

The overall performance can be measured in two ways – macro and weighted averages. The macro-average is the simple average of the different classes, calculated by the average between the classes regardless of their frequency. The weighted-average applies the number of instances in each class as weights, making the results affected by the class distribution and not an average value anymore. The macro F1 is useful in cases when we have a high imbalance between the classes and want to consider all the classes equally. On the other hand, the weighted F1 is a better estimate of the overall performance of the model than the macro F1 when the classes are imbalanced (Sokolova and Lapalme, 2009).

A.3 Dataset Distribution

Class	Before	After
Support/Praise	277	276
Troll/Oppose	692	693
Support for person	177	177
Support for party	99	99
Troll/Oppose Against Person	547	547
Troll/Oppose Against Party	146	146
Total	**969**	**969**

Table A.2: Class distribution before and after Level 1 label correction.

Table A.2 presents the distribution of samples across the Support/Praise and Troll/Oppose categories at Level 1, along with their corresponding Level 2 classes. During data preprocessing, few inconsistencies between the Level 1 and Level 2 annotations were discovered, where the Level 1 label did not match the category implied by the Level 2 label. These inconsistent labels were overridden using the corresponding Level 2-derived Level 1 label. The modified Level 1 labels were used in the experiments.

A.4 Level-2 Classification Mapping

Level-1 Class	Target	Level-2 Class
Support/Praise	Person	Support for Person
Support/Praise	Party	Support for Party
Troll/Oppose	Person	Troll/Oppose Against Person
Troll/Oppose	Party	Troll/Oppose Against Party

Table A.3: Stance and target combinations for the Level-2 classification.

A.5 Confusion Matrix and Error Analysis

Out of 969 instances, 710 (73.27%) had both stance and target correctly predicted. 177 (18.27%) instances had the correct stance but an incorrect target, while 56 (5.78%) had the correct target but an incorrect stance. The remaining 26 (2.68%) instances contained errors in both predictions. Therefore, it can be concluded that the majority of misclassified instances had issues with target identification, which suggests that the prediction of targets was more challenging than stance detection in this hierarchical pipeline.

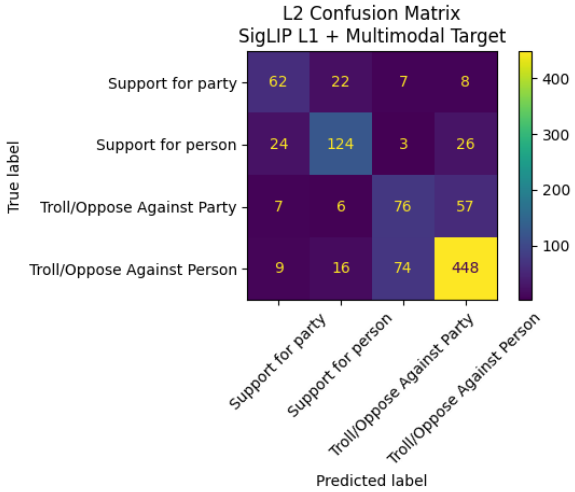


Figure A.1: Confusion matrix for the Level-2 classification, obtained by combining the SigLIP-based Level-1 stance predictions with the multimodal target predictions.

A.6 Classifier Architecture

Component	Level-1 Stance Classifier Configuration	Target Entity Classifier Configuration
Input	768-dimensional SigLIP image embedding	1536 (768+768)-dimensional fused image-text embedding
Dropout	0.3	0.3
Linear Layer	$768 \rightarrow 256$	$1536 \rightarrow 512$
Normalization	Layer Normalization	Layer Normalization
Activation	GELU	GELU
Dropout	0.2	0.2
Output Layer	Linear ($256 \rightarrow 2$)	Linear ($512 \rightarrow 2$)
Prediction	Support/Praise or Troll/Oppose	Party or Person

Table A.4: Architecture of the proposed Level-1 stance classifier and the multimodal target entity classifier.

A.7 Experimental Setup

The proposed approach was implemented in Python 3.12.13 using PyTorch library (version 2.11.0+cu128) and Hugging Face Transformers (version 5.13.1). In particular, the vision encoder was implemented using the pretrained SigLIP Vision Transformer (google/siglip-base-patch16-512), while the text encoder used was the pretrained XLM-RoBERTa-base (xlm-roberta-base). Finally, the evaluation scores including Precision, Recall, and F1 were calculated using scikit-learn library, while all computations were performed on the NVIDIA GPU with CUDA acceleration.

A.8 Code Availability

The implementation of our proposed framework is available at: https://github.com/anika-py/CPSLab2-MMLow-Political-Meme-Classification/tree/main