IndiLangTech@DravidianLangTech 2026: Hierarchical Modeling for Multi-Level Political Meme Classification

Saurabh Kumar, Vivekananda G, Sanasam Ranbir Singh, and Sukumar Nandi
Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati, Guwahati-781039, India
{saurabh1003, g.vivekanada, ranbir, sukumar}@iitg.ac.in

Abstract

Political memes are a widely used form of digital political expression in linguistically diverse regions such as South India, where visual cues, textual overlays, and cultural symbolism convey complex political narratives. The Shared Task on Multi-Level Political Meme Classification at DravidianLangTech 2026 introduces a hierarchical setting requiring stance identification (Support vs. Troll) and target-type prediction (Individual vs. Party) for Tamil and Malayalam memes. We propose a two-stage hierarchical framework based on the Gemma 3 4B Instruction model. Instead of jointly predicting both levels, two specialized models are fine-tuned: the first predicts meme stance, and its output conditions the second model for target identification, explicitly modeling the dependency between the meme content, the predicted stance, and the target type. Using LoRA-based parameter-efficient instruction tuning, our approach achieves an average F1-scores of 0.8029 for Tamil and 0.6950 for Malayalam across both levels, ranking **1st** in Tamil and **4th** in Malayalam.

1 Introduction

Memes have become a prevalent form of communication on social media platforms, enabling users to express opinions, emotions, and attitudes on various social and political issues. Typically memes combining images or videos with short text overlays convey messages in a compact and highly shareable format (Atanasov et al., 2019; Kiela et al., 2021). While many memes are created purely for entertainment, others are used to shape public opinions or to target individuals, organizations, or political groups. By combining visual indicators, textual cues, satire, and cultural references, memes can communicate complex political narratives in a concise manner. Particularly in political discourses, memes often comprise of symbolic imagery, party iconography, and culturally grounded humor to express support, criticism, or ideological positions. Such multimodal and context-dependent memes make the detection and the analysis of memes challenging (Farabi et al., 2024).

(a) Support for Individual (b) Support for Party

(c) Troll against Individual (d) Troll against Party

Figure 1: Multi-Level Political Meme Sample.

To advance research on political memes detection in Dravidian languages, DravidianLangTech@ACL 2026[1] (Rajiakodi et al., 2026) introduces a dataset and a Shared Task on Multi-Level Political Meme Classification. The task focuses on hierarchical two-level classification for Tamil and Malayalam political memes. Sample memes from the dataset are illustrated in Figure 1. The task is to identify (i) whether a meme expresses **Support** or **Troll** (Level 1), and (ii) whether the meme targets an **Individual** or a **Party** (Level 2).

In response to this task, we propose a hierarchical two-stage framework built on the Gemma 3

[1] https://www.codabench.org/competitions/11325/

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 273–278

4B Instruction model. The first stage predicts the stance of the meme, and its output is then used as contextual input for the second stage to identify the target type. This design explicitly models the dependency $P(y_2 \mid M, y_1)$, helping to reduce cross-level inconsistencies. Our approach achieves an average F1-scores of 0.8029 for Tamil and 0.6950 for Malayalam across both levels, ranking **1st** in Tamil and **4th** in Malayalam.

2 Related Work

Understanding political memes in Dravidian languages requires integrating insights from multimodal analysis, sarcasm detection, hierarchical classification, and Multimodal Large Language Models (MLLMs).

Early research on online toxicity and troll detection primarily relied on unimodal text-based natural language processing techniques. These studies formulated troll detection as a text classification task using sentiment analysis, n-gram features, and word embeddings to identify malicious or disruptive content in forums and social media posts (Fortuna and Nunes, 2018; Monakhov, 2020). While effective for standard textual data, such approaches often fail when applied to memes, where textual content may appear benign but becomes offensive or sarcastic when interpreted alongside visual context.

To overcome these limitations, subsequent work shifted toward multimodal approaches. The introduction of the Hateful Memes Challenge (Cai et al., 2019; Kiela et al., 2020) established an important benchmark for studying multimodal toxicity, demonstrating that benign visual and textual elements can combine to produce harmful meanings. Later studies explored cross-modal attention mechanisms and contrastive learning to capture complex interactions between images and embedded text in memes (Burbi et al., 2023; Arya and Bagwari, 2024).

Building on these advances, research extended to culturally rich and linguistically diverse Dravidian languages. Shared tasks at *DravidianLangTech* introduced datasets for Tamil and Malayalam troll meme classification (Suryawanshi et al., 2020; Suryawanshi and Chakravarthi, 2021; Premjith et al., 2022). Many approaches adopted early fusion strategies, combining Convolutional Neural Networks (CNNs) such as VGG16, VGG19, ResNet, and EfficientNet for visual feature extraction with transformer-based language models like mBERT, MuRIL, ALBERT, and XLNet to process multilingual or code-mixed text (Nandi et al., 2022; Das et al., 2022; Hariprasad et al., 2022; Krishna and Kumar, 2022; Hasan et al., 2022). Although these models improved meme classification performance, they generally treated the problem as a flat classification task without modeling dependencies between labels.

Recent advances in MLLMs have further improved performance on vision-language reasoning and meme understanding tasks (Liu et al., 2024; Team et al., 2024). However, most existing approaches still predict labels independently. In political meme analysis, this assumption can be limiting because the stance expressed in a meme often constrains its target. For example, a supportive meme may promote a political party or leader, whereas a trolling meme may criticize or attack a specific individual. Modeling such hierarchical relationships can therefore improve prediction consistency, as demonstrated in structured classification literature (Mohammad et al., 2016; Pramanick et al., 2021; Zhou et al., 2020; Chen et al., 2021).

Language	Level	Label	Train	Test
Tamil	1	Troll	691	175
		Support	112	26
	2	TAI	547	137
		TAP	146	37
		SFI	86	22
		SFP	24	5
Malayalam	1	Troll	477	96
		Support	23	4
	2	TAI	315	50
		TAP	110	31
		Inter	53	15
		SFI	12	4
		SFP	10	-

Table 1: Dataset statistics for Tamil and Malayalam meme images. Level-1 contains two labels: **Troll** and **Support**. Level-2 further categorizes these into Troll against Individual (**TAI**), Troll against Party (**TAP**), Support for Individual (**SFI**), Support for Party (**SFP**), and Intersection (**Inter**).

3 System Overview and Experimental Setups

This section describes the dataset, the task formulation, and the proposed hierarchical framework. We also outline the instruction-tuning strategy and

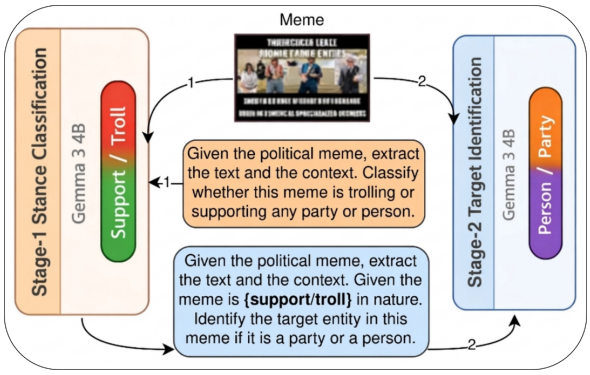

Figure 2: Two stage Hierarchical Framework for Multi-level Meme Classification

Component	Value
Base Model	gemma-3-4b-it
Fine-tuning	LoRA
LoRA Rank (r)	8
LoRA Alpha	16
Optimizer	AdamW
Learning Rate	5×10^{-5}
Scheduler	Cosine
Epochs	3
Per-device Batch Size	2
Gradient Accumulation	8
Effective Batch Size	16
Max Sequence Length	2048
Max Gradient Norm	1.0

Table 2: Training configuration for both models.

training configuration used for fine-tuning the multimodal model.

3.1 Dataset

We use the dataset released for the Shared Task on Multi-Level Political Meme Classification at DravidianLangTech 2026. The dataset consists of political meme images in two Dravidian languages: Tamil and Malayalam, annotated at two hierarchical levels.

Level 1 (Stance): Determines whether a meme expresses *Support* or *Troll* towards a political entity.

Level 2 (Target Identification): Identifies the type of political target referenced in the meme, such as an individual leader or a political party.

The Tamil dataset contains 1,003 memes (802 train, 201 test), while the Malayalam dataset contains 600 memes (500 train, 100 test). Detailed label distributions are presented in Table 1.

3.2 Hierarchical Two-Stage Framework

We implement a hierarchical two-stage framework based on the **Gemma 3 4B Instruction** model (Team et al., 2025), as illustrated in Figure 2. Gemma 3 is a multimodal large language model with strong vision–language understanding capabilities, making it suitable for analyzing memes that combine visual and textual cues. It employs a SigLIP vision encoder (Zhai et al., 2023) to process image inputs.

Instead of jointly predicting both levels, we fine-tune two specialized models that share the same pretrained backbone model but are optimized independently using instruction tuning. The first model (**Stage–1**) predicts the stance label, determining whether a meme expresses *support* or *troll*. The second model (**Stage–2**) predicts the target type, identifying whether the meme refers to an *individual* or a *party*. During the inference, the prediction from the Stage–1 model is passed as contextual input to the Stage–2 model, enabling to capture the hierarchical structure at the time of prediction.

Training instances follow an instruction-based format where the image meme and a task-specific prompt are provided as input, and the model generates a single-label response. We use the prompts described below to fine-tune the models for each stage. Unlike Tamil in the Malayalam dataset, an additional Level–2 target category namely *Intersection*, is present. To accommodate this label, the Stage–2 prompt is extended to include *Intersection* as an additional target entity.

Stage–1 Prompt

> `<image>`
> Given the political meme, analyze the visual and textual content. Classify whether the meme expresses trolling or support toward a political entity.
> Respond with only one label: “troll” or “support”.

Stage–2 Prompt

> `<image>`
> Given the political meme, analyze the visual and textual content. The meme expresses `[support/troll]`. Identify whether the target entity is a party or a person.
> Respond with only one label: “party” or “person”.

During training of the Stage–2 model, the gold Level–1 label is used as conditioning context, whereas during inference the predicted output from the Stage–1 model is supplied.

3.3 Training Configuration

We use the `Gemma 3 4B` instruction-tuned model (Team et al., 2025) as the backbone for both the stages. Both the models are fine-tuned using supervised instruction tuning with cross-entropy loss

Lang	Setups	Level-1				Level-2				Avg F1
		Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	
ta	Independent	0.9204	0.9147	0.9204	0.9159	0.7065	0.6754	0.7065	0.6885	0.8022
	Hierarchical	0.9204	0.9147	0.9204	0.9159	0.6915	0.6930	0.6915	0.6898	**0.8029**
ml	Independent	0.9400	0.9208	0.9400	0.9303	0.5000	0.4003	0.5000	0.4243	0.6773
	Hierarchical	0.9400	0.9208	0.9400	0.9303	0.4800	0.4462	0.4800	0.4596	**0.6950**

Table 3: Performance of different training setups in terms of Accuracy (**Acc**), Precision (**Prec**), Recall (**Rec**), and F1-score (**F1**) for each level on Tamil (**ta**) and Malayalam (**ml**). **Avg F1** denotes the average F1-score across both.

and *LoRA-based parameter-efficient fine-tuning*. Low-rank adapters are inserted into the backbone while the vision encoder and multimodal projection layers remain frozen, reducing the number of trainable parameters while preserving pretrained multimodal representations.

Training is conducted using the LLaMA-Factory framework[2] for efficient multimodal instruction tuning. Mixed-precision training (`bf16`) is employed together with the AdamW optimizer, cosine learning rate scheduling, and gradient accumulation to improve computational efficiency. The detailed training configuration is summarized in Table 2. All implementation details, training scripts, and prompts used in this work are publicly available in the GitHub repository[3].

4 Results and Evaluation

We evaluate our proposed hierarchical framework on the official test set of the shared task. To analyze the impact of hierarchical modeling, we compare it with an **Independent** training setup using the same `Gemma-3-4B-it` backbone.

Independent: Two separate models are instruction fine-tuned independently for each task level: one for Level-1 stance classification and another for Level-2 target identification.

Hierarchical: Our proposed method follows a two-stage framework where Model–1 predicts the stance label (support or troll), and the predicted stance is used as contextual input for Model–2 to identify the target type (person or party). To ensure a fair comparison, we reuse the Level-1 model from the Independent setup and only train the Stage–2 model to incorporate the stance information predicted by Stage–1.

Table 3 presents the evaluation results for both Tamil and Malayalam datasets. From the results, we observe that Level-1 stance classification achieves consistently high performance across both languages. For Tamil we achieve an F1-score of 0.9159, while for Malayalam the models achieve an F1-score of 0.9303. This indicates that determining whether a meme expresses support or trolling is relatively easier for the model compared to identifying the specific target.

In contrast, Level-2 target identification is more challenging. For Tamil, the Independent setup achieves an F1-score of 0.6885, while the Hierarchical model slightly improves the performance to 0.6898. Although the improvement is modest, the hierarchical setup increases the precision from 0.6754 to 0.6930, suggesting that incorporating stance information helps the model make more accurate target predictions. A similar trend is observed for the Malayalam dataset. While the Independent model achieves an F1-score of 0.4243 for Level-2, the Hierarchical framework improves the F1-score to 0.4596 along with an increase in precision from 0.4003 to 0.4462. This indicates that conditioning the target prediction on the predicted stance helps the model better identify the relevant political entity.

4.1 Error Analysis

We perform a detailed error analysis of misclassified samples across both levels of the task. For Level-1 (stance classification), the model errors are primarily attributed to the presence of sarcastic or ironic textual content, as illustrated in Figure 3(a). Such cases often convey implicit meaning that contradicts the literal interpretation of the text. Additionally, memes containing ambiguous or mixed signals—where both supportive and trolling elements are present within the same instance—pose significant challenges to the model. Examples of such cases are shown in Figure 3(b) and (c), where

[2] https://github.com/hiyouga/LlamaFactory

[3] https://github.com/saurabhdbz/IndiLangTech-2026

(a) Misclassified as Troll (b) Misclassified as Troll (c) Misclassified as Support (d) Misclassified as Troll Against Party

Figure 3: Examples of misclassified memes highlighting challenges such as sarcasm and ambiguous intent.

the model struggles to resolve conflicting cues and assigns incorrect labels.

At Level-2 (target identification), errors mainly arise when both a political party and an individual are referenced within the same meme. In such cases, the model finds it difficult to determine the primary target entity, leading to incorrect predictions. An example of this ambiguity is shown in Figure 3(d), where the meme references both the individual (*Vijay*) and his political party (*TVK*). The model incorrectly predicts the target as a party and classifies the meme as trolling against the party, highlighting its difficulty in distinguishing the dominant target in multi-entity contexts.

5 Conclusion

This paper presented a hierarchical two-stage framework for multi-level political meme classification in Tamil and Malayalam using the Gemma 3 4B Instruction multimodal model. The proposed approach decomposes the task into stance detection and target identification, where the prediction from the first stage guides the second stage, enabling structured modeling of dependencies between the two levels. Experimental results demonstrate the effectiveness of this framework, achieving average F1-scores of 0.8029 for Tamil and 0.6950 for Malayalam, and securing the 1st rank for Tamil and 4th for Malayalam in the shared task.

Our analysis further reveals key challenges in political meme understanding, including sarcasm, implicit meaning, and ambiguity arising from mixed signals or multiple referenced entities. In particular, the model struggles when both a political figure and a party are present, highlighting limitations in resolving dominant targets. These findings suggest that incorporating improved multimodal reasoning, entity grounding, and context-aware interpretation could further enhance performance in such complex settings.

References

Gaurav Arya and Ashutosh Bagwari. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12:22359–22375.

Atanas Atanasov, Gianmarco De Francisci Morales, and Preslav Nakov. 2019. Predicting the role of political trolls in social media. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1023–1034.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2014–2024.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangong Jiang. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4370–4379.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. hate-alert@ dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 51–57.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8020–8028. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Shruthi Hariprasad, Sarika Esackimuthu, Saritha Madhavan, Rajalakshmi Sivanaiah, and 1 others. 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.

Md Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Cuetnlp@ dravidianlangtech-acl2022: Investigating deep learning techniques to detect multimodal troll memes. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 170–176.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, and 1 others. 2021. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*, 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.

Achyuta Krishna and Mithun Kumar. 2022. Troll meme classification using feature extraction and transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Sergei Monakhov. 2020. Understanding troll writing as a linguistic phenomenon. In *Proceedings of SAI Intelligent Systems Conference*, pages 315–334. Springer.

Rabindra Nath Nandi, Firoj Alam, and Preslav Nakov. 2022. Teamx@ dravidianlangtech-acl2022: A comparative analysis for troll-based meme classification. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 79–85.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Preslav Nakov, Pawan Goyal, Animesh Bhattacharya, Md. Arid Hasan, Niloy Bhattacharya, and Sourya Das. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796. Association for Computational Linguistics.

B Premjith, Bharathi Raja Chakravarthi, Malliga Subramanian, B Bharathi, Soman Kp, V Dhanalakshmi, K Sreelakshmi, Arunaggiri Pandian, and Prasanna 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.

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.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Thomas Mesnard, Cassidy Hardin, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Jie Zhou, Chunping Ma, Dingkun Long, and 1 others. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117.