byteSizedLLM@DravidianLangTech 2025: Multimodal Misogyny Meme Detection in Low-Resource Dravidian Languages Using Transliteration-Aware XLM-RoBERTa, ResNet-50, and Attention-BiLSTM

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Abstract

Detecting misogyny in memes is challenging due to their multimodal nature, especially in low-resource languages like Tamil and Malayalam. This paper presents our work in the Misogyny Meme Detection task, utilizing both textual and visual features. We propose an Attention-Driven BiLSTM-XLM-RoBERTa-ResNet model, combining a transliteration-aware fine-tuned XLM-RoBERTa for text analysis and ResNet-50 for image feature extraction. Our model achieved Macro-F1 scores of 0.8805 for Malayalam and 0.8081 for Tamil, demonstrating competitive performance. However, challenges such as class imbalance and domain-specific image representation persist. Our findings highlight the need for better dataset curation, task-specific fine-tuning, and advanced fusion techniques to enhance multimodal hate speech detection in Dravidian languages.

1 Introduction

The proliferation of social media has transformed communication but has also led to the rise of harmful content, including misogynistic memes that combine text and visuals to convey discriminatory messages. Detecting such content is challenging due to its multimodal nature, implicit messaging, and linguistic diversity. Developing robust systems to identify and mitigate misogynistic memes is essential for fostering safer online spaces.

The Shared Task on Misogyny Meme Detection, part of DravidianLangTech@NAACL 2025[1], addresses this issue by focusing on memes in Tamil and Malayalam, two Dravidian languages with complex morphologies and distinct scripts. Participants are tasked with designing multimodal systems capable of analyzing both textual and visual components to classify memes as Misogynistic or Non-misogynistic. The challenges include handling transliterated text and capturing cultural nuances in linguistic expressions.

This task underscores the importance of multilingual and multimodal approaches in misogyny detection, particularly for Tamil and Malayalam, emphasizing culturally sensitive solutions in low-resource settings. Annotated social media datasets enable effective text and image processing, with a transliteration-aware fine-tuned XLM-RoBERTa-base handling textual content and ResNet-50 extracting visual features. This baseline serves as a foundation for exploring advanced architectures that integrate contextual information, with macro F1 score ensuring balanced evaluation across classes.

In this paper, we present our methodology, experimental setup, and results, demonstrating the effectiveness of our hybrid model in addressing the challenges of misogyny meme detection. We also discuss key challenges, such as transliteration, cultural nuances, and data sparsity, and propose directions for future research to enhance multilingual and multimodal misogyny detection.

2 Related Work

Advancements in multimodal image-text analysis have driven progress in hate speech detection, particularly with social media content. Early models like MOMENTA (Pramanick et al., 2021) and HateCLIPper (Kumar and Nandakumar, 2022) leveraged CLIP's vision-language encoders for cross-modal interactions, while newer methods refine alignment through textual inversion and image captioning.

MemeCLIP (Shah et al., 2024) directly utilizes CLIP's pre-trained encoders for meme processing, tackling data scarcity and class imbalance with Feature Adapters and a cosine classifier to enhance robustness.

[1] https://codalab.lisn.upsaclay.fr/competitions/20856

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 206–211

DravidianLangTech shared tasks highlight challenges in processing Tamil and Malayalam, especially with transliteration and code-mixing. Codewithzichao@DravidianLangTech-EACL2021 (Suryawanshi and Chakravarthi, 2021) employed XLM-RoBERTa and multilingual BERT for offensive language detection in Tamil, Malayalam, and Kannada, achieving strong F1 scores despite class imbalance (Li, 2021). Similarly, BPHC@DravidianLangTech-ACL2022 (V et al., 2022) focused on troll meme classification in Tamil-English code-mixed text, where MuRIL achieved a weighted F1 score of 0.74 (B et al., 2022).

Our work builds on these efforts, addressing misogyny meme detection in Tamil and Malayalam. We enhance CLIP's vision-language capabilities while tackling transliteration, code-mixing, and data sparsity, advancing multimodal analysis for Dravidian languages.

Label	Train	Dev	Test	Total
0	381	97	122	600
1	259	63	78	400
Total	640	160	200	1,000

Table 1: Statistics of the Malayalam Dataset for Misogynistic(1) and Non-Misogynistic(0) Classification

3 Dataset

The dataset for this task, provided as part of the Misogyny Meme Detection - DravidianLangTech@NAACL 2025 shared task(Ponnusamy et al., 2024), consists of multimodal memes in Malayalam and Tamil, annotated as misogynistic (1) or non-misogynistic (0). Each sample includes an image (JPG format) and transcribed text, with data split into train, development (dev), and test sets.

Table 1 and Table 2 present the data distribution for Malayalam and Tamil, respectively. This dataset benchmarks misogyny detection in low-resource languages, addressing challenges such as transliteration, code-mixing, and limited annotated data, fostering advancements in multilingual and multimodal learning.

4 Methodology

The proposed methodology utilizes a multimodal architecture to effectively handle textual and visual features for misogyny meme detection. The model combines the strengths of a transliteration-aware fine-tuned XLM-RoBERTa for text, ResNet-50 for image processing, and an attention-driven BiLSTM for multimodal feature fusion and classification.

Label	Train	Dev	Test	Total
0	851	210	267	1,328
1	285	74	89	448
Total	1,136	284	356	1,776

Table 2: Statistics of the Tamil Dataset for Misogynistic(1) and Non-Misogynistic(0) Classification

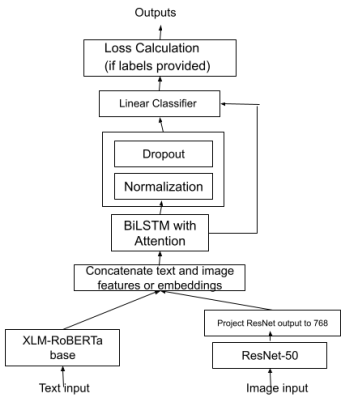

Figure 1: Architecture of the Attention-Driven BiLSTM-XLM-RoBERTa Classifier

4.1 XLM-RoBERTa Fine-Tuning with Transliteration Awareness

The XLM-RoBERTa Base model(Conneau et al., 2019) was fine-tuned on small portion of Tamil and Malayalam text from AI4Bharath(Kunchukuttan et al., 2020), achieving a perplexity of 4.9 for Tamil and 4.1 for Malayalam. To handle transliterated and mixed-script text, the IndicTrans tool(Bhat et al., 2015) was used to create three variations: native script, fully Romanized, and partially transliterated text. This preprocessing enhances text representation for diverse script inputs. The CLS token output provides a 768-dimensional embedding[2][3].

4.2 ResNet-50 for Image Feature Extraction

The visual component of the memes is handled using ResNet-50(He et al., 2016), a widely used convolutional neural network pre-trained on ImageNet. To align the visual features with the textual features, the fully connected (FC) layer of ResNet-50 is modified to project the extracted image features

[2] https://huggingface.co/bytesizedllm/TamilXLM_Roberta

[3] https://huggingface.co/bytesizedllm/MalayalamXLM_Roberta

into a 768-dimensional space. This modification ensures compatibility and seamless integration of textual and visual embeddings in the later stages of the model.

4.3 Attention-BiLSTM-XLM-RoBERTa-ResNet Classifier

We propose a hybrid Attention-Driven BiLSTM-XLM-RoBERTa-ResNet model for multimodal misogyny detection, inspired by our previous research (Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b). This architecture integrates textual and visual features to capture both linguistic and image-based patterns.

The text input is processed using a fine-tuned XLM-RoBERTa, extracting contextual embeddings:

$$\mathbf{X}_t = \textbf{XLM-RoBERTa}(input\_ids, atten\_mask) \quad (1)$$

The image input is processed using ResNet-50 to extract deep visual features:

$$\mathbf{X}_i = \textbf{ResNet-50}(image\_features) \quad (2)$$

These features are concatenated (or element-wise added, averaged, attention-weighted, etc.) and passed through a BiLSTM layer (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) to capture sequential dependencies.

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (3)$$

An attention mechanism enhances key information:

$$\alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \quad (4)$$

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (5)$$

A fully connected layer classifies the output after layer normalization and dropout:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (6)$$

The model is optimized using cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (7)$$

Figure 1 illustrates the model architecture, showcasing the integration of XLM-RoBERTa, ResNet-50, and BiLSTM with attention for enhanced multimodal classification.

5 Experimental Setup

This section describes the experimental setup for the proposed multimodal architecture, integrating XLM-RoBERTa for text, ResNet-50 for images, and BiLSTM with attention for feature fusion.

5.1 Text Processing

XLM-RoBERTa Base is fine-tuned on Tamil and Malayalam text to handle linguistic diversity in social media. IndicTrans preprocesses text into three formats: native script, fully Romanized, and partially transliterated (20–70%). Tokenization is performed with a max sequence length of 128, applying padding and truncation for batch uniformity.

5.2 Image Processing

ResNet-50, pre-trained on ImageNet, extracts visual features. The final layer is replaced with a projection layer to align 768-dimensional text embeddings. Images are resized to 224×224 pixels and normalized for consistency.

5.3 Multimodal Feature Fusion

Textual and visual embeddings are fused into a single tensor and processed via a BiLSTM, capturing cross-modal dependencies in a 512-dimensional space. An attention mechanism refines feature relevance before classification.

5.4 Training Configuration

The model is trained using AdamW with a learning rate of (learning rate 2×10^{-5} and a weight decay of 0.01, using a batch size of 16 for up to 5 epochs with early stopping. Cross-entropy loss is used for classification, and early stopping is applied based on the validation macro F1 score. Gradient clipping with a maximum norm of 1.0 ensures training stability.

6 Results and Discussion

Evaluation is based on macro F1, with accuracy and classification reports. The best macro F1 model is saved for testing. Our unique model setup for the shared task yielded notable results for both Tamil and Malayalam. For Malayalam, our second run[4] achieved a Macro F1 score of 0.8805, securing the highest score in the competition, surpassing the first-ranked team, CUET_Novice, which scored 0.8763. The best Macro F1 was obtained in our

[4] https://github.com/mdp0999/Misogyny-Meme-Detection/blob/main/test2_ml.ipynb

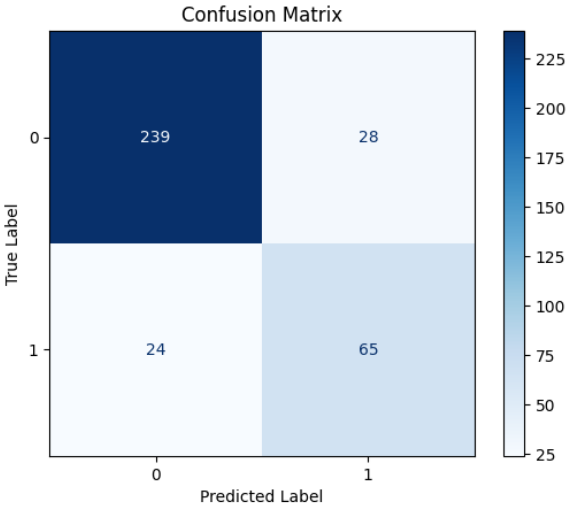

Figure 2: Confusion matrix for Task 1: Misogyny Meme Detection in Tamil.

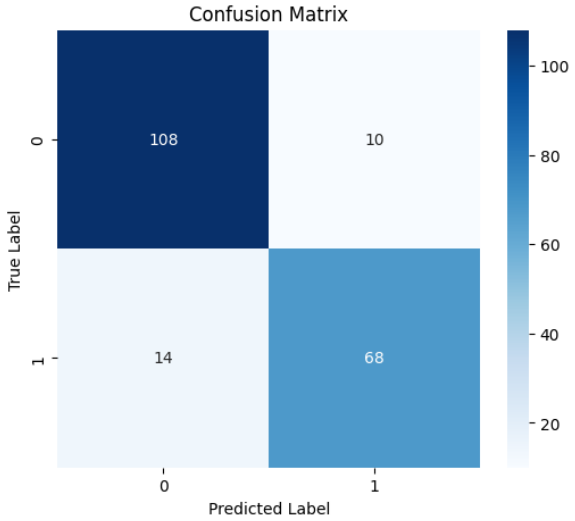

Figure 3: Confusion matrix for Task 2: Misogyny Meme Detection in Malayalam.

second run using a learning rate of 2×10^{-5}, while our first run, with a same learning rate and a customized transformer encoder-decoder architecture, resulted in a slightly lower score of 0.8391, as reflected in the task results.

In contrast, our Tamil model achieved a Macro F1 score of 0.8081, securing third place in the competition. The top score of 0.8368 for Tamil was achieved by team DLRG_RR. While our model performed well for the non-misogynistic class, its recall for the misogynistic class was lower, indicating challenges in capturing nuanced patterns associated with this class. These results suggest that class imbalance and limited training data may have hindered the model's ability to generalize effectively for Tamil.

The performance gap between Malayalam and Tamil suggests that class imbalance and script variations influenced misclassification. While the Malayalam model achieved high precision and recall across classes, the Tamil model struggled with lower recall for the misogynistic class, indicating difficulty in capturing nuanced linguistic patterns. The confusion matrices (Figures 2 and 3) highlight these challenges, emphasizing the need for better handling of class imbalance in Tamil and further refinement of feature extraction in both languages.

7 Limitations and Future Work

The primary limitation of our work was the restricted size of the training dataset due to computational constraints. This likely affected the model's ability to capture complex patterns, especially for the Tamil task. Additionally, the ResNet-50 architecture used for meme analysis was not fine-tuned for extracting features specific to misogyny memes, which may have limited its performance in visual understanding.

Future work will focus on addressing these limitations by training models on larger datasets to improve generalization and exploring meme-specific architectures for enhanced feature extraction. Furthermore, techniques to handle class imbalances more effectively will be incorporated to boost recall for minority classes. These advancements are expected to improve misogyny meme detection performance across multiple languages.

8 Conclusion

This study presents our approach to misogyny meme detection for Tamil and Malayalam languages, demonstrating strong performance for Malayalam with a top Macro F1 score of 0.8805 and competitive results for Tamil with a Macro F1 score of 0.8081. The findings emphasize the importance of addressing class imbalances, increasing data availability, and fine-tuning models for task-specific visual features. Despite its limitations, this work provides a robust foundation for future research and development in misogyny meme detection tasks. Our team, **byteSizedLLM**, remains committed to advancing solutions for such challenging multimodal tasks in low-resource languages.

References

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, 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, Dublin, Ireland. Association for Computational Linguistics.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 164–168, Kyiv. Association for Computational Linguistics.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *Preprint*, arXiv:2409.14703.

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, Kyiv. Association for Computational Linguistics.

Achyuta V, Mithun Kumar S R, Aruna Malapati, and Lov Kumar. 2022. BPHC@DravidianLangTech-ACL2022-a comparative analysis of classical and pre-trained models for troll meme classification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 151–157, Dublin, Ireland. Association for Computational Linguistics.