Fired_from_NLP@DravidianLangTech 2025: A Multimodal Approach for Detecting Misogynistic Content in Tamil and Malayalam Memes

Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto,
Jidan Al Abrar, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{064, 055, 049, 080}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

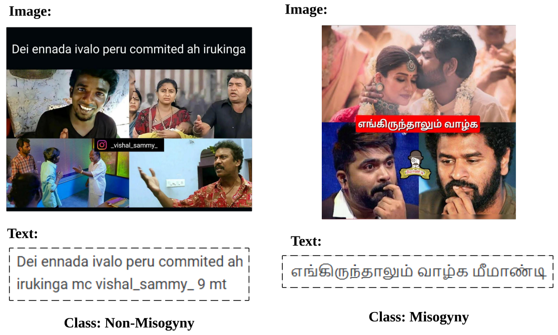
Abstract

In the context of online platforms, identifying misogynistic content in memes is crucial for maintaining a safe and respectful environment. While most research has focused on high-resource languages, there is limited work on languages like Tamil and Malayalam. To address this gap, we have participated in the Misogyny Meme Detection task organized by DravidianLangTech@NAACL 2025, utilizing the provided dataset named MDMD (Misogyny Detection Meme Dataset), which consists of Tamil and Malayalam memes. In this paper, we have proposed a multimodal approach combining visual and textual features to detect misogynistic content. Through a comparative analysis of different model configurations, combining various deep learning-based CNN architectures and transformer-based models, we have developed fine-tuned multimodal models that effectively identify misogynistic memes in Tamil and Malayalam. We have achieved an F1 score of 0.678 for Tamil memes and 0.803 for Malayalam memes.

1 Introduction

The rapid proliferation of social media has enabled the widespread sharing of memes, which are often used to express humor, ideas, or opinions. However, this medium is also increasingly being misused to propagate harmful ideologies, including misogyny. Therefore, detecting misogynistic content in memes has become essential for mitigating hate speech and ensuring online safety.

Many works have been done on harmful meme detection (Sharma et al., 2022), (Lin et al., 2024), (Gu et al., 2024), (Pramanick et al., 2021), but only a limited number of studies have specifically focused on misogyny meme detection (Srivastava, 2022), (Fersini et al., 2019), (Habash et al., 2022). Most of the existing research in misogyny detection has concentrated on high-resource languages like English, Hindi, and Arabic (Singh et al., 2024), (Srivastava, 2022), (Mulki and Ghanem, 2021), (Mahdaouy et al., 2022), leveraging large-scale datasets and advanced techniques and models. However, research in low-resource languages such as Tamil and Malayalam has been scarce (Rajalakshmi et al., 2023), (Ghanghor et al., 2021), (Chakravarthi et al., 2024), leaving a significant gap in addressing this issue in multilingual and diverse online communities.



Dei ennada ivalo peru commited ah irukinga mc vishal_sammy_ 9 mt

Class: Non-Misogyny

எங்கிருந்தாலும் வாழ்க மீமாண்டி

Class: Misogyny

Figure 1: Example of a misogynistic and a non-misogynistic meme in Tamil

To address this gap, the task of Misogyny Meme Detection was introduced as part of DravidianLangTech@NAACL 2025. For this task, the organizers have provided a dataset named MDMD (Misogyny Detection Meme Dataset) for memes in the Tamil and the Malayalam languages (Ponnusamy et al., 2024), consisting of both misogynistic and non-misogynistic memes. The details of this shared task and its findings have been thoroughly presented in the overview paper (Chakravarthi et al., 2025).

Our objective has been to develop a multimodal model that effectively detects misogynistic memes in Tamil and Malayalam by combining both visual and textual elements. To achieve this, we have employed various CNN-based architectures and transformer-based models and conducted a comparative analysis of different multimodal model configurations.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 537–542
May 3, 2025 ©2025 Association for Computational Linguistics

Our main contributions are as follows:

- We have developed fine-tuned multimodal models that can effectively detect misogynistic memes in Tamil and Malayalam.
- We have conducted a comparative analysis of various model configurations, combining different transformer-based models with CNN backbones.

The implementation details are available in this GitHub repository[1].

2 Related Work

Recent research has focused on multimodal approaches for detecting harmful content in memes, particularly misogynistic and offensive memes.

Several studies have proposed frameworks that fuse both text and image features to improve detection accuracy. For instance, the MISTRA framework has been developed by utilizing variational autoencoders for dimensionality reduction of image features and combining them with text embeddings to detect misogynous memes (Jindal et al., 2024). In (Pramanick et al., 2021), the authors have introduced MOMENTA, a multimodal deep neural network that analyzes both global and local perspectives within memes to detect harmful content. Additionally, a large-scale Hindi-English code-mixed dataset has been introduced in (Singh et al., 2024), focusing on misogynous meme detection using multimodal fusion methods.

The authors in (Gu et al., 2024) have proposed the SCARE framework, which addresses multimodal alignment by maximizing the mutual information between image and text features while enhancing intra-modal representation learning. Furthermore, in (Habash et al., 2022), an ensemble of models has been utilized by combining multiple multimodal deep learning models for detecting misogynous content.

In the field of multilingual meme detection, the DravidianLangTech-2022 shared task in (Das et al., 2022) has explored meme detection in Tamil, showing that fusing text-based and image-based models improves performance for troll meme classification. The authors in (Ghanghor et al., 2021) have focused on offensive language identification and troll meme classification in multiple Dravidian languages. Moreover, in (Chakravarthi et al., 2024), an overview of the first shared task on 'Multitask Meme Classification - Unraveling Misogynistic and Troll Memes in Online Memes' has been presented, focusing on Tamil and Malayalam memes.

3 Dataset

The Misogyny Meme Detection task of DravidianLangTech@NAACL 2025 consisted of two subtasks: one for the Tamil language and the other for the Malayalam language. We were provided with the MDMD dataset, which contains memes and text transcriptions for each language, annotated as either misogynistic or non-misogynistic (Ponnusamy et al., 2024).

Tamil Dataset			
Category	**Train**	**Dev**	**Test**
Non-Misogyny	851	210	267
Misogyny	285	74	89
Total	1136	284	356

Table 1: Dataset distribution for Tamil memes.

Malayalam Dataset			
Category	**Train**	**Dev**	**Test**
Non-Misogyny	381	97	122
Misogyny	259	63	78
Total	640	160	200

Table 2: Dataset distribution for Malayalam memes.

Tables 1 and 2 present the dataset distribution for the Tamil and Malayalam languages, respectively. The Tamil language consisted of 1,336 training samples, 284 validation samples, and 356 test samples, while the Malayalam consisted of 640 training samples, 160 validation samples, and 200 test samples.

We can see that the dataset is highly imbalanced, with non-misogynistic memes significantly outnumbering misogynistic ones. Additionally, many text transcriptions have contained code-mixed text, combining English with Tamil or Malayalam. The images have also included redundant elements such as social media logos, profile names, and icons.

4 Methodology

This section presents our approach for misogyny meme detection in Tamil and Malayalam. The methodology consists of four main components: input modalities, preprocessing, feature extraction, and cross-modal attention and fusion. Figure 2 summarizes the model architecture.

[1] https://github.com/Sajid064/Misogyny-Meme-Detection

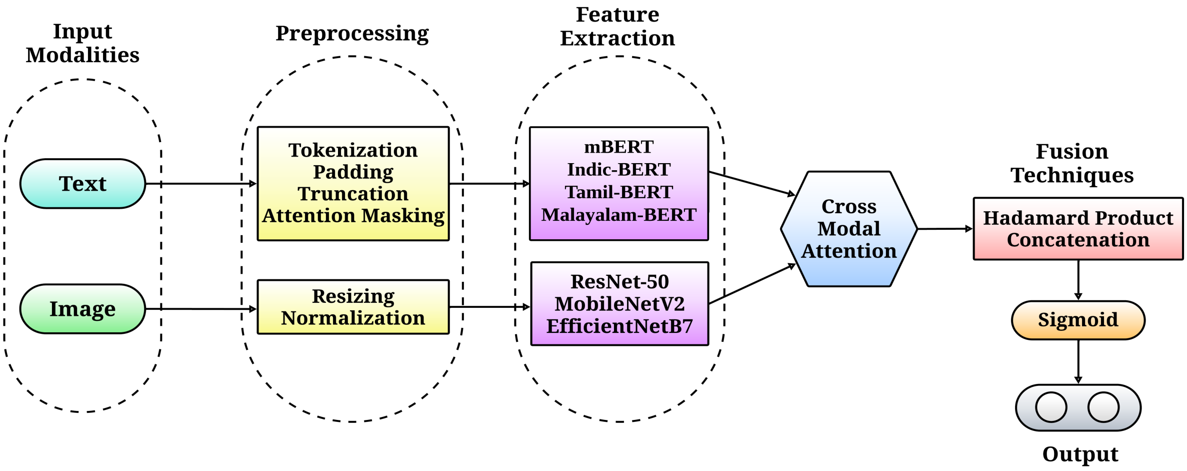


Figure 2: Model architecture of our proposed multimodal approach for misogynistic meme detection

4.1 Input Modalities

We have utilized two primary input modalities:

- **Text Modality:** Textual data has been obtained from transcriptions and processed using a transformer-based language model.
- **Image Modality:** Images corresponding to the textual descriptions have been processed using a deep convolutional neural network (CNN) backbone.

4.2 Preprocessing

4.2.1 Text Processing

The text data have been pre-processed using the BERT tokenizer to convert raw text into input token sequences. We have used padding to ensure that all sequences are of uniform length. We have truncated the sequences if they exceed the maximum length. Then, we applied attention masking to distinguish real tokens from padding tokens. The processed tokens, including the attention mask, have then been fed into a pre-trained Tamil BERT model.

4.2.2 Image Processing

The images have been resized to 224×224 pixels for consistency and to match the input requirements and normalized for faster convergence.

4.3 Feature Extraction

4.3.1 Text Feature Extraction

We have utilized multiple transformer-based models for text feature extraction. For both Tamil and Malayalam datasets, we have employed two general-purpose multilingual models (mBERT and Indic-BERT) to ensure robust multilingual representations. Additionally, we have used two language-specific BERT models: Tamil-BERT for Tamil texts and Malayalam-BERT for Malayalam texts. Each input text has been tokenized and passed through the transformer-based models, and we have extracted the `pooler_output` representation from the final transformer layer.

4.3.2 Image Feature Extraction

For image-based feature extraction, we have experimented with multiple deep CNN architectures, including ResNet50, MobileNetV2, and EfficientNetB7. These models have been initialized with ImageNet pre-trained weights, and their fully connected layers have been removed to obtain meaningful feature representations.

4.4 Cross-Modal Attention and Fusion

To effectively combine textual and visual information, a cross-modal attention mechanism has been applied so that the model can focus on the most relevant aspects of both modalities by computing attention scores between text and image features. For fusion, we have employed both the Hadamard product and concatenation techniques. The Hadamard product has been used for element-wise interaction between the attended image and text features, and the concatenated representation has been used to preserve distinct modality-specific characteristics. Finally, the fused features have then been passed through dense layers for final classification.

5 Experimental Setup

The parameter setups for our multimodal model are displayed in Table 3.

Parameter	Value
Optimizer	Adam
Loss Function	Binary Crossentropy
Learning Rate	$1e^{-4}$
Learning Rate Scheduler	Factor: 0.5
	Patience: 3
	Min lr: $1e^{-7}$
Early Stopping	Patience: 10
Batch Size	8
Epochs	100

Table 3: Training Parameter Settings

6 Experimental Findings

In this section, we have provided the experimental results of our proposed model. Table 4 shows a comparative analysis of different model configurations, combining various BERT variants with CNN backbones by evaluating the Micro-F1 score on the test samples of the Tamil (TAM) and Malayalam (MAL) datasets.

BERT Variants	CNN Backbone	F1 Score	
		TAM	MAL
mBERT	ResNet50	0.621	0.710
	MobileNetV2	0.596	0.681
	EfficientNetB7	0.644	0.742
Indic-BERT	ResNet50	0.573	0.717
	MobileNetV2	0.566	0.694
	EfficientNetB7	0.598	0.728
Tamil-BERT	ResNet50	0.647	-
	MobileNetV2	0.658	-
	EfficientNetB7	**0.678**	-
Malayalam -BERT	ResNet50	-	0.794
	MobileNetV2	-	0.773
	EfficientNetB7	-	**0.803**

Table 4: Performance comparison of different models using various BERT variants and CNN backbones

Among the general-purpose multilingual models, we have observed that the mBERT model consistently outperforms the Indic-BERT model across all the CNN backbones used (ResNet50, MobileNetV2, and EfficientNetB7). The combination of mBERT and EfficientNetB7 has achieved the highest F1 score of 0.644 for Tamil memes and 0.742 for Malayalam memes. In contrast, Indic-BERT with EfficientNetB7 has obtained a lower F1 score of 0.598 for Tamil and 0.728 for Malayalam.

For language-specific models, Tamil-BERT paired with EfficientNetB7 has demonstrated superior performance for Tamil memes by achieving the highest F1 score of 0.678 and surpassing all multilingual models. Similarly, Malayalam-BERT with EfficientNetB7 has achieved the highest F1 score of 0.803 for Malayalam memes by outperforming other configurations. These results indicate that while multilingual models like mBERT have performed well, language-specific models fine-tuned on their respective languages have yielded better results.

7 Error Analysis

From Table 1 and 2, we have observed that the distribution of Tamil memes is highly imbalanced, with a significantly larger number of non-misogynistic samples compared to misogynistic samples. This has led to lower F1 scores as our model has struggled with the minority class. In contrast, the class distribution of Malayalam memes is slightly more balanced, leading to comparatively improved performance. Additionally, the overall dataset size is quite limited, which has restricted the model's ability to generalize effectively. Another major challenge has been the presence of code-mixed text, where many transcriptions have been in English-written Tamil/Malayalam or a combination of English and Tamil/Malayalam words. This has made it harder for BERT models to extract proper features. Furthermore, a significant number of images in the dataset contained redundant elements such as social media icons, profile names, and profile photos, which have introduced noise into the learning process. These distractions have also contributed to some misclassifications.

8 Conclusion

In this paper, we have developed fine-tuned multimodal models for the detection of misogynistic memes in Tamil and Malayalam. Through a comparative analysis of various model configurations, combining transformer-based models with CNN backbones, we have found that language-specific BERT models combined with powerful CNN architectures, such as EfficientNet, achieved the highest results for both languages. In the future, we plan to experiment with more advanced models, such as Vision Transformers (ViT), and explore techniques to mitigate the dataset imbalance issue for enhanced performance.

9 Limitations

While our proposed approach has shown promising results, certain limitations have remained. The availability of labeled data has been limited, which has impacted the ability of our model to generalize effectively to unseen instances. Additionally, the approach has not explicitly accounted for the cultural and linguistic subtleties of the Tamil and Malayalam languages, which may have influenced classification accuracy. Moreover, pre-trained models have inherited biases from their training data, and the multimodal fusion process has faced challenges in capturing implicit or sarcastic expressions of misogyny.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

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.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Tianlong Gu, Mingfeng Feng, Xuan Feng, and Xuemin Wang. 2024. Scare: A novel framework to enhance chinese harmful memes detection. *IEEE Transactions on Affective Computing*, pages 1–14.

Mohammad Habash, Yahya Daqour, Malak Abdullah, and Mahmoud Al-Ayyoub. 2022. YMAI at SemEval-2022 task 5: Detecting misogyny in memes using VisualBERT and MMBT MultiModal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784, Seattle, United States. Association for Computational Linguistics.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 2359–2370.

Abdelkader El Mahdaouy, Abdellah El Mekki, Ahmed Oumar, Hajar Mousannif, and Ismail Berrada. 2022. Deep multi-task models for misogyny identification and categorization on arabic social media. *Preprint*, arXiv:2206.08407.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*.

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.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M.

2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech Language*, 78:101464.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar I. Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *International Joint Conference on Artificial Intelligence*.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Harshvardhan Srivastava. 2022. Misogynistic meme detection using early fusion model with graph network. *ArXiv*, abs/2203.16781.