One_by_zero@DravidianLangTech 2025: A Multimodal Approach for Misogyny Meme Detection in Malayalam Leveraging Visual and Textual Features

Dola Chakraborty*, **Shamima Afroz Mithi***
Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904012, u1904106, u1704039}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Misogyny memes are a form of online content that spreads harmful and damaging ideas about women. By combining images and text, they often aim to mock, disrespect, or insult women, sometimes overtly and other times in more subtle, insidious ways. Detecting Misogyny memes is crucial for fostering safer and more respectful online communities. While extensive research has been conducted on high-resource languages (HRLs) like English, low-resource languages (LRLs) such as Dravidian (e.g. Malayalam) remain largely overlooked. The shared task on Misogyny Meme Detection, organized as part of DravidianLangTech@NAACL 2025, provided a platform to tackle the challenge of identifying misogynistic content in memes, specifically in Malayalam. We participated in the competition and adopted a multimodal approach to contribute to this effort. For image analysis, we employed a ResNet18 model to extract visual features, while for text analysis, we utilized the IndicBERT model. Our system achieved an impressive F1-score of 0.87, earning us the 3rd rank in the task.

1 Introduction

Misogynistic memes have a significant influence as they contribute to normalizing and perpetuating harmful attitudes and behaviors (Paciello et al., 2021). These memes use social media's visual-textual and viral qualities to quietly embed and disseminate sexist beliefs, frequently making identification and intervention challenging. Detecting misogyny in memes poses unique challenges due to their multimodal nature. Sometimes text and image individually exhibit no offense, but combining both elements can convey implicit or contextual misogyny, making the meme offensive when taken as a whole (Chen et al., 2024; Gasparini et al., 2022). While substantial research has been conducted on misogyny detection in high-resource languages like English (Fersini et al., 2022), low-resource languages, particularly Dravidian languages such as Malayalam, remain underexplored. Malayalam, a Dravidian language spoken predominantly in the Indian state of Kerala, faces a growing issue of misogynistic memes on digital platforms, highlighting the need for targeted research.

Misogynistic memes in Malayalam often present additional challenges due to linguistic nuances, transliterated text (Malayalam written in English script), and the interplay of regional cultural references. The issue is made worse by the dearth of extensively annotated datasets in Malayalam, which makes it more difficult to create reliable detection methods. As participants in this shared task, our work makes the following notable contributions:

- Proposed a transformer-based approach to classify Malayalam misogynistic memes as *Misogynistic (Miso)* or *Non-misogynistic (NMiso)*.
- Experimented with several DL, transformer-based models to extract visual and textual features and employed late fusion to combine features from both modalities to detect misogynistic memes.

2 Related Work

Shushkevich and Cardiff (2018) analyzed tweets from Twitter for the Automatic Misogyny Identification (AMI) task at EVALITA 2018 and achieved an F1 score of 0.78 using Logistic Regression (LR) Devi and Saharia (2021) focused on misogyny detection in English, classifying texts as misogynous or not. They achieved 93.43(%) accuracy using a Bi-LSTM model. Goenaga et al. (2018) was part of the AMI-IberEval 2018 competition, identifying misogyny in English and Spanish tweets. They used a Bi-LSTM with CRF, achiev-

*Authors contributed equally to this work.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 586–591

ing 78.9(%) accuracy for English and 76.8(%) accuracy for Spanish. Srivastava (2022) participated in SemEval-2022 Task 5, specifically SubTask-A, which focused on identifying whether a meme contained misogyny. Using the ResNet-50nsfw model, they achieved a notable 7th-place ranking with an F1 score of 0.759. Rizzi et al. (2023) focused on misogyny detection in memes using unimodal and multimodal approaches, achieving an overall accuracy of 61.43(%). Their method incorporated a bias mitigation strategy based on Bayesian Optimization to improve model performance. Chinivar et al. (2024) tackled misogynistic meme detection using multimodal models on a benchmarked dataset. Their approach combined XLM-R for text and Swin for images, resulting in an F1 score of 0.7607. Singh et al. (2024) performed binary and multi-label classification of misogynistic memes. Their top-performing model, BiT (image) + MuRIL (text), for binary classification, achieved a high F1 score of 0.7319.The task described in Arango et al. (2022) involved identifying misogynous memes for the Multimedia Automatic Misogyny Identification (MAMI) task at SemEval-2022. Using a multimodal system based on the CLIP model, they achieved an F1 score of 71(%). Raha et al. (2022) addressed misogyny detection in memes as a binary classification task. Their best-performing models, VisualBERT and ViLBERT, attained an F1 score of 0.712. Using the MAMI task dataset, Ravagli and Vaiani (2022) worked on identifying misogynistic memes for SemEval-2022 Task 5 (MAMI). They combined Mask R-CNN for image processing and VisualBERT for multimodal processing. The VisualBERT (COCO) model achieved an F1 score of 0.670. Chen and Pan (2022) conducted hateful meme detection through text and image analysis. Their approach used OSCAR+RF, integrating the OSCAR Vision-Language Pre-Training Model with a Random Forest classifier, achieving an accuracy of 0.684.

3 Task and Dataset Description

In this shared task, the focus is on misogyny meme detection in Malayalam language. The task involves classifying memes as Misogynistic (labeled as 1) or Non-misogynistic (labeled as 0) in the Malayalam language. The dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024, 2025) provided by the organizers is sourced from various social media platforms.

Table1 depicts the statistical distribution of data. The training dataset contains a total of 640 samples, with 259 labeled as misogynistic and 381 as non-misogynistic. The validation dataset consists of 160 samples, including 63 misogynistic and 97 non-misogynistic samples. Out of the 200 samples in the test dataset, 78 are classified as misogynistic and 122 as non-misogynistic.

Classes	Train	Valid	Test	T_W
Mis	259	63	78	6398
NMiso	381	97	122	11004
Total	640	160	200	17402

Table 1: Dataset Statistics for Train, Validation, and Test Sets. (T_W denotes total words)

The dataset is provided in the form of an image with an associated transcription. We utilized image, text and multimodal (text + image) features to address this task. The implementation of our proposed approach has been made publicly available, and the source code can be accessed on GitHub[1].

4 Methodology

Before adopting a multimodal approach, we have focused on exploiting the visual aspects of memes by developing several CNN architectures and a Vision Transformer. For the textual aspects, we have implemented Text-CNN, Malayalam BERT, and IndicBERT. Finally, the visual and textual features are combined using fusion techniques to enhance the model's performance in detecting misogynistic content.Figure 1 depicts a schematic process in detecting fake news, illustrating each major phase.

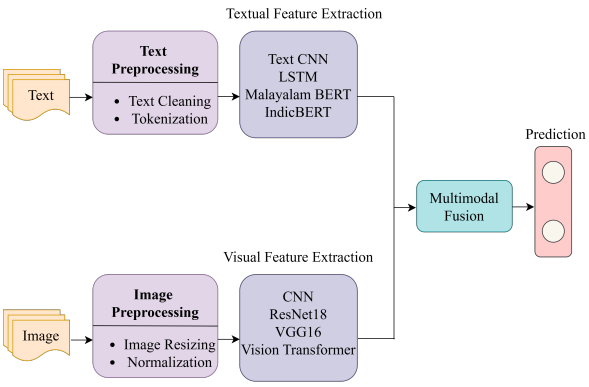


Figure 1: Schematic process of misogyny meme detection.

[1] https://github.com/DolaChakraborty12/Misogyny_Meme_Detection

4.1 Data Pre-processing

In the preprocessing step, unwanted symbols and punctuation were removed from the text. The text was then tokenized using a pre-trained BERT tokenizer, which converted the words into unique numerical representations. The sequences were padded to a fixed length of 128 tokens to ensure consistent input size across all samples. The images were resized to a fixed size of 224x224 pixels. In RGB format, the images were transformed to a size of (224x224x3). Each image was then normalized by scaling the pixel values to a range between 0 and 1. Additionally, image transformations were applied, including resizing, normalization, and conversion to tensors to prepare the data for input into the CNN model.

4.2 Visual Approach

For visual elements, several CNN-based architectures and transformer-based models were evaluated, such as Vision Transformer (ViT), VGG16, and a Convolutional Neural Network (CNN). The ViT was fine-tuned with a learning rate of 1e-5, batch size of 16, and 15 epochs, using AdamW optimizer and Binary Cross-Entropy loss. The VGG16 model was modified with custom layers and trained with a learning rate of 1e-5, batch size of 16, and 100 epochs, using categorical cross-entropy loss. Lastly, the CNN model, trained with a learning rate of 5e-5, batch size of 16, and 100 epochs, using categorical cross-entropy loss.

4.3 Textual Approach

In the textual approach, we used BiLSTM, TextCNN, LSTM + CNN, and Malayalam BERT.

- **BiLSTM**: Input text was tokenized using TensorFlow Keras (maximum vocabulary size: 10,000, input length: 100) and passed through an embedding layer (128-dimensional). It was then processed by a bidirectional LSTM layer with dropout rates of 0.8 and 0.5, followed by a dense layer with ReLU activation and L2 regularization (0.01). The model was trained using the Adam optimizer, binary cross-entropy loss, and balanced class weights, with a batch size of 16 for 10 epochs.

- **Malayalam BERT**: The model was fine-tuned using the Adam optimizer and binary cross-entropy loss with a batch size of 16 for 10 epochs, using a learning rate of 2e-5 and a weight decay of 0.01 to prevent overfitting.

- **Text-CNN**: Input text was first passed through an embedding layer (100-dimensional), followed by a convolutional layer with 128 filters and a kernel size of 5. The output was then processed by a max pooling layer, followed by a fully connected layer with 128 units and ReLU activation, and finally passed through a dropout layer (0.5) before the output layer. The model was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 32 for 100 epochs.

- **LSTM+CNN**: Input text was first processed by an embedding layer and then passed through a bidirectional LSTM layer with 128 units. The output was then fed into a convolutional layer, followed by a max pooling layer. Finally, the processed features were passed through fully connected layers before the output layer. The model was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 32 for 15 epochs.

4.4 Multimodal Approach

For visual feature extraction, multiple pretrained models, including Vision Transformer (ViT), CLIP, and ResNet-18 were utilized due to their strong performance in capturing diverse image features, ViT excels in global context understanding, CLIP aligns visual and textual features, and ResNet-18 excels in robust, hierarchical feature extraction. ViT processed the images by resizing them to 224x224 pixels and generating representations using a sequence-based transformer approach. CLIP and ResNet-18 were employed with batch sizes of 16 and learning rates set to 1e-5 and 2e-5 respectively, to extract additional visual and contextual features. The resulting image embeddings were passed through fully connected layers to reduce dimensionality and align with textual features for multimodal fusion.

For textual feature extraction, the system implemented advanced language models like Malayalam BERT and IndicBERT. These models are pretrained on large corpora of Dravidian languages, enabling them to capture language-specific syntactic and semantic features essential for accurately understanding the nuanced textual content in Malayalam memes. Malayalam BERT, fine-tuned for Malayalam text, was used to handle transliterated and native Malayalam text effectively. IndicBERT, being a multilingual model, was also implemented

for handling Malayalam effectively. Text inputs were tokenized, where sequences were padded or truncated to a fixed length, and embeddings were extracted from the model's output. The extracted embeddings were then passed through fully connected layers to transform them into a compact feature vector.

The outputs of both the visual and textual models were concatenated at a multimodal fusion layer. This integration of visual and textual features ensured that both modalities contribute effectively to the final prediction. A fully connected classification layer with a sigmoid activation was added after the fusion layer to produce the final binary prediction.

Training was conducted end-to-end with the binary cross-entropy loss function and the Adam optimizer, with a learning rate of 2e-5, a maximum sequence length of 128, and a batch size of 16. Table 2 demonstrates the hyperparameters of the best performed model(ResNet and Malayalam BERT).

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Max Length	128
Epochs	5

Table 2: Hyperparameter setup

5 Results and Analysis

Table 3 illustrates the performance of the various deep learning (DL) and transformer-based models employed on the test dataset across different approaches. The multimodal approach outperformed both visual and textual approaches. In the visual approach, the Vision Transformer (ViT) outperformed deep learning models, achieving an F1-score of 0.79. Malayalam BERT outperformed the other models in the textual method. Finally, in the multimodal approach, the fusion of ResNet (for visual features) and Malayalam BERT (for textual features) provided the best result, achieving a macro F1-score of 0.86.

6 Error Analysis

A detailed error analysis of the best-performed model is executed using quantitative and qualitative approaches.

Approach	Classifier	P	R	F1
Visual	ViT	0.98	0.56	0.79
	VGG16	0.77	0.67	0.68
	CNN	0.58	0.51	0.54
Textual	BiLSTM	0.73	0.72	0.72
	Malayalam BERT	0.41	0.80	0.54
	CNN	0.59	0.85	0.71
	LSTM + CNN	0.62	0.69	0.71
Multimodal	ResNet18 + Malayalam BERT	0.82	0.82	0.82
	ResNet18 + IndicBERT	**0.87**	**0.88**	**0.87**
	ViT + Malayalam BERT	0.88	0.79	0.81
	ViT + IndicBERT	0.86	0.86	0.86
	CLIP + Malayalam BERT	0.88	0.83	0.85

Table 3: Performance of various DL and Transformer-based models on the test set. P (Precision), R (Recall), F1 (macro F1-score).

Quantitative Analysis: Figure 2 presents the confusion matrix of the best-performing multimodal model, ResNet18 + IndicBERT. A detailed error analysis of the fine-tuned multimodal model is performed based on the confusion matrix. It is evident from the confusion matrix that, out of 200 samples, 176 are correctly predicted. The model misclassifies 11 misogynistic samples as non-misogynistic and 13 non-misogynistic samples as misogynistic.

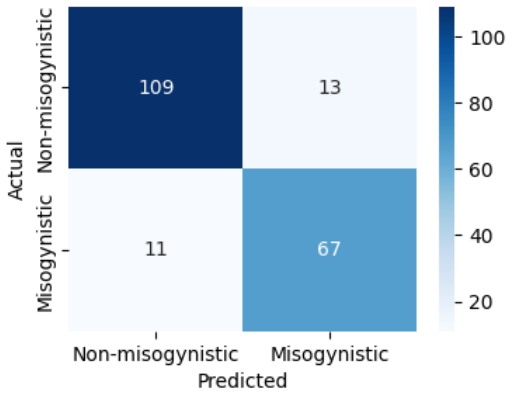


Figure 2: Confusion matrix of the best-performed model (ResNet18 + IndicBERT).

Qualitative Analysis: A comparison of actual labels and predicted labels for a particular transcription is illustrated in Figure 3. The first two samples are incorrectly predicted as misogynistic, even though they are non-misogynistic. However, the next two samples are predicted correctly as their actual labels.

The misclassifications observed in the results can be attributed to the challenges inherent in multimodal fusion, where both image and text features are integrated. While the fusion of deep learning-based image features (extracted via ResNet18) and

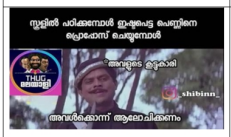
Images ID	Transcriptions	Predicted Labels	Actual Labels
	സ്കൂളിൽ പഠിക്കുമ്പോൾ ഇഷ്ടപെട്ട പെണ്ണിനെ പ്രപ്പോസ് ചെയ്യുമ്പോൾ "അവളുടെ കൂട്ടുകാരി അവൾക്കൊന്ന് ആലോചിക്കണം	1	0
	അറിയാൻ പാടില്ലാത്ത ഒരാളെകുറിച്ച് നമ്മൾ ഒരിക്കലും കുറ്റം പറയരുത് \nകാരണം അടുത്തറിയുമ്പോഴായിരിക്കും അവനൊരു പാവമാണെന്ന് പലർക്കും മനസ്സിലാവുക	1	0
	എന്തൊക്കെ ആകിയിട്ടും ഒരു മെന വരുന്നില്ലലോ നിഖിയേ..	1	1
	ഇത് ഞാൻ ചെറുതായിരുന്നപ്പോൾ ഇത് 5ആം ക്ലാസ്സ് വരെ കണ്ടു പിന്നെ ഇത്. ഇത്... ഇന്ന് രാവിലെയും കൂടെ കണ്ടു	0	0

Figure 3: Some predicted outputs by (ResNet18 + IndicBERT).

transformer-based text features (from IndicBERT) enables the model to leverage complementary information, it may also introduce certain ambiguities. The concatenation of these two distinct feature types can sometimes lead to confusion in classification, as the model must balance the influence of both modalities. The model often struggles with sarcastic statements where the textual content appears non-misogynistic but carries misogynistic undertones when paired with the image. The model fails to capture such implicit misogyny, leading to misclassification. Some memes contain misleading or ambiguous text, cultural references, slang, and region-specific humor, particularly in Malayalam, where proverbs or idiomatic expressions may have context-dependent misogynistic intent, challenging the model's classification.

7 Conclusion

This work explored the effectiveness of various DL and transformer-based models for misogyny meme detection in Malayalam. Different modalities, including textual, visual, and multimodal approaches, are systematically evaluated. ViT achieved the highest F1 score among the visual models, demonstrating its ability to capture complex visual patterns. BiLSTM outperformed other models for the textual modality, showcasing its strength in handling sequential data. However, the best overall performance is achieved through a multimodal approach that combined ResNet18 and IndicBERT, resulting in the highest F1 score of 0.87. This result highlights the significance of integrating complementary features from textual and visual modalities for addressing challenging tasks like misogyny meme detection. Future work can enhance this task by incorporating larger datasets to improve model robustness, reduce bias, and enhance generalization by exposing the model to a diverse range of misogynistic and non-misogynistic memes. Additionally, exploring single transformer-based approaches for multimodal learning and investigating large language models can improve performance. While this study focuses on Malayalam memes, our results suggest that IndicBERT outperforms language-specific models like MalayalamBERT, highlighting its potential for cross-lingual effectiveness in other Dravidian languages.

Limitations

The current model posses several weaknesses. A few of them are illustrating in the following:

- Combining ResNet18 and IndicBERT features can introduce confusion, leading to misclassifications.
- The small dataset may hinder the model's ability to capture nuanced patterns and generalize effectively.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 581–584, Seattle, United States. Association for Computational Linguistics.

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.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1864–1871.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9):1–12.

Sneha Chinivar, M. S. Roopa, J. S. Arunalatha, and K. R. Venugopal. 2024. Identification of misogynistic memes using transformer models. In *Proceedings of International Conference on Advanced Communications and Machine Intelligence*, pages 107–116, Singapore. Springer Nature Singapore.

Maibam Debina Devi and Navanath Saharia. 2021. Misogynous text classification using svm and lstm. In *Advanced Computing*, pages 336–348, Singapore. Springer Singapore.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Díaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Pérez, and Olatz Perez-de Viñaspre. 2018. Automatic misogyny identification using neural networks. In *IberEval@ SEPLN*, pages 249–254.

Marinella Paciello, Francesca D'Errico, Giorgia Saleri, and Ernestina Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in human behavior*, 116:106655.

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.

Tathagata Raha, Sagar Joshi, and Vasudeva Varma. 2022. IIITH at SemEval-2022 task 5: A comparative study of deep learning models for identifying misogynous memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 673–678, Seattle, United States. Association for Computational Linguistics.

Jason Ravagli and Lorenzo Vaiani. 2022. JRLV at SemEval-2022 task 5: The importance of visual elements for misogyny identification in memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 610–617, Seattle, United States. Association for Computational Linguistics.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Elena Shushkevich and John Cardiff. 2018. Misogyny detection and classification in english tweets: The experience of the itt team. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:182.

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. *Preprint*, arXiv:2203.16781.