CUET_Novice@DravidianLangTech 2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Malayalam Language

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama and **Ashim Dey**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904013, u1904008, u1804039}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Memes, combining images and text, are a popular social media medium that can spread humor or harmful content, including misogyny—hatred or discrimination against women. Detecting misogynistic memes in Malayalam is challenging due to their multimodal nature. A Shared Task on Misogyny Meme Detection, organized as part of DravidianLangTech@NAACL 2025, aimed to address this issue by promoting the advancement of multimodal machine learning models for classifying Malayalam memes as misogynistic or non-misogynistic. In this work, we explored visual, textual, and multimodal approaches for meme classification. CNN, ResNet50, Vision Transformer (ViT), and Swin Transformer were used for visual feature extraction, while mBERT, IndicBERT, and MalayalamBERT were employed for textual analysis. Additionally, we experimented with multimodal fusion models, including IndicBERT+ViT, MalayalamBERT+ViT, and MalayalamBERT+Swin. Among these, our MalayalamBERT+Swin Transformer model performed best, achieving the highest weighted F1-score of 0.87631, securing **1st** place in the competition. Our results highlight the effectiveness of multimodal learning in detecting misogynistic Malayalam memes and the need for robust AI models in low-resource languages.

1 Introduction

Misogynistic content fosters hostility and discrimination, particularly targeting specific genders, and poses a significant challenge to creating safe and inclusive online environments. The rise of social media has accelerated the spread of such content, often as text-visual memes. Detecting misogyny in multimodal formats is challenging, as intent depends on text-image interplay. In Malayalam, linguistic complexity adds to the difficulty, requiring advanced tokenization and semantic analysis for effective detection. Addressing these linguistic intricacies is crucial for building robust misogyny detection models. The Misogynistic Meme Detection Shared Task, conducted as part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025), aimed to tackle these challenges by identifying misogynistic content in Tamil and Malayalam memes.

Our participation focused specifically on Malayalam memes. Through this work, we aimed to address the unique challenges posed by the multimodal nature of memes and the intricacies of Malayalam text. Our key contributions include:

- Utilized the Swin Transformer for visual feature extraction, leveraging its advanced capabilities for image representation, and Malayalam-BERT for extracting textual features, given its effectiveness in capturing the nuances of the Malayalam language.
- Additionally, we evaluated the performance of models that were trained exclusively on textual or visual data, which helped us to understand the relative contributions of each modality.

This work not only contributes to the broader goal of misogyny detection in underrepresented languages but also emphasizes the importance of multimodal approaches in tackling the nuanced challenges of meme classification. The code is available at https://github.com/Khadiza13/Misogyny-NAACL2025.

2 Related Work

In recent years, there has been a growing focus among NLP researchers on identifying trolling, hostility, offensive language, and abusive content on social media platforms. While early studies primarily focused on analyzing textual information ((Anzovino et al., 2018) (Sadiq et al., 2021)

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 550–555

(Ishmam and Sharmin, 2019)), recent works have explored multimodal approaches that consider both textual and visual features embedded in memes. (Jha et al., 2024) introduced MultiBully-Ex, a dataset for multimodal explanations in code-mixed cyberbullying memes, combining visual and textual data. Similarly, (Hasan et al., 2022) demonstrated that the CNN-Text+VGG16 combination outperformed other multimodal models with an F1-score of 0.49 for meme detection. (Barman and Das, 2023) utilized mBERT for textual features, ViT for visual features, and MFCC for audio features to tackle abusive language detection. (Rahman et al., 2024) introduced a hybrid ConvLSTM+BiLSTM+MNB model, which obtained the highest macro F1-score of 71.43%. (Ahsan et al., 2024) presented MIMOSA, a new multimodal dataset for detecting Bengali aggression, containing 4,848 annotated memes classified into five aggression categories. They proposed the Multimodal Attentive Fusion (MAF) method. Similarly, (Mahesh et al., 2024) focused on identifying misogynistic memes in Tamil and Malayalam. Their models, including mBERT+ResNet-50 and MuRIL+ResNet-50, obtained macro F1-scores of 0.73 and 0.87, respectively. (Osama et al., 2024) highlighted mBERT's strong performance in misinformation detection for low-resource languages such as Malayalam. Additionally, (Rehman et al., 2025) presented a multimodal framework for detecting misogynistic content using attention, graph-based refinement, and lexicon-based features, achieving notable improvements on benchmark datasets.(Gu et al., 2022) explored ensemble models for misogyny classification, like Naive Bayes and gradient boosting.

3 Task and Dataset Description

A misogynistic meme combines visual content and text to demean, stereotype, or offend women, often spreading harmful ideologies on social media (Ponnusamy et al., 2024). The goal of this task is to classify misogynistic memes by leveraging both visual and textual information (**?**). The organizers provided a dataset containing two types of memes (Misogynistic and Non-misogynistic) in the Tamil and Malayalam languages (Chakravarthi et al., 2024). Here, Table 1 provides the distribution of samples across training, development, and test sets. The dataset is presented as an image accompanied by a corresponding caption.

Class	Train	Dev	Test	Total	Words
Misogynistic	259	63	78	400	3735
Non-misogynistic	381	97	122	600	6560
Total	640	160	200	1000	10295

Table 1: Statistical distribution of our dataset.

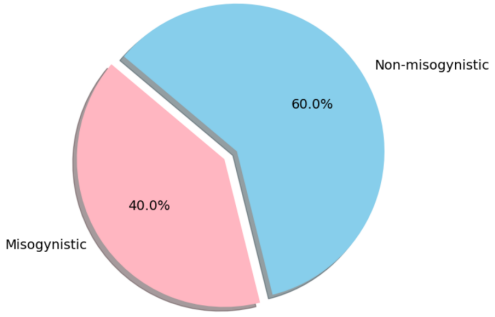


Figure 1: Percentage distribution of two different classes.

Participants can use either the image, the caption, or both to complete the classification task. We employed image, text, and multimodal (image + text) features to tackle the given task.

4 Methodology

The aim of this study is to detect misogynistic content in multimodal Malayalam memes. Initially, we exploit the visual aspects of the memes. Subsequently, the textual information is processed using Malayalam-specific language models, and finally, both modalities are combined through a fusion mechanism to make robust classification decisions. Figure 2 offers a clear visualization of our methodology, highlighting the essential steps in our approach.

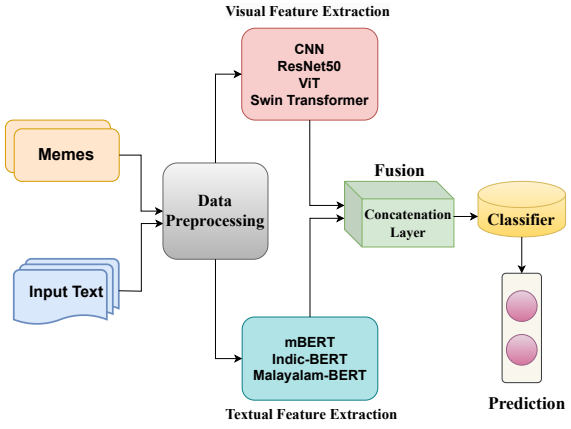


Figure 2: Abstract view of our methodology.

4.1 Data Preprocessing

In this step, the Malayalam text undergoes tokenization using specialized tokenizers from the Malayalam-BERT model. The text is transformed into numerical representations with a maximum sequence length of 128 tokens. Special tokens ([CLS], [SEP]) are added as required by the transformer architecture. For image preprocessing, all memes are transformed into a size of 224×224 pixels, normalized using standard ImageNet statistics (Deng et al., 2009), and converted to RGB format to maintain consistency across the dataset.

4.2 Visual Approach

For visual feature extraction, we first use a CNN with 8 layers, followed by ResNet-50 (He et al., 2016), which is pre-trained on ImageNet. After that, we employ the Vision Transformer (Dosovitskiy et al., 2020) and finally employ the Swin Transformer model (Liu et al., 2021), which utilizes a hierarchical structure with shifted windows for efficient processing of visual information. The model, pre-trained on ImageNet, processes the meme images to generate 1024-dimensional feature vectors. This architecture was chosen for its proven effectiveness in capturing both local and global visual features.

4.3 Textual Approach

The textual component of memes is first processed using BERT-Base Multilingual Cased (Devlin et al., 2018) leveraging pre-trained weights, token resizing, dropout, and a classification layer for meaningful representations. After this, we employ Indic-BERT (Kakwani et al., 2020) and then Malayalam-BERT (Tabassum et al., 2024), a transformer-based model trained for Malayalam, generates 768-dimensional feature vectors, excelling in linguistic nuance and contextual understanding.

4.4 Multimodal Approach

Our multimodal approach combines the visual and textual features through fusion strategy. The visual features from Swin Transformer (1024-dimensional) and textual features from Malayalam-BERT (768-dimensional) are concatenated to form a 1792-dimensional vector. This combined representation is then processed through a two-layer neural network classifier. The first layer reduces the dimensionality to 512, followed by ReLU activation and dropout (0.1) for regularization. The final layer produces binary classification outputs for misogyny detection. The training protocol uses AdamW (learning rate: 5e-5, batch size: 16) for 5 epochs, using binary cross-entropy loss, a learning rate scheduler with warmup steps, and gradient clipping for stability. Table 2 shows the list of tuned hyperparameters used in the experiment.

Hyperparameter	Value
Batch Size	16
Learning Rate	5e-5
Optimizer	AdamW
Epochs	5
Dropout Rate	0.1
Weight Decay	0.01

Table 2: Overview of optimized hyper-parameters.

5 Results & Discussion

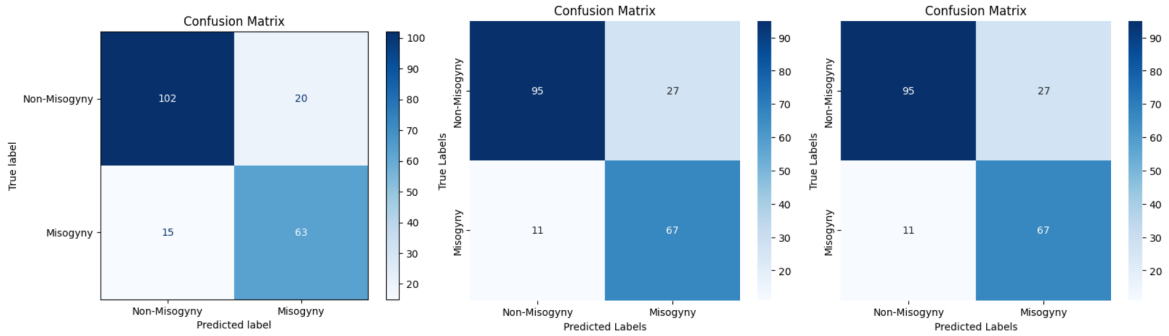
This section presents a comparative performance analysis of various experimental approaches for classifying memes. The effectiveness is primarily assessed based on the weighted f1-score, while precision and recall are also considered in some cases. Table 3 presents a summary of the precision (P), recall (R), and F1 (f1) scores for each model on the test set. The results show that, Swin Transformer and Malayalam-BERT performed best in visual and textual models, respectively, with an F1-score of 0.78. However, the top classification performance was seen in the multimodal models, where combining Malayalam-BERT and Swin Transformer resulted in the highest F1-score of 0.88. These findings highlight the superiority of multimodal models in meme classification by combining text and visual features.

Approach	Classifier	P	R	f1
Visual	CNN	0.62	0.51	0.56
	ResNet50	0.91	0.13	0.22
	ViT	0.87	0.68	0.76
	Swin Transformer	0.76	0.81	0.78
Textual	mBERT	0.69	0.58	0.63
	Indic-BERT	0.62	0.82	0.71
	Malayalam-BERT	0.71	0.86	0.78
Multimodal	Indic-BERT+ViT	0.73	0.83	0.78
	Malayalam-BERT+ViT	0.78	0.81	0.80
	Malayalam-BERT+Swin	**0.87**	0.78	**0.88**

Table 3: Performance of different models on test set.

5.1 Quantitative Discussion

The results underscore the effectiveness of transformer-based architectures in identifying misogynistic content. The confusion matrix in



(a) Visual (b) Texual (c) Multimodal

Figure 3: Confusion matrix of three different approaches.

Figure 3 provides a detailed breakdown of our model's performance. Here, the visual model correctly classifies 102 Non-Misogyny and 63 Misogyny samples but misclassifies 20 instances of Non-Misogyny as Misogyny and 15 Misogyny instances as Non-Misogyny. The textual model improves classification, correctly predicting 95 Non-Misogyny and 67 Misogyny samples, though it misclassifies 27 Non-Misogyny samples. The multimodal model (Malayalam-BERT + Swin Transformer) outperforms both unimodal models, correctly classifying 113 Non-Misogyny and 61 Misogyny instances, with fewer misclassifications (9 false positives and 17 false negatives). These results affirm that multimodal models improve precision and recall in detecting misogynistic memes.

5.2 Qualitative Discussion

Figure 4 presents sample predictions from our best-performing Malayalam-BERT+Swin Transformer model. In the first instance, the model incorrectly classified the sample as non-misogynistic (label 0). This misclassification might have occurred because the text, although seemingly neutral, could have contained subtle contextual cues or sarcasm that the model failed to pick up on. In contrast, the second sample, which was genuinely non-misogynistic (label 0), was misclassified as misogynistic (label 1). This could be attributed to the image associated with the text, which may have included visual elements such as expressions, body language, or symbols that the model interpreted as indicative of misogyny. Furthermore, cultural norms and societal stereotypes in the visual context may have influenced inaccurate predictions. While the model captures linguistic and contextual cues well, it struggles with nuanced cases involving sarcasm, cultural context, or visual ambiguity.

English Translation: To the wife who wakes up in middle of the night putting on makeup Husband: Why are you Wearing makeup at this time??Wife: Thats my phone face lock dude... Now it won't turn on without makeup on

Actual: 1
Predicted: 0

English Translation: WhatsApp status of most people who have seen it for a few days

Actual: 0
Predicted: 1

Figure 4: Examples of some wrongly classified sample of the best model.

6 Conclusion

This work presented the details of the methods and performance analysis of the models for detecting misogynistic memes in Malayalam, exploring visual, textual, and multimodal fusion techniques. The results revealed that the Malayalam BERT+Swin Transformer model got the highest weighted F1-score of 0.88, demonstrating that multimodal fusion significantly enhances model performance. In the future, we plan to explore audio and video modalities, advanced fusion strategies, extend the dataset and ensemble models for better robustness, especially in low-resource languages. Transfer learning, domain-specific knowledge, and addressing social and cultural biases will also enhance the model's adaptability, fairness, and generalization.

Limitations

A primary limitation of this study lies in the reliance on pre-trained models for both visual and textual features, which may not fully capture the nuances of Malayalam-specific cultural context or meme content. While our multimodal approach performs well, the models used are limited by their generalization capabilities when handling domain-specific or low-resource language memes. Additionally, the dataset used for training may not be comprehensive enough to account for all variations in meme content, which could impact the robustness of the model. Furthermore, the impact of cultural norms, humor, and sarcasm—which are often deeply embedded in Malayalam social discourse—has not been explicitly analyzed. Misogynistic content can sometimes be expressed subtly through irony, satire, or culturally specific references, making it difficult for AI models to detect intent accurately. Future work with a larger, more diverse dataset, the exploration of specialized Malayalam language models, and a deeper analysis of sarcasm and cultural factors in error cases could enhance model accuracy and generalizability. While the current dataset was balanced and did not require data augmentation, it would be crucial to incorporate data augmentation techniques when dealing with larger and imbalanced datasets. By generating synthetic examples through text or image transformations, data augmentation could help address class imbalance and improve the model's ability to generalize across different classes. This would ensure better performance, especially in situations where certain classes are underrepresented, ultimately leading to a more robust and reliable model for real-world applications.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Shubhankar Barman and Mithun Das. 2023. hatealert@DravidianLangTech: Multimodal abusive language detection and sentiment analysis in Dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Qin Gu, Nino Meisinger, and Anna-Katharina Dick. 2022. Qinian at semeval-2022 task 5: Multi-modal misogyny detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 736–741.

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.

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.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_goodfellows@ dravidianlangtech eacl2024: A transformer-based approach for detecting fake news in dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

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.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu Sang Choi, and Byung-Won On. 2021. Aggression detection through deep neural model on twitter. *Future Generation Computer Systems*, 114:120–129.

Nafisa Tabassum, Sumaiya Aodhora, Rowshon Akter, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Punny_punctuators@ dravidianlangtech-eacl2024: Transformer-based approach for detection and classification of fake news in malayalam social media text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 180–186.