CUET-823@DravidianLangTech 2025: Shared Task on Multimodal Misogyny Meme Detection in Tamil Language

Arpita Mallik, Ratnajit Dhar, Udoy Das[†], Momtazul Arefin Labib, Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004023, u2004008}@student.cuet.ac.bd, udoy.d@eastdelta.edu.bd,
{u1904111, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Misogynous content on social media, especially in memes, present challenges due to the complex reciprocation of text and images that carry offensive messages. This difficulty mostly arises from the lack of direct alignment between modalities and biases in large-scale visio-linguistic models. In this paper, we present our system for the Shared Task on Misogyny Meme Detection - DravidianLangTech@NAACL 2025. We have implemented various unimodal models, such as mBERT and IndicBERT for text data, and ViT, ResNet, and EfficientNet for image data. Moreover, we have tried combining these models and finally adopted a multimodal approach that combined mBERT for text and EfficientNet for image features, both fine-tuned to better interpret subtle language and detailed visuals. The fused features are processed through a dense neural network for classification. Our approach achieved an F1 score of 0.78120, securing 4th place and demonstrating the potential of transformer-based architectures and state-of-the-art CNNs for this task.

1 Introduction

The concept of memes have evolved into a powerful form of cultural transmission on the internet. Despite being widely debated in academic circles, memes have been adopted extensively, as evidenced by a surge of interest since 2011 and over 1.9 million Google search results for 'Internet meme,' which has highlighted their great cultural significance (Shifman, 2013).

Although memes have often been viewed as harmless entertainment, the internet has become a space where the harassment of women and marginalized groups has been documented widely in both academic and popular press. Feminist research has revealed that online sexism and harassment have frequently been reframed as "acceptable" form of humor (Drakett et al., 2018). The need to address multimodal issues has been highlighted by the fact that hateful content targeting women is not only found in text but also in visual, audio, or combined forms (Singhal et al., 2022).

The purposive focus of this paper has been to detect misogyny in Tamil-English code-mixed memes on social media platforms. The DravidianLangTech@NAACL 2025 conference (Chakravarthi et al., 2025) has introduced a dataset under the Shared Task on Misogyny Meme Detection (Ponnusamy et al., 2024), which has consisted of multimodal content with textual and visual elements labeled as misogynistic or non-misogynistic, enabling the identification of misogyny in a multimodal context.

To achieve our goal, we have applied various image and text augmentation techniques, such as brightness adjustment, grayscale transformations, and back-translation, and have evaluated various models, including transformer-based models (mBERT, IndicBERT), image models (ResNet, ViT, and EfficientNet), and multimodal models combining mBERT with EfficientNet, mBERT with ViT, and IndicBERT with ResNet. The results have shown that the multimodal models have outperformed the unimodal ones. With an F1 score of 0.78120, our approach has ranked 4th in the competition. The main contributions of this work have been:

- We have made comparison of various multimodal and unimodal models, to find the best suitable one for the given dataset.
- We have applied different types of data augmentation techniques, including back-translation, to address class imbalance.
- We have developed a preprocessing pipeline involving text transliteration and image enhancements to improve data quality.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 403–407

Further implementation details can be accessed via the GitHub repository: [1].

2 Related Work

In the existing research on the topic of misogyny meme detection, only a few approaches have specifically addressed the issue, which is a major concern on social media platforms all over the world. Prior studies in this domain can be categorized based on their approach—text-based, image-based, and multimodal—as well as their use of machine learning, deep learning, and pre-trained models. Additionally, research efforts have focused on binary vs. multi-label classification and bias mitigation.

Nozza et al. (2021) has shown that hate speech detection models are not effective across different types of hate speech targets, highlighting the need for specialized approaches and datasets for detecting misogyny.

Recent studies have explored multimodal techniques to improve classification accuracy. For instance, Shaun et al. (2024) proposed a multimodal method for classifying Tamil and Malayalam memes as "Misogynistic" or "Non-Misogynistic", using Multinomial Naive Bayes, with outputs combined using weighted probabilities. This proved the effectiveness of combining modalities for low-resource languages.

Building on the success of prompt learning in NLP, some researchers have investigated prompt-based approaches for identifying harmful memes (Jindal et al., 2024). These methods involved converting images into textual representations to reduce the semantic gap between the text and images in memes.

Other studies have used advanced fusion techniques. Attanasio et al. (2022) presented a system using Perceiver IO for late fusion in misogynous meme detection. It combines ViT for images and RoBERTa for text, handling both binary and multi-label classification. The approach outperformed baseline models and showed the effectiveness of Perceiver IO for multimodal fusion. Hakimov et al. (2022) proposed a pre-trained CLIP model to extract text and image features, which are then combined with an LSTM layer. Pramanick et al., 2021 also used CLIP embeddings, along with intra-modality attention and cross-modality fusion in their proposed model MOMENTA, which is a multimodal deep learning model for detecting harmful memes along with their targets.

3 Data

We have utilized the dataset provided under the Shared Task on Misogyny Meme Detection - DravidianLangTech@NAACL 2025 (Ponnusamy et al., 2024). The dataset has been segmented into training, development, and test sets containing 1136, 284, and 356 samples, respectively. It primarily consists of code-mixed Tamil-English memes, the type of language commonly observed in online communication. The dataset consists of significantly lower number of misogynistic memes compared to non-misogynistic ones, as shown in Table 1.

Sets	Misogyny	Not-misogyny	Total
Train	285	851	1,136
Development	74	210	284
Test	89	267	356
Total	448	1,328	1,776

Table 1: Label Distribution for Misogyny Meme Detection Dataset.

4 Methodology

4.1 Data Preprocessing

In terms of image preprocessing, all images have been resized to a consistent dimension of $224 \times 224 \times 3$ pixels. Contrast and brightness have also been adjusted with specific control parameters to enhance image quality.

For text preprocessing, unwanted symbols, punctuation, numbers, URLs, and emojis have been removed. Tamil stopwords have been filtered out to preserve meaningful content. Since the dataset has comprised text in Tamil, English, and Tamil written in English script, the preprocessed text has been transliterated into Tamil using the Indic Transliteration library[2].

4.2 Data Augmentation

Targeted data augmentation techniques have been applied to tackle the class imbalance in our dataset. Since the misogynistic memes have been significantly fewer than the non-misogynistic ones, as

[1] https://github.com/ratnajit-dhar/CUET-823-Multimodal-Misogyny-Meme-Detection

[2] https://pypi.org/project/indic-transliteration/

reflected in Table 1, focus has been placed solely on augmenting the misogynistic memes to balance the dataset.

For image data, the torchvision library[3] has been used to apply brightness adjustment, grayscale transformation, and posterization. For text, back-translation has been employed using the deep-translator library[4] with two pipelines: Tamil → English → Tamil and Tamil → Malayalam → Tamil. Additionally, synonym replacement and paraphrasing augmentation techniques have been experimented with, but these methods have not yielded satisfactory results. Table 2 shows the class distribution after data augmentation technique has been applied.

Class	Before	After
Misogynistic	285	855
Non-Misogynistic	851	851

Table 2: Dataset distribution before and after augmentation.

4.3 Overview of Experimented Models

4.3.1 Unimodal model

We have fine-tuned mBERT and IndicBERT for textual features, tokenizing with a max length of 512 tokens. The models have been trained on 20 topics with a batch size of 8, using the AdamW optimizer and a learning rate of 2e-5.

For image data, we have experimented with Vision Transformers (ViT), ResNet, and EfficientNet. The images were converted to PIL format, resized to 224×224, converted to tensors, and normalized using standard ImageNet values. The models have been trained for 20 epochs with a batch size of 16, using the Adam optimizer with a learning rate of 1e-4. The final fully connected layer has been replaced to match the number of unique classes. The architecture of the unimodal text model is shown in Figure 1, while the unimodal image model is illustrated in Figure 2

4.3.2 Multimodal model

Building on these unimodal baselines, we have explored multimodal models, such as mBERT with ResNet, mBERT with ViT, and IndicBERT with EfficientNet. These multimodal architectures allow the model to learn complex interactions between

[3] https://pytorch.org/vision/stable/index.html
[4] https://pypi.org/project/deep-translator/

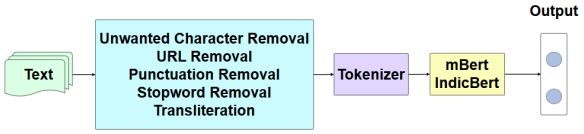

Figure 1: Unimodal Architecture for Text Data Processing and Classification.

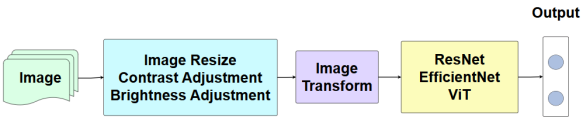

Figure 2: Unimodal Architecture for Image Data Processing and Classification.

textual and visual inputs, as shown in Figure 3. Our best-performing multimodal model, mBERT + EfficientNet, was fine-tuned using a 512-sequence-length mBERT tokenizer and images resized to 224×224 pixels. The 768-dimensional text embeddings and 1280-dimensional image features have been combined via a fully connected layer. The model was trained for 20 epochs with a batch size of 16, and the learning rate was set to 1e-4. For the mBERT + ViT model, text and image features were concatenated using 768 dimensions from both modalities. In the IndicBERT + ResNet model, the concatenation involved 768-dimensional text embeddings and 2048-dimensional image features.

5 Results and Analysis

This section presents the outcomes of our misogyny meme classification task, comparing unimodal and multimodal approaches to highlight their effectiveness in addressing classification challenges.

We have evaluated the performance of our models using weighted precision, recall, F1 score, and macro-averaged F1 score (Macro-F1). The Macro-F1 score is considered as the primary metric for assessing the final performance of the systems.

5.1 Comparative Analysis

We have assessed the performance of various models and found that among the unimodal text classifiers, mBERT performed better than IndicBERT with a higher F1 score of 0.69 compared to 0.62. For unimodal image classifiers, ResNet achieved a better score than EfficientNet and ViT.

When we combined mBERT with EfficientNet in a multimodal setup, the model achieved the highest F1 score of 0.78 with a precision of 0.80 and a recall of 0.77. This shows the strength of com-

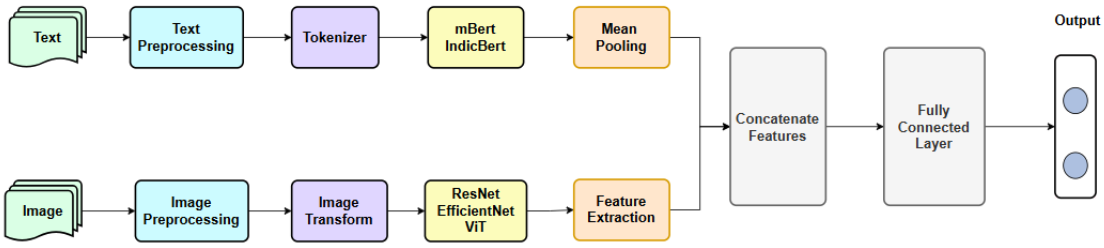

Figure 3: Multimodal Architecture for Text-Image Classification.

bining textual and visual modalities for improved performance.

We have also evaluated the performances of mBERT + ResNet, IndicBERT + ViT, and various other combinations of multimodal models. However, Table 3 shows only the three best performing approaches.

5.2 Error Analysis

To further analyze model performance, we present the confusion matrix in Figure 4. It shows that our best multimodal model (mBERT + EfficientNet) correctly classified 247 non-misogynistic and 54 misogynistic memes. However, it misclassified 20 non-misogynistic memes as misogynistic (false positives) and 35 misogynistic memes as non-misogynistic (false negatives). This suggests that the model may sometimes rely on certain words or patterns rather than the overall meaning, leading to incorrect predictions.

	Classifier	Macro Average		
		P	R	F1
Unimodal (Text)	mBert	0.68	0.71	0.69
	IndicBert	0.64	0.61	0.62
Unimodal (Image)	ResNet	0.77	0.74	0.74
	EfficientNet	0.76	0.69	0.72
	ViT	0.71	0.63	0.65
Multi-modal	mBert+ EfficientNet	**0.80**	**0.77**	**0.78**
	mBert+ViT	0.79	0.72	0.75
	IndicBert+ ResNet	0.79	0.73	0.75

Table 3: Performance of different systems on the test dataset.

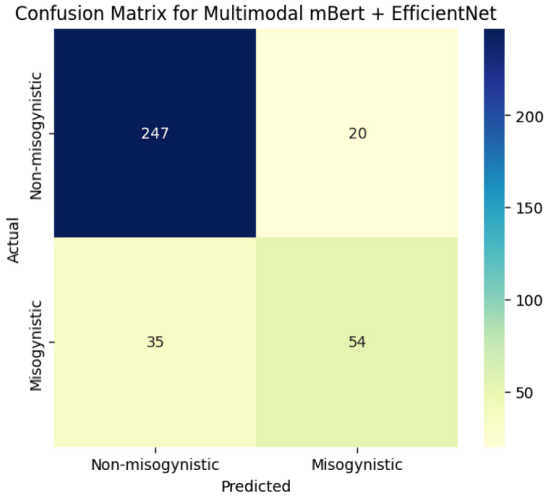

Figure 4: Confusion Matrix of the Multimodal mBert-EfficientNet Model.

6 Conclusion

In this research, we have evaluated various unimodal models and then compared various combinations of multimodal models for detecting misogynistic content in memes. Regardless of the challenges from the limited dataset, multimodal models have surpassed unimodal approaches, highlighting the importance of incorporating textual and visual information. We have discovered that the multimodal mBert + EfficientNet has performed the best among the other multimodal approaches, with an F1 score of 0.78. Future work will focus on expanding the dataset, improving data augmentation techniques, and better utilizing both text and images to detect subtle misogynistic content more effectively.

Limitations

The dataset provided for our task has been relatively small and imbalanced, with an inadequate

number of misogynistic memes. Although data augmentation techniques have been applied, the dataset size has still impacted performance, especially for the minority class. Additionally, although contrast and brightness adjustments were applied to improve image quality, but they could not fully eliminate noise and inconsistencies, leading to some misclassifications. Lastly, our model's performance could be further improved by training on more nuanced and subtle examples, which are currently underrepresented in the training data.

Ethics Statement

In this study, we have developed our methodology following the highest ethical practices. By contributing to the identification of misogynistic content in Tamil-English code-mixed memes, we hope to make the internet a safer and more inclusive place. We are committed to sharing our findings to prevent online misogyny while respecting linguistic and cultural diversity.

References

Giuseppe Attanasio, Debora Nozza, Federico Bianchi, et al. 2022. Milanlp at semeval-2022 task 5: Using perceiver io for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 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.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media–online sexism and constructions of gender in internet memes. *Feminism & psychology*, 28(1):109–127.

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*.

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.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for 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. *arXiv preprint arXiv:2109.05184*.

H Shaun, Samyuktaa Sivakumar, R Rohan, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@ lt-edi 2024: A svm-resnet50 approach for multitask meme classification-unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.