InnovationEngineers@DravidianLangTech 2025: Enhanced CNN Models for Detecting Misogyny in Tamil Memes Using Image and Text Classification

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Pooja Sree M**[1],
Palanimurugan V[1],**Roshini Priya K** [1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{poojasreem,palanimuruganv,roshinipriyak}.22aid@kongu.edu

Abstract

The rise of misogynistic memes on social media posed challenges to civil discourse. This paper aimed to detect misogyny in Dravidian language memes using a multimodal deep learning approach. We integrated Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM), EfficientNet, and a Vision Language Model (VLM) to analyze textual and visual information. EfficientNet extracted image features, LSTM captured sequential text patterns, and BERT learned language-specific embeddings. Among these, VLM achieved the highest accuracy of 85.0% and an F1-score of 70.8, effectively capturing visual-textual relationships. Validated on a curated dataset, our method outperformed baselines in precision, recall, and F1-score. Our approach ranked 12th out of 118 participants for the Tamil language, highlighting its competitive performance. This research emphasizes the importance of multimodal models in detecting harmful content. Future work can explore improved feature fusion techniques to enhance classification accuracy.

Keywords: LSTM, BERT, EfficientNet, Vision Language Model, Meme Classification

1 Introduction

Social media platforms have developed into places for entertainment and idea sharing, but they have also made it possible for harmful information, such as misogynistic memes, to proliferate. These text-and-image memes frequently spread harmful viewpoints and disrupt online conversation. In Dravidian languages, which are spoken in southern India and nearby areas and are frequently underrepresented in language processing studies, it is particularly difficult to detect such material. By creating a binary classification model that classifies information as either abusive or non abusive, this work aims to detect misogyny in memes across Dravidian languages. Using a multimodal deep learning architecture, our method analyzes both the text and images in memes by combining multiple powerful models: Vision-Language Models (VLM) to comprehend the relationship between images and text, EfficientNet to extract image features efficiently, BERT to process language-specific text embeddings, and LSTM to capture the flow of text sequences. The VLM achieves the highest accuracy among these models, highlighting the significance of analyzing both visual and textual information simultaneously. Our model performs well on a curated dataset, surpassing previous methods and helping to combat abusive content in low-resource languages.

2 Literature Survey

The growing prevalence of online hate speech, particularly in the form of misogyny and trolling, has led to a surge in research focusing on the detection of such harmful content. The detection of misogyny in online content, particularly through memes, has been extensively studied in recent years.

Ponnusamy et al. (2024) introduced an annotated dataset for misogyny detection in Tamil and Malayalam memes. The study highlights challenges in identifying offensive content in low-resource languages due to cultural and contextual nuances. The dataset, sourced from social media, is manually labeled and evaluated using NLP techniques. Their work emphasizes the importance of multimodal approaches, considering both textual and visual elements in memes. This research aids in developing AI systems to detect online misogyny and improve fairness in content moderation.

Jindal et al. (2024) introduced MISTRA, a novel approach that combines text and image features for misogyny detection, emphasizing the effectiveness of fusion models to detect subtle forms of misogyny in multimodal platforms like memes and images. Chinivar et al. (2024) proposed V-LTCS, focusing

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 249–253
May 3, 2025 ©2025 Association for Computational Linguistics

on the importance of selecting appropriate backbone networks for multimodal misogynous meme detection, which significantly impacts the performance of detecting misogynistic content in memes. Raja et al. (2023) applied a transfer learning approach with adaptive fine-tuning for fake news detection in Dravidian languages, offering a methodology that could be valuable for language-specific challenges in misogyny detection within multilingual contexts. Kumari et al. (2024) introduced M3Hop-CoT, employing a multimodal, multi-hop chain-of-thought process for meme identification, which emphasizes the contextual relationships between visual and textual elements for improved misogyny detection.

Srivastava (2024) proposed an early fusion model with graph networks for meme detection, demonstrating the advantages of combining multimodal features before processing to capture complex patterns in memes.Rizzi et al. (2023) discussed biases in misogyny detection models and stress the importance of addressing these biases for fairer and more equitable detection systems. Chakravarthi et al. (2024) provided a comprehensive evaluation of multitask meme classification, focusing on detecting both misogynistic and troll content in memes, aiming to improve classification accuracy and address various forms of online abuse.

Anzovino et al. (2018) explored automatic identification and classification of misogynistic language on Twitter, offering early insights into the challenge of detecting misogynistic content in text-based platforms. Plaza-Del-Arco et al. (2020) extended this work by focusing on the detection of misogyny and xenophobia in Spanish tweets, contributing to multilingual approaches in hate speech and misogyny detection. Frenda et al. (2019) investigated online hate speech against women, highlighting challenges in identifying misogyny and sexism on platforms like Twitter. Chakravarthi et al. (2025) presented the findings of the misogyny meme detection task in Dravidian languages at NAACL 2025, analyzing various machine learning and deep learning models. The study highlights dataset creation, annotation challenges, and the importance of multimodal approaches for effective detection.

Kiela et al. (2020) introduced the Hateful Memes Challenge, emphasizing the detection of hate speech in multimodal memes, which is a pivotal work in the multimodal analysis of harmful online content. Zhu (2020) enhanced multimodal transformer models for the Hateful Meme Challenge, showcasing how external labels and pretraining can improve meme classification accuracy. Zia et al. (2021) expanded on this by classifying memes beyond hate speech, tackling the challenge of detecting misogynistic and sexist memes specifically. Aloysius and Tamil Selvan (2023) addressed the reduction of false negatives in multi-class sentiment analysis, which is an important consideration for improving the accuracy of automated sentiment detection models, including those used for identifying misogynistic content.

3 Dataset Description

The dataset is made up of a series of pictures and a matching CSV file with three essential elements as given image id, labels and transcriptions. Each image in the folder is uniquely identified by its image id, which makes sure that every image is easily identifiable. The labels field indicates the type of content by classifying each image as either non-misogynistic or misogynistic. The textual information linked to each image is included in the transcriptions field, which gives the visual data context.

4 Methodology

Our model was developed as part of the DravidianLangTech 2025 Shared Task for detecting misogyny in Tamil memes. The task challenged participants to create robust multimodal models capable of processing both text and visual elements.

4.1 Dataset Preprocessing

Dataset preprocessing involved handling missing values, encoding labels, and splitting the dataset into training (1,136 memes), validation (284), and testing (356), totaling 1,776 memes. Text preprocessing included lowercasing, removing special characters, stopwords, and padding sequences. Image preprocessing involved resizing, standardizing pixel values, and data augmentation. These steps ensure standardized, noise-free input for optimal model performance. The dataset split is given below in Table 1.

4.1.1 Text Preprocessing

In order to prevent case sensitivity, all text was converted to lowercase, cleaning and standardizing the text data for model input. Only relevant Tamil and English characters were left after special characters, emojis, URLs, and other unnecessary

Dataset Split	Number of Memes
Training Set	1,136
Validation Set	284
Test Set	356
Total	1,776

Table 1: Dataset Description

symbols were eliminated using regular expressions. They removed stopwords such as "enna" (what) and "ipo" (now), and they padded or cleaned sequences to a specified length of 100. To expand the vocabulary and maintain uniformity throughout the dataset, words that appeared less than five times were substituted with placeholder tokens.

4.1.2 Image Preprocessing

Image data was standardized to ensure compatibility with the deep learning model. First, all images were resized to a target size of 224x224 pixels, which is the required input size for the EfficientNetB0 model. To enhance training efficiency, pixel values were normalized to a range of [0, 1] by dividing by 255, ensuring faster convergence during model training. Additionally, images were processed in batches, converting each image into an array and stacking them into a single dataset for efficient handling during training. These preprocessing steps ensured that the image data was optimized for input into the model.

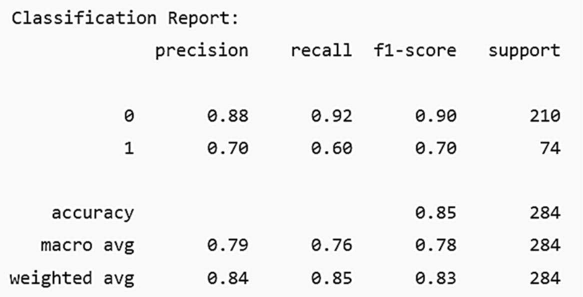
4.2 Models

4.2.1 Vision Language Model

A Vision-Language Model (VLM), which achieves 85% accuracy and a 70.8 F1-score, successfully combines text and images. Although accuracy indicates general correctness, efficiency could be improved by controlling class imbalance and increasing precision. The classification report is shown in Figure 1. VLMs excel in tasks like image captioning, visual quality assurance, and image-text matching, leveraging CNNs for images and Transformer-based models for text.

4.2.2 EfficientNet

EfficientNet optimizes accuracy and efficiency using compound scaling for tasks like object detection and image classification. With 82% accuracy and a 72.5 F1 score, it balances precision and recall well.The classification report is shown in Figure 2. Further tuning and data augmentation could enhance performance while maintaining reliability in reducing false predictions.

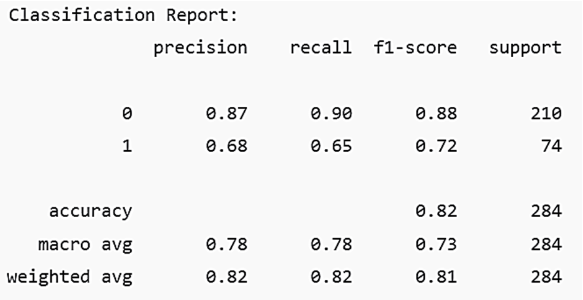


```
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.92      0.90       210
           1       0.70      0.60      0.70        74

    accuracy                           0.85       284
   macro avg       0.79      0.76      0.78       284
weighted avg       0.84      0.85      0.83       284
```

Figure 1: Classification Report for VLM



```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       210
           1       0.68      0.65      0.72        74

    accuracy                           0.82       284
   macro avg       0.78      0.78      0.73       284
weighted avg       0.82      0.82      0.81       284
```

Figure 2: Classification Report for EfficientNet

4.2.3 BERT

BERT enhances semantic understanding for NLP tasks using a bidirectional transformer. With 79.5% accuracy and a 65 F1 score, it performs well but struggles with class imbalance. Techniques like oversampling, class weighting, domain-specific models, hyperparameter tuning, and data augmentation can improve performance.

4.2.4 LSTM

LSTM, a type of RNN, excels in sequential tasks like NLP and time series forecasting by retaining long-term dependencies. With 75% accuracy and a 67.5 F1 score, it performs well but can improve on imbalanced data. Enhancements like more layers, bidirectional LSTMs, fine-tuning, class balancing, pre-trained embeddings, or attention mechanisms can boost performance.

5 Workflow

The workflow begins with text and image preprocessing, where BERT extracts textual features and CNN captures image features. These features are fused and fed into a training model comprising LSTM, EfficientNet, and VLM. LSTM processes sequential text patterns, EfficientNet refines image features, and VLM models text-image interactions. After training, the model is evaluated for accuracy,

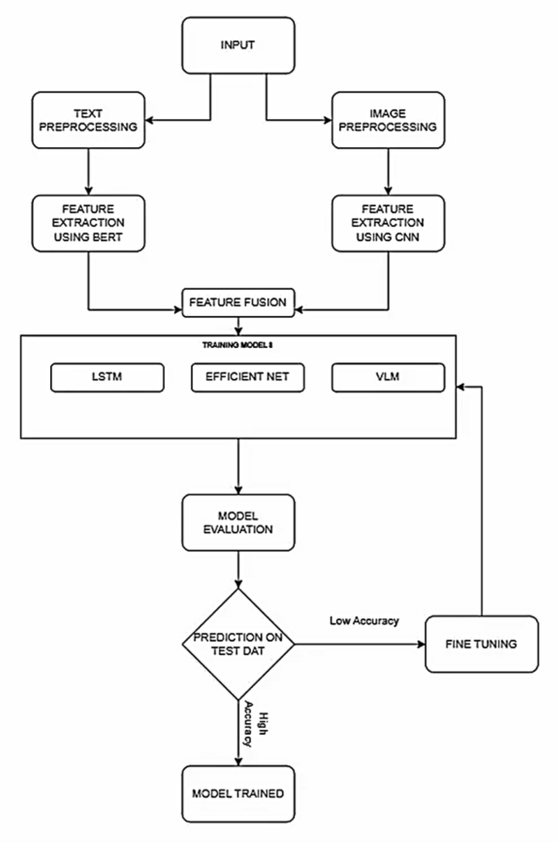


Figure 3: Proposed System Workflow

precision, recall, and F1-score. If accuracy is high, the model is finalized. Otherwise, fine-tuning is performed to enhance performance. Figure 3 illustrates this iterative process for effective misogyny detection in Dravidian language memes.

6 Result and Discussion

This study investigated the performance of many deep learning architectures for multimodal classification tasks, such as Vision-Language Models (VLM), BERT, LSTM, and EfficientNet. The comparison analysis demonstrates how various feature extraction and fusion techniques affect attaining peak performance.And our model ranked 12th out of 118 participants in the Tamil language track, showcasing its competitive performance in multimodal misogyny detection.The model accuracy performance comparision has been shown in Table 2.

VLM achieved the highest accuracy (85.0%), demonstrating its effectiveness in integrating textual and visual features. However, EfficientNet outperformed other models in F1-score (72.5%), indicating a better balance between precision and recall. While BERT excelled in text-based tasks, it lagged behind multimodal approaches. LSTM, though less accurate overall, maintained a competitive F1-score (67.5%), making it suitable for sequential feature extraction.The preprocessing of images and implementation details can be found in our GitHub repository InnovationEngineers Misogyny meme detection.

Models	Accuracy	F1-Score
VLM	85.0	70.8
EfficientNet	82.0	72.5
BERT	79.5	65.0
LSTM	75.0	67.5

Table 2: Model Performance Comparison

7 Conclusion

This study emphasized how Vision-Language Models (VLM) are a great help for multimodal categorization problems. VLM performs better than other models, exhibiting a superior capacity to integrate both textual and visual data with an accuracy of 85%. A better balance between precision and recall is demonstrated by EfficientNet's superior F1-score (72.5%), but VLM's total performance emphasizes the value of multimodal learning in improving classification accuracy. In sequential and text-based feature extraction tasks, BERT and LSTM offer useful insights, but multimodal techniques like VLM and EfficientNet outperform them.

References

C. Aloysius and P. Tamil Selvan. 2023. Reduction of false negatives in multi-class sentiment analysis. *Bulletin of Electrical Engineering and Informatics*, 12(2):1209–1218.

M. Anzovino, E. Fersini, and P. Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

B. R. Chakravarthi, S. Rajiakodi, R. Ponnusamy, K. Pannerselvam, A. K. Madasamy, R. Rajalakshmi, H. Ramakrishna Iyer Lekshmi Ammal, A. Kizhakkeparambil, S. S. Kumar, B. Sivagnanam, and C. Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the LT-EDI@EACL 2024*.

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.

S. Chinivar, M. S. R., J. S. A., and K. R. V. 2024. V-ltcs: Backbone exploration for multimodal misogynous meme detection. *Natural Language Processing*, 100109.

S. Frenda, B. Ghanem, M. Montes y Gómez, and P. Rosso. 2019. Online hate speech against women: automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing*, 100073.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

G. Kumari, K. Jain, and A. Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the SemEval-2022 Task 5 (MAMI Task)*.

F.-M. Plaza-Del-Arco, M.D. Molina-González, L.A. Ureña-López, and M.T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

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.

E. Raja, B. Soni, and S. K. Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 106877.

G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, and E. Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(6):103474.

H. Srivastava. 2024. Misogynistic meme detection using early fusion model with graph network. In *Proceedings of the SemEval-2022 Task 5*.

R. Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv:2012.08290 [cs.CL]*.

H. B. Zia, I. Castro, and G. Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the Workshop on Online Abuse and Harms (WOAH 2021)*, pages 1–10.