Code_Conquerors@DravidianLangTech 2025: Multimodal Misogyny Detection in Dravidian Languages Using Vision Transformer and BERT

Pathange Omkareshwara Rao, Harish Vijay V, Ippatapu Venkata Srichandra, Neethu Mohan, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4aie22039@cb.students.amrita.edu, harishvijay0204@gmail.com,
ippatapuvenkatasrichandra@gmail.com, s_sachinkumar@cb.amrita.edu

Abstract

This research focuses on misogyny detection in Dravidian languages using multimodal techniques. It leverages advanced machine learning models, including Vision Transformers (ViT) for image analysis and BERT-based transformers for text processing. The study highlights the challenges of working with regional datasets and addresses these with innovative preprocessing and model training strategies. The evaluation reveals significant improvements in detection accuracy, showcasing the potential of multimodal approaches in combating online abuse in underrepresented languages.

Keywords: Misogyny Detection, Dravidian Languages, Tamil, Malayalam, Multimodal Techniques, Vision Transformers (ViT), BERT-Based Models, Regional Datasets, Machine Learning, Online Abuse.

1 Introduction

The study explores the underrepresented area of misogyny detection in Dravidian languages, particularly Tamil and Malayalam. Online misogyny is a growing social issue, with an increasing volume of content targeting women on digital platforms.This work was submitted to the Misogyny Meme Detection - DravidianLangTech@NAACL 2025 competition, organized by DravidianLangTech, where it achieved a **rank of 13 in Malayalam and 15 in Tamil languages**, demonstrating the effectiveness of the proposed multimodal approach in detecting misogyny in Dravidian languages. Existing methods often fail to address the nuances of regional languages and multimodal content. This research builds upon prior works that used unimodal or traditional models and introduces a robust multimodal approach. By fusing features from visual and textual data, the proposed method seeks to improve detection accuracy. The paper presents a detailed examination of the challenges, including data inconsistencies and the need for advanced preprocessing techniques, while providing solutions to these issues.Code available at: GitHub repository.

2 Literature Review

Another methodology was proposed by Gu et al. (2022a), where they introduced a multi-modal and multi-task Variational Autoencoder (VAE) on the same dataset. Their method achieved F1-scores of 0.72 for Task A and 0.634 for Task B. Despite these encouraging results, the study lacked statistical analysis of the model's outcomes, which limited its scope for deeper insights.The work by Singh et al. (2023) investigated the MAMI dataset using various state-of-the-art models. They combined a pretrained BERT model with ViT, achieving an F1-score of 0.874. However, the methodology faced limitations such as the lack of interpretability and the extended time required for training and deployment.

The authors Mahesh et al. (2024) focused on the LT-EDI @EACL 2024 dataset, applying mBERT+ResNet-50 and MuRIL+ResNet-50 models. These approaches achieved F1-scores of 0.73 and 0.87 for Tamil and Malayalam datasets, respectively.

Another methodology proposed by Gu et al. (2022b) combined joint image and text classification, resulting in a macro F1-score of 0.665. However, their work did not explore varying threshold values or class weights, and they struggled to develop an effective model that leveraged image data effectively.The authors Shanmugavadivel et al. (2023) explored abusive comment detection data and applied various machine learning, deep learn-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 361–365

ing, and transformer-based approaches. Among their methods, the Random Forest model achieved a macro F1-score of 0.42. A key limitation of their work was the lack of experimentation with different pretrained BERT models.

Another methodology was proposed by Shaun et al. (2024), who used a multinomial Naive Bayes approach for textual data and a ResNet-50 model for pictorial data. Their approach on Tamil and Malayalam datasets achieved an F1-score of 0.82. Similarly, Koutlis et al. (2023) introduced a new deep-learning-based architecture called Memefier, which was tested on datasets such as Facebook Hateful Memes, Memotions 7k, and MultiOFF.

Finally, another methodology was developed by Boinepelli et al. (2020), who applied various machine learning and deep learning methods to the SemEval-2020 dataset. Their CNN-LSTM model achieved an F1-score of 59.04

3 Data Description

The dataset used for the misogyny detection task focuses on two Dravidian languages: Tamil Ponnusamy et al. (2024) Chakravarthi et al. (2024) and Malayalam Ponnusamy et al. (2024) Chakravarthi et al. (2024). It is divided into development and training datasets, each containing images and corresponding metadata in Excel sheets. The metadata includes three columns: `image_id`, `label`, and `transcription`. Here, label '1' indicates misogynistic content, while label '0' represents non-misogynistic content. However, discrepancies were observed in the original dataset between the number of images in the folders and the entries in the Excel sheets. To address this, consolidated datasets were created by aligning the images with their corresponding metadata. Below, we describe both the original and consolidated datasets.

3.1 Original Development Dataset

Folder	Images	Metadata	Labels
Malayalam	160	160	0: 97, 1: 63
Tamil	282	284	0: 210, 1: 74

Table 1: Details of the original development dataset.

Kumar et al. (2017) The table-1 contained separate folders for Tamil and Malayalam memes. Each folder consisted of images and metadata entries, with slight inconsistencies in the Tamil dataset between the image folder count and Excel sheet count.

3.2 Original Training Dataset

Folder	Images	Metadata	Labels
Malayalam	639	640	0: 381, 1: 259
Tamil	1135	1136	0: 851, 1: 285

Table 2: Details of the original training dataset.

The table-2 showed a larger discrepancy, particularly in the Tamil subset, with 1135 images in the folder but 1136 entries in the metadata. Despite this, the dataset provided a robust training foundation.

3.3 Consolidated Dataset

Folder	Images	Metadata	Labels
Malayalam	460	460	0: 252, 1: 168
Tamil	787	787	0: 591, 1: 196

Table 3: Details of the consolidated dataset.

The inconsistencies in the original dataset were resolved by aligning images with their respective metadata. The table-3 forms the primary input for model training and evaluation (Chakravarthi et al., 2025). The figure-1 illustrates the overall architecture of the multimodal approach used for misogyny detection in this research. The model takes in both image and text data, which undergo respective preprocessing steps such as resizing, normalization, tokenization, and padding. The preprocessed data is then fed into specialized feature extractors - Vision Transformer (ViT) for images and BERT/RoBERTa for text. The extracted features are then fused and passed through a classification layer to predict whether the input is misogynistic or non-misogynistic. The details of the individual components and training setup are described in the Methodology section.

4 Proposed Methodology

The methodology integrates preprocessing, training setup, and architecture to detect misogynistic content in memes effectively. The figure-1 shows the proposed methodology.

4.1 Architecture

Our multimodal approach processes image and text data simultaneously. The Vision Transformer Remya et al. (2024) serves as the image encoder, processing image patches as sequences for global context extraction. For text encoding, BERT (Malay-

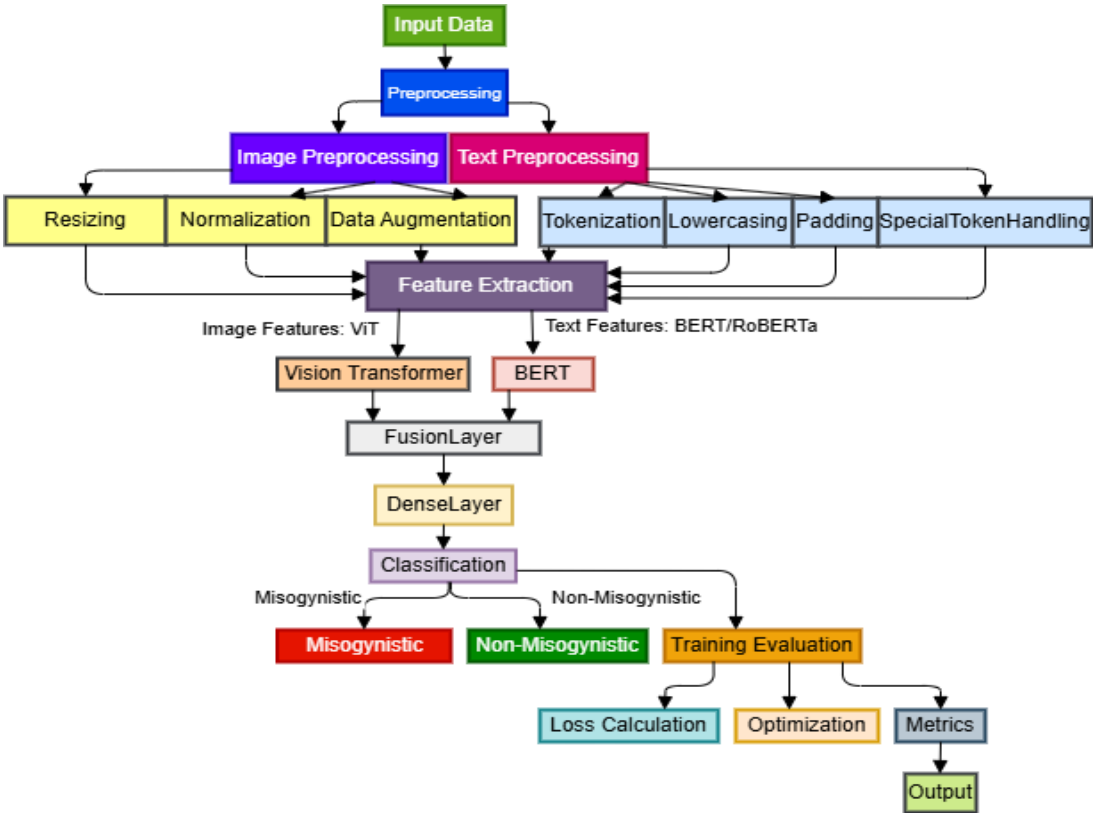

Figure 1: Overview of the Multimodal Misogyny Detection Model.

Image Model	Text Model	F1 Score (Malayalam)	F1 Score (Tamil)	Data Type
ViT	BERT-base-uncased	**0.882251**	**0.822728**	Validation
Swin Transformer	XLM-RoBERTa	0.881944	0.808841	Validation
ResNet50	RoBERTa-base	0.822145	0.797273	Validation
EfficientNet-B3	XLM-RoBERTa	0.779364	0.789389	Validation
EfficientNet-B0	DeBERTa-base	0.757143	0.779304	Validation

Table 4: Performance comparison of top models for Malayalam and Tamil datasets.

alamAbeera et al. (2023)) or RoBERTa (Tamil) extracts semantic features from captions. The features from both encoders are concatenated through a fully connected layer, forming a unified representation that feeds into a classification layer for binary prediction (misogynistic/non-misogynistic).

4.2 Preprocessing

4.2.1 Image Preprocessing

Images are resized to 224x224 pixels, normalized to [0,1], and augmented using random rotations, flips, color jittering, and cropping to prevent overfitting.

4.2.2 Text Preprocessing

Text undergoes tokenization with model-specific tokenizers (BERT, RoBERTa), uniform padding, lowercasing, and special token addition ([CLS], [SEP]).

4.3 Training Setup

Training utilizes CUDA-enabled GPUs with 15 epochs, learning rates of 2×10^{-5} (text) and 3×10^{-4} (image), binary cross-entropy loss, Adam optimizer, and batch size of 32.

4.4 Models

For Malayalam, **ViT + BERT-base-uncased** achieved the highest F1 score of **0.882251**, while for Tamil, **ViT + RoBERTa-base** scored **0.822728**.

5 Results and Discussion

5.1 Experimental Setup

The model hyperparameters (table-5) were selected through experimentation to balance performance and efficiency. A learning rate of 2e-5 was chosen for the AdamW optimizer, ensuring stable convergence for BERT-based text encoders, while a higher rate of 3e-4 was used for image encoders

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	16
Dropout Rate (Fusion)	0.5
Dropout Rate (Classifier)	0.3
Fusion Dimension	512
Weight Decay	AdamW
Number of Epochs	15
Scheduler Factor	0.5
Scheduler Patience	2

Table 5: Model Hyperparameters.

(ViT/ResNet-50) to accommodate their larger parameter space. The batch size of 16 was tailored to GPU memory constraints, maintaining gradient stability. The fusion module combines text and visual embeddings into a 512-dimensional space, with a dropout rate of 0.5 to prevent overfitting. The classifier applies an additional dropout rate of 0.3 to enhance generalization. These dropout rates were empirically determined to balance regularization and feature retention. Training was conducted for 15 epochs, with a ReduceLROnPlateau scheduler reducing the learning rate by half if the validation F1-score stagnated for two epochs. This adaptive approach ensured robust convergence. Ablation studies confirmed that alternative configurations (e.g., higher dropout or larger batch sizes) resulted in lower validation F1-scores, validating the chosen hyperparameters.

5.2 Results

Data	Malayalam	Tamil
Validation	0.9115	0.8079
Test	0.75649	0.66142

Table 6: Model Performance Across Languages. *Note: All values represent F1 scores.*

The table-6 shows our evaluation of misogyny detection models, we focused on the F1 score as the primary performance metric. Our experimentation involved multimodal combinations, including ResNet-50 + BERT-base uncased and Vision Transformer (ViT) + BERT-base uncased, to assess their effectiveness in detecting misogyny in text and images.For the Tamil language Selvan et al. (2015), the macro F1 scores obtained were: the ResNet-50 + BERT-base uncased combination achieved a validation F1 score of 0.8079 and a test F1 score of 0.66142. In the case of Malayalam, the Vision Transformer (ViT) + BERT-base uncased combination achieved a validation F1 score of 0.9115 and a test F1 score of 0.75649.

5.3 Performance Comparison:

The performance of the top models for Malayalam and Tamil datasets is summarized in Table-4. The ViT + BERT-base-uncased model achieved the highest validation F1 score of 0.882 for Malayalam and 0.823 for Tamil. However, the test performance (shown in Table-6) revealed a drop in F1 scores, indicating potential overfitting or a domain gap between the training and test datasets.

6 Discussion

From Table-4 we could analyse that Tamil, ResNet-50 + BERT-base uncased performed best, but the drop in test performance suggests overfitting or a domain gap. For Malayalam, ViT + BERT-base uncased showed consistent performance, highlighting ViT's ability to capture visual details and BERT's multilingual text processing. The results emphasize the importance of selecting appropriate models for multimodal tasks and reaffirm the potential of multimodal approaches in addressing misogyny detection in regional languages. Upon examining misclassified cases, we observed that the model struggled with ambiguous memes where the text and image conveyed conflicting messages. For instance, memes with sarcastic or culturally specific humor were often misclassified. This highlights the need for better contextual understanding and cultural nuance in future models.

7 Conclusion

This study enhances misogyny detection in Dravidian languages using a multimodal approach. ResNet-50 + BERT-base uncased achieved F1 scores of 0.8079 (validation) and 0.66142 (test) for Tamil, while ViT + BERT-base uncased scored 0.9115 and 0.75649 for Malayalam. ViT captured visual patterns effectively, and BERT-base uncased handled multilingual text well. The performance gap highlights generalization challenges. Future work includes leveraging mBERT/IndicBERT, reducing domain gaps, and refining multimodal fusion to foster inclusive digital spaces.

8 Limitations

Our proposed methodology faced limitations in handling different domains, such as Tamil and Malayalam. mBERT or IndicBERT could have provided better contextual understanding and improved generalization.

References

Yimeng Gu, Ivan Castro, and Gareth Tyson. Mmvae at semeval-2022 task 5: A multi-modal multi-task vae on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710, 2022a.

S. Singh, A. Haridasan, and R. Mooney. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, 2023.

S. Mahesh et al. 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, 2024.

Q. Gu, N. Meisinger, and A.-K. Dick. 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, 2022b.

K. Shanmugavadivel et al. Kec_ai_nlp@dravidianlangtech: Abusive comment detection in tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299, 2023.

H. Shaun et al. 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, 2024.

C. Koutlis, M. Schinas, and S. Papadopoulos. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 586–591, 2023.

S. Boinepelli, M. Shrivastava, and V. Varma. Sis@iiith at semeval-2020 task 8: An overview of simple text classification methods for meme analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1190–1194, 2020.

R. Ponnusamy et al. 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, 2024.

B.R. Chakravarthi et al. 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, 2024.

S. S. Kumar, M. A. Kumar, and K. P. Soman. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings 5*, pages 320–334. Springer International Publishing, 2017.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 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, 2025.

S. Remya, T. Anjali, S. Abhishek, S. Ramasubbareddy, and Y. Cho. The power of vision transformers and acoustic sensors for cotton pest detection. *IEEE Open Journal of the Computer Society*, 2024.

V. P. Abeera, S. Kumar, and K. P. Soman. Social media data analysis for malayalam youtube comments: Sentiment analysis and emotion detection using ml and dl models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, September 2023.

A. Selvan, M. Anand Kumar, and K. P. Soman. Sentiment analysis of tamil movie reviews via feature frequency count. In *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS 15)*. IEEE, 2015.