TEAM_STRIKERS@DravidianLangTech2025: Misogyny Meme Detection in Tamil Using Multimodal Deep Learning

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2], **Mohamed Arsath H**[1], **Ramya K**[1], **Ragav R**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{mohamedarsathh.22aid, ramyak.22aid}@kongu.edu
ragavr.22aid@kongu.edu

Abstract

This study focuses on detecting misogynistic content in memes under the title Misogynistic Meme Detection Using Multimodal Deep Learning. Through an analysis of both textual and visual components of memes, specifically in Tamil, the study seeks to detect misogynistic rhetoric directed towards women. Preprocessing and vectorizing text data using methods like TF-IDF, GloVe, Word2Vec, and transformer-based embeddings like BERT are all part of the textual analysis process. Deep learning models like ResNet and EfficientNet are used to extract significant image attributes for the visual component. To improve classification performance, these characteristics are then combined in a multimodal framework employing hybrid architectures such as CNN-LSTM, GRU-EfficientNet, and ResNet-BERT. The classification of memes as misogynistic or non-misogynistic is done using sophisticated machine learning and deep learning approaches. Model performance is evaluated using metrics like as Accuracy, Precision, Recall, F1-Score, and Macro Average F1-Score. This study shows how multimodal deep learning can effectively detect and counteract negative narratives about women in digital media by combining natural language processing with image classification.

1 Introduction

The causes that propel the terrible part of misogyny online are inciting animosity and discriminating ideas toward women. Online communication is dominated by memes, which are extremely hard to spot because of the way they use both overt and covert misogynistic rhetoric in their words and visuals. Therefore, it is necessary to spot misogynistic material in memes. With an emphasis on examining both linguistic and visual elements, this study aims to categorize Tamil memes as either misogynistic or non-misogynistic. Misogynistic terms and phrases hidden in meme material will be detected using sophisticated natural language processing algorithms. Textual input is converted into intelligible representations for classification using transformer-based models like BERT and vectorization techniques like TF-IDF, GloVe, and word2vec. Relevant image features will be extracted for the visual analysis using deep learning models such as ResNet and EfficientNet. Modern multimodal deep learning models are integrated in the study, and metrics like Accuracy, Precision, Recall, F1-Score, and Macro F1-Score are used to assess how well they perform. This exacting approach guarantees a thorough comprehension of linguistic subtleties and cultural settings. The study advances AI strategies for addressing misogyny in regional languages by concentrating on Tamil. The groundwork for future initiatives to combat gender-based discrimination in online spaces is laid by this work, which also emphasizes the value of automated technologies in promoting digital civility.

2 Literature Survey

Cuervo and Parde (2022) highlights the paucity of research on multimodal systems intended to identify misogynistic content. Misogynistic memes are a common problem on social media that combine graphics with disparaging text to spread damaging messages. The authors apply contrastive learning in the context of SemEval 2022 Task 5, which is concerned with identifying sexist memes, by utilizing OpenAI's CLIP model, which is well-known for its efficacy in multimodal tasks. Even if the built model doesn't perform at its best, the tests offer insightful exploratory information that advances our knowledge of how to identify misogynistic content in memes.

Rizzi et al. (2023) explore methods for identifying misogynistic content in social media memes. The study highlights the introduction of a bias es-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 671–675

timation technique and a Bayesian optimization strategy, resulting in a 61.43% improvement in prediction accuracy. They assess multiple unimodal and multimodal approaches, addressing challenges associated with specific meme archetypes. The authors emphasize the necessity for further research to mitigate model biases and enhance detection techniques.

Multimodal hate content detection has become more difficult due to the development of misogynistic memes. Multimodal misogyny is more difficult to detect than text-based sexism, even when using balanced datasets such as MAMI, which has 12,000 annotated memes Singh et al. (2023). Even with improvements, contextual ambiguity remains a challenge for models. However, performance is greatly enhanced by domain-specific pretraining, especially when BERT is used in conjunction with attention-based techniques.

In today's digital age, memes have emerged as a popular medium for online expression, humor, sarcasm, and social commentary. Yet, beneath their surface, they often carry troubling elements like misogyny, gender-based bias, and harmful stereotypes. To address these issues, Ponnusamy et al. (2024) delves into the world of online misogyny among Tamil and Malayalam-speaking communities, creating an annotated dataset with comprehensive guidelines. By analyzing memes, the authors reveal the complexities of gender bias and stereotypes, highlighting their manifestations and impact. This dataset and its detailed guidelines are invaluable resources for understanding the prevalence, origins, and nuances of misogyny, aiding researchers, policymakers, and organizations in formulating effective strategies to combat gender-based discrimination and promote equality and inclusivity. It provides profound insights that inform strategies for fostering a more equitable and safe online environment. This study is a crucial step in raising awareness and tackling gender-based discrimination in the digital realm.

A thorough summary of the shared work at LT-EDI@EACL 2024, which focused on classifying troll memes and misogynistic content in Tamil and Malayalam, may be found at Chakravarthi et al. (2024). According to the study, 52 teams entered the competition, and four systems—three for Malayalam and four for Tamil meme classification—were presented. The results of the shared work highlight the prevalence of troll and misogynistic content on the internet today and investigate the computational methods used to identify it. For Tamil and Malayalam, the best-performing model received macro F1 scores of 0.73 and 0.87, respectively.

Online sexism is a serious social problem that turns digital spaces into unfriendly places for women. In order to overcome this difficulty, Hashmi et al. (2024) suggest a strategy for identifying misogynous content in bilingual (English and Italian) online conversations. Explainable artificial intelligence (LIME), multilingual fine-tuned transformers, and FastText word embeddings are all used in the study. Their strategy performs better on important measures like accuracy and F1-score, demonstrating the possibility of cutting-edge approaches to successfully tackle online misogyny.

Angeline et al. (2022) combine BERT embeddings with Long Short-Term Memory (LSTM) networks to present a novel method for identifying sexist discourse on social media. While LSTM achieves an accuracy of 86.15% in capturing long-term dependencies in text, their study focuses on the contextual interpretation of tweets using BERT. This demonstrates how well deep learning models detect hazardous information on Twitter and other networks.

Habash et al. (2022) created a deep learning system to identify misogynistic memes. VisualBERT and two MMBT models are among the ensemble of multi-modal models used by the system. The two subtasks it tackles are identifying sexist memes (sub-task A) and classifying them into four categories: violence, objectification, shaming, and stereotyping (sub-task B). In sub-task A, their system outperformed the baseline model by a wide margin with an F1-score of 0.722.

Mahadevan et al. (2022) suggested a feature extraction-based method for detecting misogynous memes utilizing transformer models such as BERT and RoBERTa. By combining textual and visual components from memes, their multimodal training approach improves misogyny detection. The system demonstrated its efficacy in practical applications by performing remarkably well, placing fourth in Subtask A and ninth in Subtask B.

Pro-Cap, a novel probing-based captioning technique for hostile meme detection, was presented by Cao et al. (2023). It uses pre-trained vision-language models (PVLMs) without any fine-tuning. Pro-Cap generates image captions that contain crucial information for identifying hateful content by leveraging queries linked to hateful content to pro-

voke a frozen PVLM. This method's effectiveness and generalizability were validated by its good performance on three benchmarks.

3 Task Description

The shared task focuses on detecting misogynistic content in Tamil-language memes. Our objective is to categorize a dataset of misogynistic and non-misogynistic memes into two groups: Misogynistic and Non-Misogynistic. In order to develop models that can precisely detect misogynistic content while addressing issues specific to Tamil memes, this work entails assessing both textual and visual features. Chakravarthi et al. (2025) Our system showcased its performance in this demanding multimodal classification test by securing 17th place out of 118 participants.

4 Dataset Description

The dataset contains a total of 1,776 memes, split into three subsets 1,136 for training, 356 for testing, and 284 for development make up the dataset employed in this study. Three essential elements are present in every meme in the dataset: an image ID, a label, and transcriptions. The label denotes whether the meme is categorized as Misogynistic or Non-Misogynistic the picture ID is the unique identifier for each meme image, and the transcriptions are the textual material included in the meme. In order to aid in the creation and assessment of machine learning models intended to identify harmful online content, this dataset was created especially for the categorization of misogynistic content in memes, with an emphasis on Tamil language memes.

Dataset	No. of Memes
Train	1136
Validation	284
Test	356

Table 1: Tamil Dataset Description

5 Data Pre-processing

The data preprocessing pipeline for Misogynistic Meme Detection in Tamil using Multimodal Deep Learning is designed to efficiently handle both textual and visual data: text preprocessing involves converting Tamil text to lowercase, then removing URLs, HTML tags, special characters, and

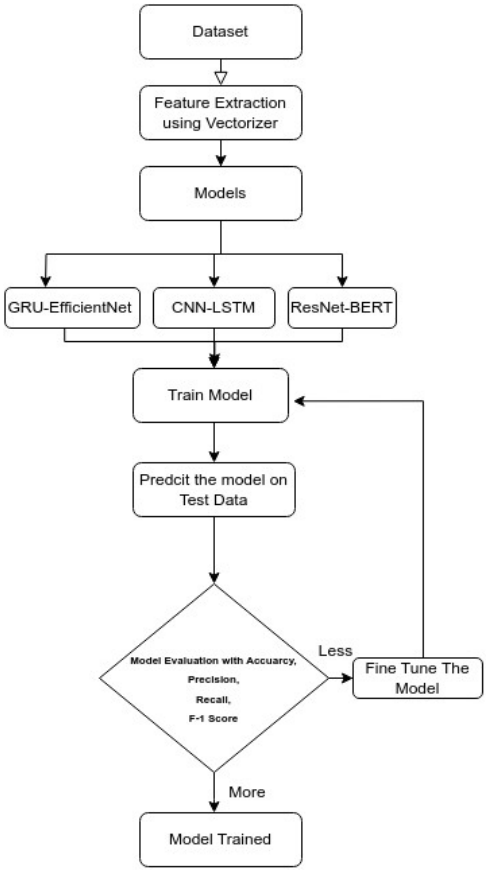

Figure 1: Proposed System Workflow

numbers to preserve meaningful content; tokenization is carried out using a Tamil-specific tokenizer, and stopword removal is applied using a curated list of Tamil stopwords; Tamil is a morphologically rich language, so stemming and lemmatization are handled using TamilMorph to normalize words; short words and duplicate words within a sentence are eliminated to reduce noise; image preprocessing involves resizing meme images to a fixed size, normalizing pixel values, and applying data augmentation techniques like rotation, flipping, and brightness adjustments to improve generalization.Tesseract OCR is used to extract Tamil text embedded in images using OCR preprocessing, and the retrieved text is then cleaned using the same procedure. Lastly, multimodal preprocessing improves the accuracy of sexist content detection in Tamil memes by combining linguistic and visual data to guarantee high-quality input for deep learning models.

6 Model Evaluation

The objective of this project is to classify memes as Misogynistic or Non-Misogynistic using a multimodal approach, incorporating both textual and

visual components. The trained model combines Natural Language Processing (NLP) techniques with computer vision methods to process and classify memes.

This experiment used a multimodal deep learning system that combined computer vision and NLP techniques to classify memes as misogynistic or non-misogynistic. CNNs like ResNet and EfficientNet extracted image features, while LSTM and GRU models processed Tamil-English code-mixed text with vectorization techniques like TF-IDF, GloVe, and Word2Vec. The CNN-LSTM model achieved the highest accuracy of 77.1%, leveraging CNN's spatial feature extraction and LSTM's ability to capture long-range text dependencies. The GRU-EfficientNet model followed with 76.8% accuracy, while the ResNet-BERT model scored 72.9%, likely due to BERT's limitations with code-mixed text. Models relying solely on text or image features failed to capture subtle contextual aspects, especially in cases involving sarcasm or culturally specific expressions. Despite these challenges, multimodal fusion effectively captured both textual and visual cues, proving to be a powerful strategy for detecting harmful online content, as demonstrated by accuracy, precision, recall, and macro average F1-score.

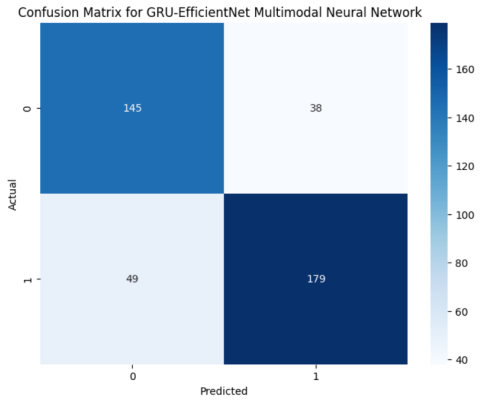

Figure 2: Confusion Matrix for GRU-EfficientNet

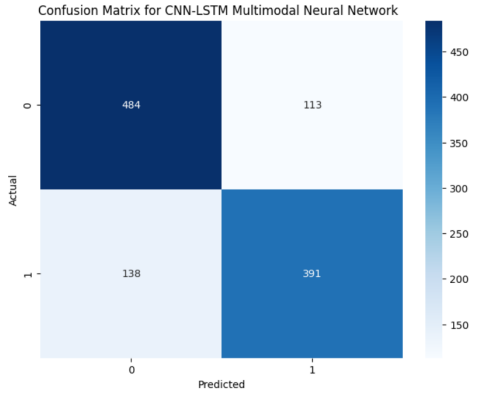

Figure 3: Confusion Matrix for CNN-LSTM

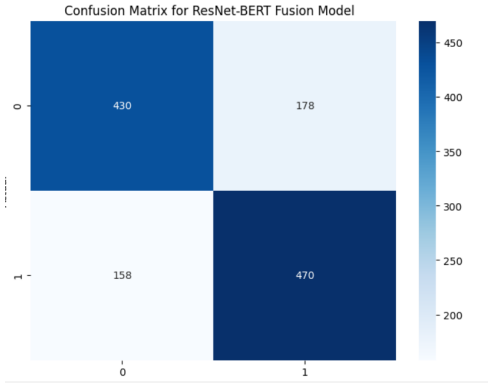

Figure 4: Confusion Matrix for ResNet-BERT

Model	Accuracy	F1 Score
GRU-EfficientNet Multimodal Neural Network	76.8%	0.62
CNN-LSTM Multimodal Neural Network	77.1%	0.68
ResNet-BERT Fusion Model	72.9%	0.63

Table 2: Model Accuracy and Macro Average F1 Score

7 Limitations

Despite encouraging outcomes, the models had trouble with complex situations where feature fusion was occasionally insufficient, such as irony, subtle misogyny, and culturally distinctive emotions. The reduced accuracy of the ResNet-BERT model was probably caused by BERT's shortcomings when it came to code-mixed Tamil-English text that lacked domain-specific fine-tuning. Furthermore, the models' reliance on huge datasets and powerful computers may restrict their scalability. By using contrastive learning strategies, bigger and more varied datasets, and refined transformer-based models specifically designed for Tamil social media material, future research can overcome these constraints.

8 Conclusion

This project combined picture classification and natural language processing (NLP) to assess both visual and linguistic features, resulting in a multimodal deep learning framework for misogynistic meme detection. The CNN-LSTM Multimodal Neural Network achieved the highest accuracy of 77.1% out of all the architectures we examined, followed by the GRU-EfficientNet model. Accuracy and macro F1-score demonstrated that models that combined text and image features performed better than single-modal approaches. These findings demonstrate the effectiveness of multimodal learning for challenging categorization tasks and show how sophisticated deep learning techniques can enhance the detection of dangerous content.

References

R. S . Angeline, D. Nurjanah, and H. Nurrahmi. 2022. Misogyny speech detection using long short-term memory and bert embeddings. In *2022 4th International Conference on Informatics and Computational Sciences (ICOIACT)*, pages 155–159.

R. Cao, M. S. Hee, A. Kuek, W. H. Chong, R. K.-W. Lee, and J. Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection.

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.

Cuervo and N. Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. pages 785–792.

M. Habash, Y. Daqour, M. Abdullah, and M. Al-Ayyoub. 2022. Ymai at semeval-2022 task 5: Detecting misogyny in memes using visualbert and mmbt multimodal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784.

E Hashmi, S. Y. Yayilgan, M. M. Yamin, and M. Ullah. 2024. Enhancing misogyny detection in bilingual texts using explainable ai and multilingual fine-tuned transformers. *Complex & Intelligent Systems*, 11(1).

S. Mahadevan, S. Benhur, R. Nayak, M. Subramanian, K. Shanmugavadivel, K. Sivanraju, and B. R. Chakravarthi. 2022. Transformers at semeval-2022 task 5: A feature extraction based approach for misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 550–554.

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.

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

Singh, A. Haridasan, and R. J. Mooney. 2023. Female astronaut: Because sandwiches won't make themselves up there: Towards multimodal misogyny detection in memes.