MNLP@DravidianLangTech 2025: Transformer-based Multimodal Framework for Misogyny Meme Detection

Shraddha Chauhan
Department of ECE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
shraddha.20224147@mnnit.ac.in

Abhinav Kumar
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

A meme is essentially an artefact of content-usually an amalgamation of a image and text that spreads like wildfire on the internet, usually shared for amusement, cultural expression, or commentary. They are very much similar to an inside joke or a cultural snapshot that reflects shared ideas, emotions, or social commentary, remodulated and reformed by communities. Some of them carry harmful content, such as misogyny. A misogynistic meme is social commentary that espouses negative stereotypes, prejudice, or hatred against women. The detection and addressing of such content will help make the online space inclusive and respectful. The work focuses on developing a multimodal approach for categorizing misogynistic and non-misogynistic memes through the use of pretrained XLM-RoBERTa to draw text features and Vision Transformer to draw image features. The combination of both text and images features are processed into a machine learning and deep learning model which have attained F_1-scores 0.77 and 0.88, respectively Tamil and Malayalam for misogynist Meme Dataset.

1 Introduction

Memes have become one of the most powerful ways of expressing on social media, often full of humor, satire, and cultural commentary. Seemingly innocuous in nature, memes can be a means of disseminating harmful content (Weber et al., 2020) that enforces gender biases and stereotypes. Thus, it is crucial to identify such content in order to bring about a more respectful and inclusive digital space (Rao and Kalyani, 2022; Kumar et al., 2021). The classification of misogyny in Tamil and Malayalam memes faces different types of challenges because of linguistic as well as cultural characteristics of these languages (Fersini et al., 2022). As Dravidian languages are low resource languages which possess unique syntactic structures and vocabulary and idiomatic expressions (Singh et al., 2025). There is a scarcity of labeled datasets within these languages, which complicates the training of Machine and Deep learning models. Lack of datasets makes it hard to develop effective machine learning models because they require enough good-quality and labelled data to learn complex patterns of misogyny in social media memes.

Adding text and images to memes makes it even more complex as its often the use of visual elements to relay context and meaning which might be hard for text-only models to understand. The multimodal approach is necessitated to capture the complexity between the textual and visual components of memes (H et al., 2024). With deep learning models such as transformers in text and image feature extraction, along with feature fusion strategies, it is possible to raise the classification accuracy of misogynistic memes. Therefore, this work proposes a multimodal deep learning-based model for the identification of misogynist memes.

The rest of the paper is arranged as follows: Section 2 contains related work, Section 3 discusses the dataset & task, and Section 4 presents the proposed methodology. In Section 5, the outcome of the proposed model is listed and concluded in Section 6, and Section 7 details the limitations of the proposed framework.

2 Related Work

Detection of misogynistic content in memes in low-resource languages has started to gain greater attention in the recent past (Kumar et al., 2021; Ponnusamy et al., 2024). Memes that convey meaning using textual and visual elements introduce an interesting challenge (Priyadharshini et al., 2022). In recent years, several studies focused on multimodal approaches that have combined text as well as image data for classifying hate speech and misogynistic content. Recent studies have adapted techniques like data augmentation or transfer learning

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 335–340
May 3, 2025 ©2025 Association for Computational Linguistics

from related languages to improve the models' performance on those low-resource languages (Joshi et al., 2020). For the detection of misogyny in meme analysis, datasets play an important role. A prominent MIMIC dataset (Singh et al., 2024) focuses on low-resource Hindi-English code-mixed language, thus allowing the detection of misogyny. A study by (Rizzi et al., 2023) explores various approaches in detecting misogynistic content in memes. They compares different approaches in the integration of textual and visual data, one unimodal and the other multimodal approach, respectively.

Multimodal approaches (Jindal et al., 2024) are vital while considering memes since they are based on the visual component and textual information. Convolutional Neural Networks were traditionally applied to extract features from images. The recent alternative for image processing has been the Vision Transformers (Dosovitskiy et al., 2014). Deep learning-based techniques are used in various studies like (Garcia et al., 2021), where the use of textual and visual information enhances performance. The study (Kiela et al., 2019) proved the multimodal embeddings effectiveness in cross-modal retrieval and classification tasks, where textual and visual features are combined to improve the understanding of meme content. Misogyny in memes is challenging due to the complexities present in meme content, which often involves humor, cultural references, and contextual elements (Chakravarthi et al., 2024). Traditional machine learning models struggle to capture these complexity. Study by (Suryawanshi et al., 2020) and (Beigi et al., 2020) addresses these challenges by incorporating both image and text features for detecting hate speech. However, these approaches face limitations due to the paucity of labeled data in Tamil and Malayalam, and labeled datasets are often small and imbalanced.

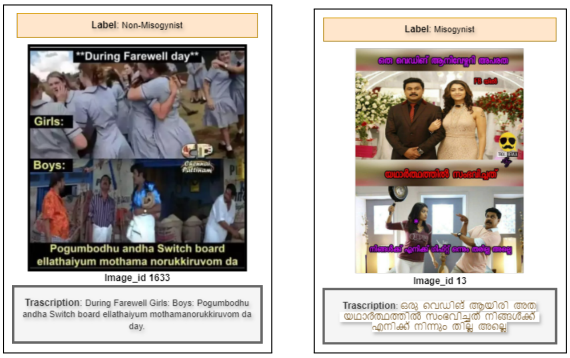
Data imbalance is a massive concern for various tasks and becomes more acute in the context of misogynistic meme detection (Hossain et al., 2022; Gasparini et al., 2022). Several methods have been proposed to address such imbalance, such as oversampling the minority class or applying cost-sensitive learning techniques (Buda et al., 2018). Synthetic data generation and semi-supervised learning have been explored in order to enhance classifier performance in cases where the training data is limited (Zhang et al., 2020). These techniques can be specifically helpful when used with meme datasets in low-resource languages, mitigating the imbalance problem of the data.

3 Dataset & Task

The Tamil dataset distribution consists of 1,133 memes for training and 356 memes for testing purposes and Malayalam dataset consists of 640 training and 200 test data are shown in Table 1. Each meme in the dataset contains both pictorial content and overlaid text (Chakravarthi et al., 2025). The memes are classified based on the presence or absence of misogyny. However, the dataset is imbalanced, with a significantly higher number of non-misogynistic memes compared to misogynistic ones for both Tamil and Malayalam datasets.

Table 1: Datasets and their distribution.

Dataset	Label	Train	Val	Test
Tamil	Misogynistic	285	74	89
	Non-Misogynistic	848	210	267
Total		1,133	284	356
Malayalam	Misogynistic	259	63	78
	Non-Misogynistic	381	96	122
Total		640	159	200


Label: Non-Misogynist
During Farewell day
Girls:
Boys:
Pogumbodhu andha Switch board ellathaiyum mothama norukkiruvom da
Image_id 1633
Label: Misogynist
ഒരു വെഡിങ് ആനിവേഴ്സറി അപരത
FB യിൽ
യഥാർത്ഥത്തിൽ സംഭവിച്ചത്
നിങ്ങൾക്ക് എനിക്ക് ഗിഫ്റ്റ് ഒന്നും തരില്ല അല്ലെ
Image_id 13

Trascription: During Farewell Girls: Boys: Pogumbodhu andha Switch board ellathaiyum mothamanorukkiruvom da day.

Trascription: ഒരു വെഡിങ് ആയിരി അത യഥാർത്ഥത്തിൽ സംഭവിച്ചത് നിങ്ങൾക്ക് എനിക്ക് നിന്നും തില്ല അല്ലെ

Figure 1: Examples of misogyny detection in memes from Tamil and Malayalam datasets.

A sample memes from the Tamil and Malayalam can be seen in Figure 1 with their transcription and label.

4 Methodology

This section explores the pre-processing, feature extraction, feature fusion and training machine learning and deep learning models for classification of misogyny memes. The framework illustrating these methodologies is depicted in Figure 2. The design of the work is as follows: (i) Extract image and text features using transfer learning with the pre-trained ViT-Base-Patch16-224 and XLM-RoBERTa, respectively. (ii) Fuse the extracted text and image features for further processing. (iii) Train machine learning and deep learning models on the fused features to classify misogynistic content.

4.1 Pre-processing

The text extracted from the transcription of memes consists of noise, such as stopwords, digits, and punctuation which do not contribute to the classification. English stopwords available at NLTK library, Tamil and Malayalam stopwords available at github repositories[1] are used as references to remove English, Tamil and Malayalam stopwords, respectively. Vision Transformer preprocesses Tamil and Malayalam memes by resizing the images to a fixed input size of (224×224) pixels.

4.2 Feature extraction

The pre-processed text data is transformed into feature vectors using feature extraction techniques. This work utilizes XLM-RoBERTa to extracts features from text. The process begins with tokenizing the text into subword units using Byte-Pair Encoding, ensuring effective handling of rare words and diverse languages. Special tokens like <s> and </s> are added to structure the sequence, and each token is mapped to a high-dimensional embedding that incorporates token, positional, and segment information. These embeddings are passed through multiple transformer layers, where self-attention captures relationships between tokens, and feedforward networks refine the contextual representations. Hidden states generated at each layer provide rich, contextualized embeddings for each token, while the state of the <s> token often serves as a global representation for the input. The final output is a set of dense feature vectors for Tamil and Malayalam text.

The resized images are split into non-overlapping 16x16 pixel patches, flattened, and projected into a high-dimensional space using a linear layer, creating patch embeddings. These embeddings, combined with positional encodings, are passed through multiple self-attention layers of the Vision Transformer, where relationships between patches are learned. This process captures both local and global contextual features from the meme, enabling effective representation for classification task.

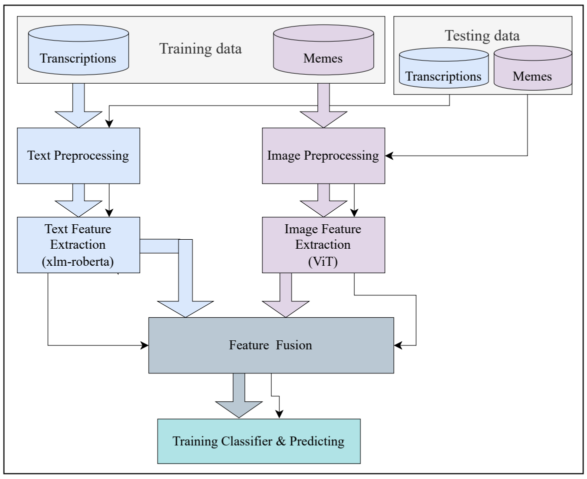


Figure 2: Block diagram of experimental work.

4.3 Feature Fusion & Classification

To fuse text and image features, we employ a simple concatenation-based feature fusion strategy. This fused representation allows the model to utilize both textual and visual modalities. To classify misogynistic memes, we utilized machine learning models and deep learning architectures. The machine learning models included K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB), which were trained using the fused feature embeddings derived from text and image modalities. For deep learning-based classification, we employed Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks to capture temporal dependencies in the fused features.

The Multimodal Classifier (MMC) integrates text and audio features for hate speech detection using a two-stage deep learning model. It consists of separate two-layer fully connected subnetworks for text and image features, each utilizing ReLU activation, batch normalization, and dropout (0.3) for regularization. The extracted modality-specific features are concatenated and processed through a three-layer fusion network, which learns inter-modal relationships before classification using softmax activation. The model is trained for 180 epochs end-to-end with binary cross-entropy loss, using the Adam optimizer (learning rate = 5e-5) and batch size = 32. The hidden dimension is 256, ensuring effective feature representation and robust multimodal learning.

[1] https://github.com/stopwords-iso/stopwords-iso

5 Results

Tables 2 and 3 show the Accuracy (Acc), Precision (Pre), Recall (Rec), and F_1-score (F_1) achieved by various classifiers on the Tamil and Malayalam datasets, respectively. The machine learning models considered include KNN, SVM, RF, and NB. For deep learning approaches, LSTM, GRU, and MMC were used.

Table 2: Performance of classifiers on fused embeddings for Tamil data

Classifier	Acc	Pre	Rec	F1
KNN	0.85	0.86	0.71	0.75
SVM	0.83	0.80	0.72	0.75
RF	0.83	0.88	0.66	0.69
NB	0.72	0.69	0.74	0.69
LSTM	0.83	0.77	0.75	0.76
GRU	0.81	0.75	0.71	0.73
MMC	0.81	0.75	0.79	**0.77**

Table 3: Performance of classifiers on fused embeddings for Malayalam data

Classifier	Acc	Pre	Rec	F1
KNN	0.85	0.90	0.81	0.83
SVM	0.85	0.85	0.85	0.85
RF	0.85	0.86	0.83	0.84
NB	0.81	0.81	0.81	0.81
LSTM	0.88	0.87	0.87	0.87
GRU	0.88	0.87	0.88	**0.88**
MMC	0.86	0.86	0.86	0.86

Among the models, the MMC outperformed others on the Tamil dataset, achieving an F_1-score of 77%, demonstrating its capability to process both textual and visual features effectively. On the Malayalam dataset, the GRU achieved the highest performance, with an F_1-score of 88%, indicating its superior ability to handle textual intricacies in this low-resource language whereas proposed MMC model achieved comparable performance with F_1-score of 0.86. Figures 3 and 4 illustrate the confusion matrices for the best-performing classifiers.

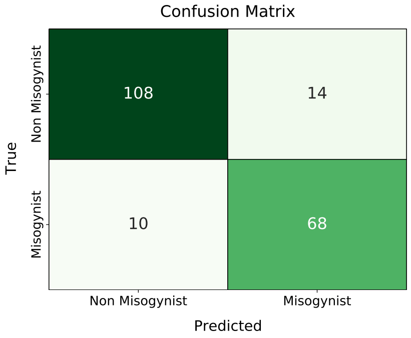


Figure 3: Confusion matrix of MMC for Tamil dataset.

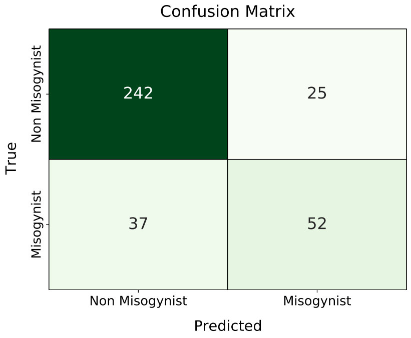


Figure 4: Confusion matrix of GRU for Malayalam dataset.

6 Conclusion

Misogyny refers to the discrimination against women, has been a persistent issue in society, affecting both offline and online spaces. There is increasing concern about its appearance in digital media. We apply transformer-based models to detect misogynistic content in a code-mixed Tamil and Malayalam. The results were encouraging: with an F_1 score of 77% for Tamil and an F_1 score of 88% for the Malayalam. This indicates that the model was successful in detecting misogynistic content, even where code-mixing was problematic and training data were limited. These results illustrate the capabilities of multimodal deep learning methods to understand and identify misogyny in low resource languages.

7 Limitations

Despite achieving strong performance in misogynistic meme classification for Tamil and Malayalam, our approach has certain limitations. First, the dataset size may not be sufficient to generalize across all variations of misogynistic memes, particularly those with subtle, implicit biases or sarcasm, which can be difficult for models to detect. The model also struggles with code-mixed content, low-resolution images, and heavily stylized fonts, which can distort both textual and visual understanding. While Malayalam models perform better than Tamil models, this discrepancy could stem from dataset composition, linguistic variations, or image-text alignment differences. In the future, a stronger system can be developed by addressing these limitations.

The code for the proposed framework is

available at: *https://github.com/Cshraddha153/Transformer-based-Multimodal-Framework-for-Misogyny-Meme-Detection.git*

References

Gita Beigi, Haiyong Ho, Kristina Lerman, and Benjamin C. Wallace. 2020. Hate speech detection in memes: A survey of approaches, datasets, and challenges. In *Proceedings of the 2020 IEEE/ACM International Conference on Web Search and Data Mining (WSDM)*, pages 322–330.

Maciej Buda, Alan Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. In *Proceedings of the International Conference on Computational Intelligence and Neuroscience*, pages 1–6.

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.

Alexey Dosovitskiy, Philipp Fischer, Jörg R Springenberg, Martin Riedmiller, and Marc Brockschmidt. 2014. Discriminative unsupervised feature learning with exemplar convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 764–772.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Vanessa Garcia, Yi Chang, Yifan Xu, and Enrique Alfonseca. 2021. Multimodal meme classification: Leveraging textual and visual features for robust meme detection. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, pages 3276–3280.

I. Gasparini, A. Singh, G. Vasan, A. Narayan, and A. Deshmukh. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. In *Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM 2022)*, pages 1267–1274, Chennai, India. IEEE.

Shaun H, Samyuktaa Sivakumar, Rohan R, 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, St. Julian's, Malta. Association for Computational Linguistics.

M. Hossain, M. Islam, M. Rahman, and S. Shahin. 2022. Memosen: A multimodal dataset for meme sentiment analysis in bengali. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 1651–1658, Marseille, France. European Language Resources Association.

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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Douwe Kiela, Mohamed Elhoseiny, Lin Zhang, Marco Baroni, and Geoffrey Hinton. 2019. Supervised multimodal hashing for scalable cross-modal retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1740–1751.

Abhinav Kumar, Pradeep Kumar Roy, and Jyoti Prakash Singh. 2021. A deep learning approach for identification of arabic misogyny from tweets. In *FIRE (Working Notes)*, pages 831–838.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

B Narasimha Rao and V Kalyani. 2022. A study on positive and negative effects of social media on society. *Journal of Science & Technology (JST)*, 7(10):46–54.

Giorgia Rizzi, Wei Zhang, and Martin Hall. 2023. Detecting misogyny in memes: A comparative study of unimodal and multimodal approaches. In *Proceedings of the International Conference on Multimodal Analysis*, pages 199–208.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2025. Misogynistic attitude detection in youtube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech Language*, 89:101682.

Ravi Singh, Aman Sharma, and Vikash Gupta. 2024. Misogyny identification in multimodal internet content (mimic). *Journal of Social Media Studies*, 12(4):345–368.

Amit Suryawanshi, Sachin Jadhav, and Venkatesh Rajendran. 2020. Hate speech detection in memes: A comparative analysis of image and text-based approaches. In *Proceedings of the 2020 International Conference on Computational Intelligence and Data Science (ICCIDS)*, pages 287–292.

Mathias Weber, Christina Viehmann, Marc Ziegele, and Christian Schemer. 2020. Online hate does not stay online–how implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in human behavior*, 104:106192.

Yu Zhang, Xue Xu, Hongzhi Yang, Zhenyu Yu, and Xiang Yu. 2020. An overview of deep learning-based approaches for multimodal hate speech detection. *Journal of Artificial Intelligence Research*, 69:545–576.