teamiic@DravidianLangTech 2025: Transformer-Based Multimodal Feature Fusion for Misogynistic Meme Detection in Low-Resource Dravidian Language

Harshita Sharma, **Simran**, **Vajratiya Vajrobol**, **Nitisha Aggarwal**,

Institute of Informatics and Communication, University of Delhi, Delhi, India

harshita.sharma@iic.ac.in, simran.2022@iic.ac.in, tiya101@south.du.ac.in, nitisha@south.du.ac.in

Abstract

Misogyny has become a pervasive issue in digital spaces. Misleading gender stereotypes are getting communicated through digital content. This content is majorly displayed as a text-and-image memes. With the growing prevalence of online content, it is essential to develop automated systems capable of detecting such harmful content to ensure safer online environments. This study focuses on the detection of misogynistic memes in two Dravidian languages, Tamil and Malayalam. The proposed model utilizes a pre-trained XLM-RoBERTa (XLM-R) model for text analysis and a Vision Transformer (ViT) for image feature extraction. A custom neural network classifier was trained on integrating the outputs of both modalities to form a unified representation. This model predicts whether the meme represents misogyny or not. This follows an early-fusion strategy since features of both modalities are combined before feeding into the classification model. This approach achieved promising results using a macro F1-score of 0.84066 on the Malayalam test dataset and 0.68830 on the Tamil test dataset. In addition, it is worth noting that this approach secured Rank 7 and 11 in Malayalam and Tamil classification respectively in the shared task of Misogyny Meme Detection (MMD). The findings demonstrate that the multimodal approach significantly enhances the accuracy of detecting misogynistic content compared to text-only or image-only models.

1 Introduction

Misogyny on social media has become an alarming concern in recent years. Unfortunately, digital platforms have become a breeding ground for misogynistic content. This issue is not merely confined to explicit hate speech but extends to more insidious forms like memes that trivialize or normalize sexism. A meme is a cultural idea, joke, trend, or piece of content (often in the form of an image, text, or video) that spreads rapidly through social media. Memes hold the power to shape societal perceptions and reinforce harmful gender stereotypes as these are often humorous, highly engagable, and shareable. Studies have shown that such content exacerbates existing inequalities and fosters a culture of misogyny (Jane, 2017; Banet-Weiser and Miltner, 2015). Addressing this issue is critical for promoting digital civility and safeguarding the rights and dignity of women in online spaces. The detection and mitigation of misogynistic content presents several unique challenges.

Over the years, significant efforts have been made to combat online misogyny. International conventions, policy frameworks, and platform-specific moderation mechanisms are designed. However, the effectiveness of these measures remains limited. Despite advancements in artificial intelligence (AI) and natural language processing (NLP), the detection of misogynistic memes remains an underexplored area. This study focuses on detecting misogynistic memes by addressing their textual and visual complexities. By utilizing a novel dataset provided by the organizers of this shared task this study is focused on extracting robust feature representations using XLM-R (Conneau) and ViT (Dosovitskiy, 2020) embeddings, paired with a carefully designed model architecture to capture nuanced patterns across both classes. The macro-average F1 score ensures balanced evaluation, mitigating class imbalance. The repository is available on GitHub[1]. The goal is to develop models capable of classifying misogynistic content and identifying contextual cues in such memes. Through this research, we seek to contribute to the growing body of knowledge in the field of hate speech detection, advancing the understanding of multimodal misogyny and paving the way for more effective content moderation strategies.

[1]This is link to the code for submission - Sharma (2025)

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 556–560

2 Related Work

The detection of misogynistic content in online platforms has been a growing area of research, particularly in the context of multimodal data such as memes. Early research on misogyny detection and hate speech primarily focused on textual data. (Waseem et al., 2017) highlights NLP-based hate speech detection. With the rise of visual communication, memes have become a prominent medium for spreading hateful content. Zhu et al. (2021) introduced a multimodal approach for detecting visual hate speech in memes, demonstrating that combining textual and image-based features enhances classification performance and Sai et al. (2022) examines different fusion strategies for integrating textual and visual cues showing late fusion approaches. Similarly, Mathew et al. (2021) examined hate speech detection in multilingual settings, highlighting the limitations of unimodal approaches.

Recent research has focused on detecting misogynistic memes in low-resource Dravidian languages, particularly Tamil and Malayalam. Ponnusamy et al. (2024) introduced the MDMD dataset for misogyny detection in Tamil and Malayalam memes, providing a valuable resource for understanding gender bias in these communities. Chakravarthi et al. (2024) organized this shared task, reporting the best macro F1 scores of 0.73 for Tamil and 0.87 for Malayalam. Earlier work by Ghanghor et al. (2021) addressed offensive language identification and troll meme classification in Dravidian languages, achieving weighted F1 scores of 0.75 for Tamil and 0.95 for Malayalam in offensive language detection.

The effectiveness of deep learning in misogyny detection has been well-documented. Johnson and Khoshgoftaar (2019) surveyed deep learning technique in imbalanced class settings. ViTs and CLIP Radford et al. (2021) have shown improvements in multimodal classification performance. Some text-only models (Annamoradnejad and Zoghi, 2024) achieved an F1 score of 0.76, while image-only models (Mathew et al., 2021) reached F1 score of 0.72, demonstrating the limitations of unimodal models. Multimodal models.

Studies such as Sharma et al. (2022) highlight the difficulty of detecting sarcasm and implicit hate speech in multimodal content. Furthermore, Davidson et al. (2017) and Chakravarthi (2022) emphasize the importance of cross-cultural sensitivity in hate speech detection, particularly in multilingual settings. Furthermore, the dataset's inherent bias towards specific cultural and linguistic contexts may limit the model's applicability to global audiences. Singh et al. (2023) employed XLM-RoBERTa to detect hate speech and its targets, demonstrating its effectiveness in multilingual contexts. The Social Media Sexist Content (SMSC) database (Buie and Croft, 2023) and addressing the bias in the dataset (Zhang et al., 2023) provides a valuable benchmark for future research, These studies show progress in multimodal misogyny detection but highlight the need for more research in multilingual, low-resource languages like Tamil and Malayalam.

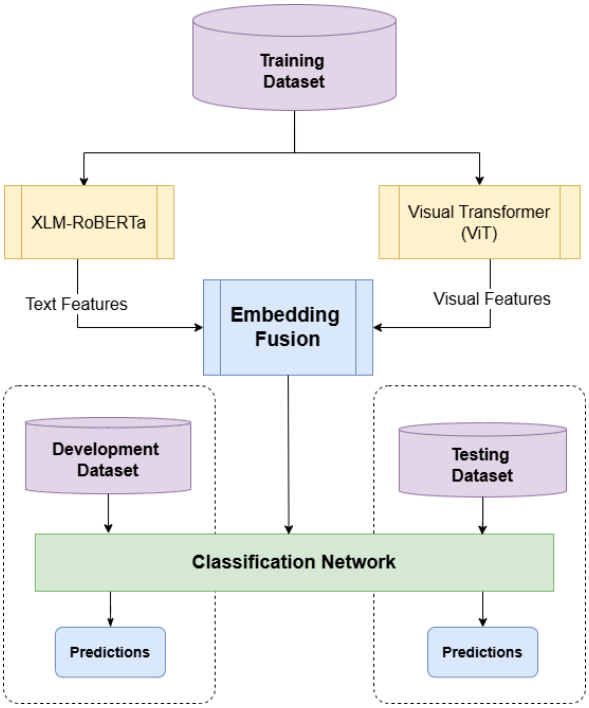

Figure 1: Process Flow of the proposed model for Misogyny Meme Detection.

3 Methodology

The proposed methodology employs a multimodal approach for misogyny detection in Tamil and Malayalam memes using a labeled dataset of images with transcriptions as shown in Figure 1.

3.1 Dataset and data pre-processing

The provided dataset is a unique collection of misogyny memes explicitly focusing on the Tamil and Malayalam languages. Figure 2 represented the transcriptions provided in the dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024, 2025) include both monolingual and bilingual con-

image_id	labels	transcriptions
1008	0	Sight Adichifying College Staff Expectation Reality nivasmc@fb
1176	1	RUKKu ioh IHL~NW irukkuzingafun irukku PHOTOURID Ungala Leggings Poda venamnu solala Short Tops potutu poda venam nu dha solrom nallavae ila asigama irukum Wear panunga ana Long tops potu wear panunga
76	0	Seven Screen Studio @7screenstudio Considering overflowing passes requests & safety constraints

Figure 2: Attribute of the provided Tamil Training Dataset

image_id	labels	transcriptions
888	0	ഈ ചാടി ഓടി നടക്കണ മനുഷ്യനാണോടാ നിങ്ങളിത്രേം കാലം ബെഞ്ചിലിട്ടത് നീയൊക്കെ പുഴുത്ത് ചാകും വേണ്ടത്. സന്തോഷിച്ചാട്ടെ സന്തോഷിച്ചാട്ടെ.
554	1	മലയാള സിനിമയുടെ ഭാവി വടറാണി ഇവൾ തന്നെ നല്ല കുഴിഞ്ഞ പൊക്കിൾ
556	1	ഒന്ന് പെറ്റത് ആണെങ്കിലും .. മുലയും വയറും ചാടിയതാണെങ്കിലും ചില മലഞ്ചരക്കുകളുടെ ഡിമാൻഡ് ഒരിക്കലും കുറയില്ല

Figure 3: Attribute of the provided Malayalam Training Dataset

tent. Dataset contains code-mixed text that combines Tamil and English. Similarly, Figure 3 combines Malayalam and English languageges. This mix of linguistic elements reflects real-world online discourse and adds complexity to the task.

The dataset was released in three divisions for each language: train, development, and test. Each subset is organized within a folder containing a set of images and an accompanying CSV file with the following structure: 1)image_id: A unique identifier assigned to each image in the folder. 2) labels: Binary annotations indicating whether the image represents misogyny or not. 3)transcriptions: The text transcription is extracted from the corresponding image. In the case of the test dataset, the labels were not provided.

During the initial data analysis, the class distributions in both the Tamil and Malayalam training datasets were examined. The Tamil dataset contains 851 non-misogynistic and 285 misogynistic samples. Malayalam dataset consists of 381 non-misogynistic and 259 misogynistic samples. Instead of applying data balancing techniques such as oversampling or undersampling, the decision was made to retain the original distribution to preserve the natural characteristics of the data. This approach mirrors real-world scenarios, where misogynistic content is generally less prevalent compared to non-misogynistic content (Buie and Croft, 2023). Specifically, the F1-score was used instead of accuracy, as it provides a more reliable measure of performance in cases where class distributions are not perfectly uniform (Johnson and Khoshgoftaar, 2019).

3.2 Preprocessing and Feature Extraction

XLM-R tokenizer was used to tokenised transcriptions of textual data. Each tokenized input was padded and truncated to ensure consistency. The embedding extraction of the text was done by using XLM-R, embeddings were extracted by averaging the last_hidden_state representations of tokens. Raw image files were resized, normalized using ViT. Embeddings were computed using the last_hidden_state of the image model to ensure robust representation of visual features.

3.3 Feature Fusion

The embeddings from text - XLM-R (768 dimensions) and image - ViT (768 dimensions) modalities were concatenated to create a unified representation. Since, here the features from text and images were combined into a single vector and no decision-making occurred at the individual modality level, it can be considered as early-fusion technique. It is a feature-level fusion. This fusion allowed for simultaneous utilization of textual and visual modalities, enhancing the model's ability to classify multimodal inputs effectively.

3.4 Model architecture and training

A custom neural network named MultimodalClassifier, was designed with the following component.
1) Input Layer: Accepts the 1536-dimensional concatenated feature vector.
2) Hidden Layer: A fully connected layer with 512 units and ReLU activation to capture non-linear interactions between the fused features.
3) Output Layer: A fully connected layer that outputs logits corresponding to the number of classes.
The fused embedding was passed through this fully connected neural network classifier. The training process incorporated some model optimizing strategies. CrossEntropyLoss was employed as the loss function. To effectively handle multi-class classification, softmax to the logits was applied. The Adam optimizer (learning rate of 0.001) was used to ensure efficient weight updates. To handle data efficiently, training samples were organized into batches of size 32 using PyTorch's DataLoader. Shuffling was applied to maintain randomness during batch sampling and enhance the model's generalization. The training process was conducted

over ten epochs. For each epoch, a forward pass was executed to compute predictions, followed by a backward pass to calculate gradients by minimizing the loss. The optimizer then updated the model weights. The model's predictions were stored to compute F1-Score along with the loss at the end of each epoch, ensuring a comprehensive evaluation of model performance.

3.5 Evaluation and Metrics

The models went through established evaluation metrics to ensure performance across all classes. In addition to accuracy, the F1-Score (macro) was calculated to provide a holistic view of the model's performance. F1-Score (macro) also gives accurate evaluation for the case of class imbalance. The final predictions of the test dataset in both the languages were evaluated against the macro f1 score metrics.

4 Results and Discussion

The proposed model in this study was evaluated on the provided dataset of misogynistic memes. The results indicate a significant scope of improvement in detecting misogynistic memes when combining both text and image features compared to unimodal approaches, Text-Only (Annamoradnejad and Zoghi, 2024) or Image-Only (Mathew et al., 2021) as it's shown in Sharma et al. (2022) which is also a multimodal approach.

Language	F1-Score	Precision	Recall	Accuracy
Tamil	0.6615	0.6580	0.6659	0.7324
Malayalam	0.7968	0.7964	0.8100	0.8000

Table 1: Performance metrics for Tamil and Malayalam development datasets.

The findings of the shared task indicate satisfactory performance, as reflected in the evaluation metrics for the development dataset. For Tamil and Malayalam, the evaluation metrics including the F1-scores are shown in Table: 1. This research aims to contribute to the expanding field of hate speech detection by enhancing the understanding of multimodal misogyny and informing more effective content moderation strategies. The final results on the test dataset yielded a macro F1-score of 0.84066 for Malayalam language and 0.6883 for Tamil language (see Tables: 3 and 2).

Team	Macro F1-score	Rank
Shraddha	0.70501	10
teamiic	**0.6883**	**11**
InnovationEngineer	0.68782	12

Table 2: F1- score for the Tamil Dataset

Team	Macro F1-score	Rank
CUET-NLP_MP	0.84118	6
teamiic	**0.84066**	**7**
byteSizedLLM	0.83912	8

Table 3: F1- score for the Malayalam Dataset

5 Conclusion

This study demonstrates the effectiveness of a multimodal approach in detecting misogynistic memes, It is highlighting the combining of textual and visual features enhances classification accuracy. By employing XLM-R for text analysis and ViT embeddings for image feature extraction, the model successfully identified nuanced representations of misogynistic content in Tamil and Malayalam memes. The feature fusion strategy significantly contributed to strong classification performance, achieving macro F1-scores of 0.84066 for Malayalam and 0.6883 for Tamil and securing Rank 7 and Rank 11 respectively. These results validate the ability of this approach to tackle the challenges of detecting misogyny in multimodal content. However, the limited size of the labeled data led to the use of pre-trained models like XLM-R and ViT. The reliance of these models on large, diverse datasets for pre-training may hinder their ability to capture highly nuanced, language-specific, or evolving forms of misogyny in the targeted communities. This dependency on pre-trained models may also affect the generalizability of the results to broader online contexts. To overcome these challenges in the future, it may be possible to get a dataset that includes more culturally diverse and real-world samples for enhancing the model's generalizability across different linguistic and societal contexts. As harmful discourse continues to evolve, ongoing advancements in multimodal AI models will play a pivotal role in creating safer, more inclusive online environments.

References

Issa Annamoradnejad and Gohar Zoghi. 2024. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249:123685.

Sarah Banet-Weiser and Kate M. Miltner. 2015. #Mas-

culinitySoFragile: culture, structure, and networked misogyny. *Feminist Media Studies*, 16(1):171–174.

Hannah Buie and Alyssa Croft. 2023. The social media sexist content (smsc) database: A database of content and comments for research use. *Collabra: Psychology*, 9(1):71341.

Bharathi Raja Chakravarthi. 2022. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

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.

A Conneau. Unsupervised cross-lingual representation learning at scale.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Emma Jane. 2017. *Misogyny Online: A short (and brutish) history*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: a benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science*, 3(2):122.

Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Hyunsung Kim, and Raj Sharma. 2022. Sarcasm detection over social media platforms using hybrid auto-encoder-based model. *Electronics*, 11(18):2844.

H. Sharma. 2025. teamiic@dravidianlangtech 2025.

Karanpreet Singh, Vajratiya Vajrobol, and Nitisha Aggarwal. 2023. Iic_team@ multimodal hate speech event detection 2023: Detection of hate speech and targets using xlm-roberta-base. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 136–143.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating biases in hate speech detection from a causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625.

Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2021. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221, Online. Association for Computational Linguistics.