HerWILL@DravidianLangTech 2025: Ensemble Approach for Misogyny Detection in Memes Using Pre-trained Text and Vision Transformers

Neelima Monjusha Preeti[1,2], Trina Chakraborty [1,3], Noor Mairukh Khan Arnob[1,4], Saiyara Mahmud[1,4], Azmine Toushik Wasi[1,3]

[1]STEM Team, HerWILL Inc., [2]Jahangirnagar University,
[3]Shahjalal University of Science and Technology, [4]University of Asia Pacific

Correspondence: arnob@uap-bd.edu

Abstract

Misogynistic memes on social media perpetuate gender stereotypes, contribute to harassment, and suppress feminist activism. However, most existing misogyny detection models focus on high-resource languages, leaving a gap in low-resource settings. This work addresses that gap by focusing on misogynistic memes in Tamil and Malayalam, two Dravidian languages with limited resources. We combine computer vision and natural language processing for multi-modal detection, using CLIP embeddings for the vision component and BERT models trained on code-mixed hate speech datasets for the text component. Our results show that this integrated approach effectively captures the unique characteristics of misogynistic memes in these languages, achieving competitive performance with a Macro F1 Score of 0.7800 for the Tamil test set and 0.8748 for the Malayalam test set. These findings highlight the potential of multimodal models and the adaptation of pre-trained models to specific linguistic and cultural contexts, advancing misogyny detection in low-resource settings. Code available at https://github.com/HerWILL-Inc/NAACL-2025

1 Introduction

Misogynistic memes on social media contribute to harmful gender stereotypes and perpetuate inequalities, creating hostile online environments (Chen et al., 2024). These memes amplify sexism, often resulting in online harassment and gender-based cyberbullying (Cai, 2024; Wang and Elfira, 2024). The anonymity and humor inherent in memes offer a unique space to critique societal norms without direct confrontation, providing a platform for both harmful content and progressive movements. While misogynistic memes reinforce oppressive stereotypes, they simultaneously highlight the need for greater awareness, policy intervention, and cultural change. Feminist movements have leveraged

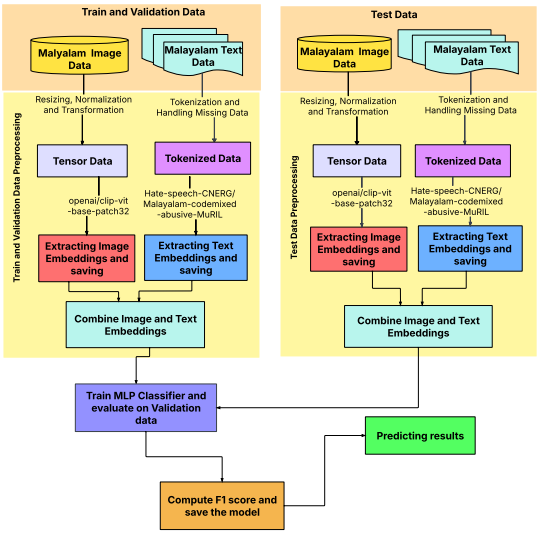

Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

these platforms to amplify their voices, demonstrating that social media is not only a battleground for gender dynamics but also a powerful space for feminist discourse (Zhu, 2024; Chen et al., 2024). This dual role of memes—simultaneously a tool of oppression and resistance—underscores the complexity of social media's impact on societal norms (Khosravi-Ooryad, 2024).

Misogyny detection is rapidly advancing through the integration of natural language processing (NLP) and computer vision (CV). In NLP, techniques such as classification, severity scoring, and text rewriting help identify harmful language, assess its intensity, and promote respectful discourse. Automated systems on social media platforms effectively detect and flag misogynistic messages, while multilingual models expand detection across languages and cultural contexts (Sheppard et al., 2024; Guzman Cabrera et al., 2024). In CV, models like CLIP (Chen and Chou, 2022) integrate visual and linguistic features to improve detection

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 433–438

accuracy in memes, which combine both text and images. Innovations in image sentiment analysis and graph convolutional networks further enhance the ability to identify misogynistic content. However, there is a significant gap in models for low-resource settings, where the need is greater due to limited datasets and underrepresented languages. Most existing models focus on high-resource languages, leaving a void in addressing gender-based online abuse in these contexts. Although the shared task on Multitask Meme Classification at LT-EDI@EACL 2024 (Chakravarthi et al., 2024) attempted to alleviate this gap, it left room for further improvements.

In this work, we tackled the multi-modal misogyny meme detection task at the Third Shared Task of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025 (Chakravarthi et al., 2025), focusing on Tamil and Malayalam memes. Our approach integrated vision and text modalities to effectively analyze memes, which often blend visual and linguistic elements to convey implicit and explicit misogynistic content. For the vision component, we leveraged CLIP embeddings, a model pre-trained on diverse internet data, to capture the contextual and thematic cues of misogynistic memes. Given CLIP's exposure to web-based imagery, we hypothesized that it would be well-suited for identifying visual patterns associated with misogyny. For the text component, we utilized language-specific BERT models trained on code-mixed hate speech datasets tailored for Tamil and Malayalam. Recognizing that misogynistic memes frequently contain elements of hate speech expressed in code-mixed language, we aimed to capture the linguistic nuances unique to these languages. We combined the embeddings and fine-tuned an MLP classifier, achieving competitive results for both Tamil and Malayalam. Our findings highlight the effectiveness of integrating vision-based CLIP embeddings with language-specific text models for misogyny detection in low-resource languages. This work underscores the importance of multimodal approaches and adapting pre-trained models to specific linguistic and cultural contexts. The implications of this study extend beyond hate speech detection, as it demonstrates the potential for cross-modal learning in tackling social media toxicity, reinforcing the need for AI-driven interventions in promoting safer digital spaces.

2 Problem Description

In this shared task, we were assigned to classify whether a given meme is misogynistic or not. The recently developed MDMD (Misogyny Detection Meme Dataset) dataset (Ponnusamy et al., 2024) was provided for this task. The dataset contains two portions: Malayalam and Tamil. In the Malayalam section, there are 640 memes on the training set, 160 memes on the dev set, and 200 memes on the test set. There are 1136, 284, and 356 memes on the train, dev, and test set of the Tamil part of the dataset. Each meme is recorded as an image file accompanied by textual transcriptions. The shared task was divided into two sub-tasks: Malayalam and Tamil. The training and development set was provided with labels during competition. We submitted our predictions on the test set for the Malayalam sub-task. Solutions were evaluated using macro average F1-score.

3 System Description

Data Pre-processing. In order to guarantee compatibility with pre-trained models and to enable proper embedding extraction, the preprocessing stage involves collecting both textual and visual data. The classification model then uses these embeddings as inputs; as outlined in Figure 1. The procedure includes using the corresponding pre-trained models to handle text and image input separately.

For handling the image data we used the OpenAI CLIP model (`openai/clip-vit-base-patch32`) (Radford et al., 2021) for image embedding extraction. First, Images are resized to 224x224 pixels and normalized into the standard form. Then preprocessed images are transformed into tensors to ensure compatibility with the image encoder model. Finally, extracted embeddings are stored as dictionary with the key-value pair of image-id and embedding tensors.

For handing text data, we employed the pre-trained language model `Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL` (Das et al., 2022). The transcriptions are tokenized using the corresponding tokenizer. For uniformity, inputs are padded or truncated to a maximum length of 128 tokens. The embeddings are then extracted and stored similarly to image embeddings indexed by image-id. Due to saving image and text embeddings on-disk, we did not have to extract embeddings(which is a time consuming process)

Table 1: Performance in the Validation Set Across Different Models for the Malayalam Dataset

Language Model	Vision Model	F1 Score	Accuracy
ai4bharat/IndicBERTv2-MLM-only	openai/clip-vit-base-patch32	0.8753	0.8812
PosteriorAI/dravida_llama2_7b	zer0int/CLIP-GmP-ViT-L-14	0.8896	0.8938
./malayalam-codemixed-abusive-MuRIL	openai/clip-vit-base-patch32	**0.8940**	**0.9000**

Table 2: Performance in the Test Set Across Different Models for the Tamil Dataset

Language Model	Vision Model	F1 Score	Accuracy
ai4bharat/IndicBERTv2-MLM-only	openai/clip-vit-base-patch32	0.7643	**0.8455**
PosteriorAI/dravida_llama2_7b	zer0int/CLIP-GmP-ViT-L-14	**0.7800**	0.8427
./tamil-codemixed-abusive-MuRIL	openai/clip-vit-base-patch32	0.7575	0.8174

every time we trained our model.

Models. Since misogyny is a form of hate speech, we selected language models pre-trained on offensive corpora to enhance detection performance. For Malayalam, we used Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL (Das et al., 2022), specifically trained to identify abusive code-mixed Malayalam text. Additionally, we experimented with ai4bharat/IndicBERTv2MLM-only (Doddapaneni et al., 2023), a model trained on 23 Indic languages, including Tamil and Malayalam, to evaluate its generalization capability. To leverage embeddings from a modern LLM, we tested PosteriorAI/dravida_llama2_7b (PosteriorAI, 2024), which has been trained on Kannada, Telugu, Malayalam, and Tamil corpora. For Tamil text encoding, we used Hate-speech-CNERG/tamil-codemixed-abusive-MuRIL (Das et al., 2022), a model specifically designed for offensive Tamil text detection. Given that the dataset consists of memes (internet-based multimodal data), we employed openai/clip-vit-base-patch32 (Radford et al., 2021) as an image encoder, leveraging its training on diverse internet images. To explore potential improvements with a larger model, we also experimented with zer0int/CLIP-GmP-ViT-L-14 (zer0int, 2023). For classification, we trained a lightweight Multi-Layer Perceptron (MLP), ensuring time- and memory-efficient classification while effectively integrating the multimodal embeddings.

Implementation Details. We designed our MLP model for classifying Malayalam memes with an efficient and structured architecture. The model starts with a Linear layer of size $[1280 \times 1024]$, followed by a LeakyReLU activation and Dropout ($p = 0.3$) to mitigate overfitting. Next, we included another Linear layer of size $[1024 \times 512]$, again paired with LeakyReLU and Dropout ($p = 0.3$). Finally, a Linear layer of size $[512 \times 1]$ is followed by a Sigmoid activation function to generate the final classification probability. We set the batch size to 32 and found that training for 10 epochs was sufficient for convergence. We used a learning rate of 0.0005, which provided a good balance between stability and learning speed. Our MLP model had 1,410,323 trainable parameters, making it both lightweight and effective for the task.

4 Experimental Findings

4.1 Malayalam Results

The performance metrics of various models trained on the Malayalam portion of the MDMD dataset are presented in Table 1. The IndicBERT model underperformed due to its limited exposure to offensive and misogynistic language in Malayalam, making it less effective for this task. Similarly, despite its large size (7 billion parameters), the dravida_llama2 model failed to achieve top results, likely because its pre-training corpus lacked sufficient misogynistic text. Our hypothesis that a hate-aware language model would be more effective is strongly supported by the results in Table 1, where the malayalam-codemixed-abusive-MuRIL model achieved the highest F1 score of 0.8940 on the validation set.

Based on this strong validation performance, we submitted our final solution using a combination of predictions from the malayalam-codemixed-abusive-MuRIL (110 million parameters) and the vision-based clip-vit-base-patch32 model. This approach secured **2nd place** in the shared task, achieving an

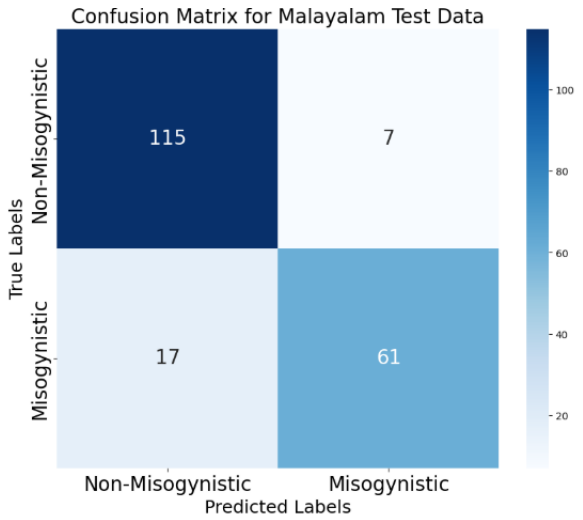

Figure 2: Performance of our proposed model on the malayalam test data

impressive F1 score of 0.8748 on the test set. The confusion matrix in Figure 2 further illustrates the model's effectiveness, correctly classifying a total of 176 out of 200 memes. These results highlight the importance of task-specific pretraining and multimodal integration for improving misogyny detection in low-resource languages.

4.2 Tamil Results

As shown in Table 2, we performed experiments on the released test set with labels to find F1 score and accuracy on the test set. It is clear that the `dravida_llama2` achieves the best F1 score in combination with the large `CLIP-GmP-ViT-L-14` model. The `IndicBERTv2` model achieved the best accuracy of 0.8455 despite being a lightweight model (278 Million Parameters) compared to `dravida_llama2` (7 Billion Parameters). Although the `tamil-codemixed-abusive-MuRIL` model was trained on offensive text, it performed relatively poorly on the test set.

5 Discussion

Our experimental findings highlight several key challenges and insights in misogynistic meme detection for Tamil and Malayalam. One of the primary challenges was achieving robust performance on the Tamil subset of the MDMD dataset. Unlike high-resource languages, where models benefit from extensive labeled datasets, Tamil's low-resource nature limits the effectiveness of pretrained models, leading to underfitting and weaker generalization. Furthermore, our results indicate that increasing model size does not necessarily lead to better performance. The 7-billion-parameter LLaMA2 model, despite its scale, did not outperform smaller transformer-based models fine-tuned on task-specific data. This reinforces the No Free Lunch Theorem (Wolpert, 1996), suggesting that model selection should be guided by domain relevance rather than sheer size. In this task, models explicitly trained on code-mixed and abusive language data demonstrated superior performance over general-purpose large language models. Our findings highlight the value of integrating vision and text for misogyny detection in memes. CLIP embeddings captured contextual cues from images, while fine-tuned BERT models processed code-mixed text. This multimodal approach effectively addressed both implicit and explicit misogynistic content.

Overall, this study highlights the necessity of curating high-quality, domain-specific datasets for low-resource languages and refining model architectures to suit the nuances of code-mixed social media text. Future research should explore adaptive pretraining strategies, knowledge distillation, and cross-lingual transfer learning to enhance performance. Expanding the dataset with more diverse examples and including user interaction patterns could further improve the robustness of misogyny detection systems in Tamil and Malayalam.

6 Conclusion

In this study, we evaluated various models for detecting misogynistic content in Malayalam and Tamil memes. Our results highlight the effectiveness of hate-aware language models, with `malayalam-codemixed-abusive-MuRIL` achieving the highest performance in Malayalam, securing second place in the shared task. This underscores the importance of incorporating offensive text data for low-resource languages. For Tamil, the `dravida_llama2` model combined with `CLIP-GmP-ViT-L-14` yielded the best F1 score, demonstrating the advantages of domain-adapted models. However, models like `IndicBERT` and `tamil-codemixed-abusive-MuRIL` showed mixed results, emphasizing that model size alone is insufficient—training data quality and architecture must be balanced. Our findings contribute to improving misogyny detection in marginalized language communities and lay the groundwork for future advancements in low-resource NLP.

Limitations

One key limitation identified in this study is the absence of diverse offensive and misogynistic words in the available corpora of the Tamil and Malayalam languages. Due to this shortcoming, the models trained on these incomplete corpora perform imperfectly in such an important task as misogynistic meme classification. The CLIP image encoder used in this study was trained on internet data, which is skewed towards developed nations (Radford et al., 2021). Therefore, CLIP lacked the cultural knowledge contained in images of Tamil and Malayalam memes. An image encoder trained on images relevant to Tamil and Malayalam contexts can be developed in the future to achieve better accuracy in misogynistic meme detection. The small dataset size (only 2,776 memes) affects model generalization in this task. Other than curating more data, data augmentation using generative models can be a promising direction to improve the results.

Broader Impact Statement

The findings of this study provide valuable insights for developing systems to detect misogyny in online spaces. Specifically, building a system for detecting misogynistic memes in low-resource languages like Tamil and Malayalam could help reduce the prevalence of misogyny in online communities, particularly those belonging to marginalized groups. However, an important ethical consideration in this area is ensuring the privacy and reputation of individuals depicted in the memes. To mitigate potential harm, we recommend that the research dataset not be publicly released, as doing so could inadvertently perpetuate misogyny. The MDMD dataset should be strictly used for research purposes, with appropriate safeguards in place to protect the privacy and well-being of all stakeholders involved.

References

Ziyan Cai. 2024. The characteristics and causes of the phenomenon of "misogyny" in contemporary chinese online social platforms: Taking weibo and red as examples. *Communications in Humanities Research*, 26(1):113–122.

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.

Lei Chen and Hou Wei Chou. 2022. Rit boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from clip model and data-centric ai principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1864–1871. ACM.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM conference on hypertext and social media*, pages 32–42.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Rafael Guzman Cabrera, Jose Carmen Morales Castro, Angelica Hernandez Rayas, Jose Ruiz Pinales, and Jose Merced Lozano Garcia. 2024. Automatic detection of misogyny on x using artificial intelligence. *DYNA*, 99(6):562–562.

Sama Khosravi-Ooryad. 2024. Memeing back at misogyny: emerging meme-feminism, visual tactics, and aesthetic world-building on iranian social media. *Feminist Media Studies*, 24(5):984–1003.

Rahul Ponnusamy, Kathiravan Pannerselvam, R Saranya, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, S Bhuvaneswari, Anshid Ka, 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.

PosteriorAI. 2024. Dravida llama 2 7b. Accessed: 2025-01-29.

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.

Brooklyn Sheppard, Anna Richter, Allison Cohen, Elizabeth Smith, Tamara Kneese, Carolyne Pelletier, Ioana Baldini, and Yue Dong. 2024. Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 427–452, Bangkok, Thailand. Association for Computational Linguistics.

Ying Wang and Mina Elfira. 2024. *International Review of Humanities Studies*, 9(1).

David H. Wolpert. 1996. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

zer0int. 2023. Clip-gmp-vit-l-14. `https://huggingface.co/zer0int/CLIP-GmP-ViT-L-14`. Accessed: 2025-01-29.

Xuanxuan Zhu. 2024. Feminism on social media: Generalized misogyny. *SHS Web of Conferences*, 199:02012.