Dll5143@DravidianLangTech 2025: Majority Voting-Based Framework for Misogyny Meme Detection in Tamil and Malayalam

Sarbajeet Pattanaik
IIIT Allahabad
Prayagraj, 211015, India
mcl2023008@iiita.ac.in

Ashok Yadav
IIIT Allahabad
Prayagraj, 211015, India
rsi2021002@iiita.ac.in

Vrijendra Singh
IIIT Allahabad
Prayagraj, 211015, India
vrij@iiita.ac.in

Abstract

Misogyny memes pose a significant challenge on social networks, particularly in Dravidian-scripted languages, where subtle expressions can propagate harmful narratives against women. This paper presents our approach for the "Shared Task on Misogyny Meme Detection," organized as part of DravidianLangTech@NAACL 2025, focusing on misogyny meme detection in Tamil and Malayalam. To tackle this problem, we proposed a multi-model framework that integrates three distinct models: M1 (ResNet-50 + google/muril-large-cased), M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert), and M3 (ResNet-50 + ai4bharat/indic-bert). The final classification is determined using a majority voting mechanism, ensuring robustness by leveraging the complementary strengths of these models. This approach enhances classification performance by reducing biases and improving generalization. Our model achieved an F1 score of 0.77 for Tamil, significantly improving misogyny detection in the language. For Malayalam, the framework achieved an F1 score of 0.84, demonstrating strong performance. Overall, our method ranked 5th in Tamil and 4th in Malayalam, highlighting its competitive effectiveness in misogyny meme detection.

1 Introduction

The rapid proliferation of social media has revolutionized communication, enabling individuals to share information, ideas, and opinions instantly (Yadav and Singh, 2025) (Yadav and Singh, 2024). However, this unprecedented connectivity has also led to a surge in harmful online content, including hate speech, trolling, and misogyny, which disproportionately targets women and other marginalized groups (Lin et al., 2024). Among the diverse forms of online expression, memes have emerged as a powerful yet controversial medium. Detecting misogyny in memes is a complex task that demands a nuanced understanding of multimodal data, as memes typically combine text and images to convey meaning (Suryawanshi et al., 2020). While misogyny meme detection in major languages like English has seen significant advancements, there is a pressing need to extend this effort to languages written in Dravidian languages like Tamil and Malayalam (Shaun et al., 2024). The introduction of datasets such as MDMD for Tamil and Malayalam memes (Ponnusamy et al., 2024) and benchmark data sets for the detection of misogynistic content (Gasparini et al., 2022) highlighted the importance of annotated resources. Key initiatives included the development of frameworks like MISTRA for misogynous meme classification (Jindal et al., 2024),and shared tasks such as (Chakravarthi et al., 2024). To address these challenges, the Shared Task on Misogyny Meme Detection is organized as part of DravidianLangTech@NAACL 2025. This task challenges participants to develop multimodal models for analyzing textual and visual elements of social media memes. We secured 5th in Tamil and 4th in Malayalam.

The paper is organized as follows: Section 2 covers task statistics and the dataset. Section 3 explains our methodology. Section 4 details the experimental setup and evaluation metrics, with Section 4.1 comparing model performance. Section 5 concludes our findings, and Section 6 discusses limitations and future improvements in misogyny meme detection for Tamil and Malayalam. Our implementation is available on GitHub. [1]

2 Task and Dataset Description

The Shared Task on Misogyny Meme Detection is organized as part of DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2025). This task aims to

[1] https://github.com/Deeplearninglabiiita/DravidianLangTech-.git

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 278–286
May 3, 2025 ©2025 Association for Computational Linguistics

advance multimodal machine learning techniques to identify misogyny in memes. Participants are tasked with designing innovative systems capable of comprehensively interpreting both the textual and visual elements of memes. By combining these modalities, the systems must accurately classify memes as either misogynistic or non-misogynistic. The unique aspect of this task lies in its focus on two Dravidian languages, Tamil and Malayalam, which adds a layer of complexity due to linguistic diversity, cultural nuances, and limited resources in these languages.

The datasets for this shared task were drawn from (Ponnusamy et al., 2024). Table 1 provides statistics on the dataset used for the shared task in Tamil and Malayalam..

Table 1: The statistics of the used dataset in DravidianLangTech@NAACL 2025

Category	Tamil		Malayalam	
	Misogyny	Non-misogyny	Misogyny	Non-misogyny
Train	285	851	259	381
Val	74	210	63	97
Test	89	267	78	122
Total	448	1328	400	600

3 Proposed Framework

We have proposed a framework that utilizes three distinct models: M1 (ResNet-50 + google/muril-large-cased), M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert), and M3 (ResNet-50 + ai4bharat/indic-bert). Subsequently, we employed a majority voting mechanism for the final classification.

3.1 M1 (ResNet-50 + google/muril-large-cased)

In this study, we proposed a multimodal architecture to classify memes as **misogyny** or **non-misogyny** by leveraging both textual and visual information. The text modality is processed using the pre-trained MuRIL encoder. Given an input textual content T, it is first tokenized using MuRIL's tokenizer and then passed through the MuRIL encoder to extract contextualized embeddings:

$$\mathbf{T}_{\text{emb}} = \text{MuRIL}(T), \tag{1}$$

where $\mathbf{T}_{\text{emb}} \in \mathbb{R}^{n \times d}$, n is the number of tokens after padding or truncation, and d is the embedding dimension (768 for MuRIL). The embedding corresponding to the `[CLS]` token, which summarizes the entire text, is extracted and projected into a lower-dimensional feature space using a linear layer:

$$\mathbf{T}_{\text{feat}} = \text{ReLU}(\mathbf{W}_T \cdot \mathbf{T}_{\text{emb}}^{[CLS]} + \mathbf{b}_T), \tag{2}$$

where $\mathbf{T}_{\text{feat}} \in \mathbb{R}^{256}$ is the projected text feature, $\mathbf{W}_T \in \mathbb{R}^{256 \times 768}$ is a learnable projection matrix, and $\mathbf{b}_T \in \mathbb{R}^{256}$ is the bias term.

The visual modality is processed using ResNet-50, which is pre-trained on ImageNet. Each input image I is resized and normalized before passing through the ResNet-50 network. Features are extracted from the penultimate layer by removing the classification head:

$$\mathbf{I}_{\text{emb}} = \text{ResNet-50}(I), \tag{3}$$

where $\mathbf{I}_{\text{emb}} \in \mathbb{R}^{2048}$ represents the high-dimensional image embedding. To align the dimensionality of image and text features, a linear projection is applied to reduce the dimensionality:

$$\mathbf{I}_{\text{feat}} = \text{ReLU}(\mathbf{W}_I \cdot \mathbf{I}_{\text{emb}} + \mathbf{b}_I), \tag{4}$$

where $\mathbf{I}_{\text{feat}} \in \mathbb{R}^{256}$, $\mathbf{W}_I \in \mathbb{R}^{256 \times 2048}$, and $\mathbf{b}_I \in \mathbb{R}^{256}$ are learnable parameters. The features from both modalities are concatenated to form a joint multimodal representation:

$$\mathbf{F} = [\mathbf{T}_{\text{feat}}; \mathbf{I}_{\text{feat}}], \tag{5}$$

where $\mathbf{F} \in \mathbb{R}^{512}$ is the fused feature vector.

The fused feature vector $\mathbf{F}$ is passed through a classification head, which consists of a dropout layer followed by a dense layer for binary classification:

$$y = \sigma(\mathbf{W}_C \cdot \mathbf{F} + b_C), \tag{6}$$

where $\mathbf{W}_C \in \mathbb{R}^{1 \times 512}$, $b_C \in \mathbb{R}$, σ is the sigmoid activation, and $y \in [0, 1]$ represents the probability of the meme being misogynistic. To classify the meme, a threshold $\tau = 0.5$ is applied:

$$\hat{y} = \begin{cases} 1, & \text{if } y \geq \tau \text{ (Misogyny)}, \\ 0, & \text{otherwise (Non-Misogyny)}. \end{cases}$$

3.2 M2 (openai/clip-vit-base-patch32 + ai4bharat/indic-bert),

This sub-section presents a multimodal approach to classifying memes as misogynistic or non-misogynistic using text and image features. The model utilizes CLIP for image encoding and IndicBERT for textual feature extraction, followed by a classification layer.

Given an input text T, the text encoder (IndicBERT) extracts a feature representation:

$$\mathbf{h}_T = f_{\text{IndicBERT}}(T), \tag{7}$$

where $\mathbf{h}_T \in \mathbb{R}^{d_T}$ is the output embedding of the [CLS] token.

Given an image I, the CLIP model extracts a feature representation:

$$\mathbf{h}_I = f_{\text{CLIP}}(I), \tag{8}$$

where $\mathbf{h}_I \in \mathbb{R}^{d_I}$ represents the image embedding from the CLIP model.

The extracted text and image features are concatenated to form a fused representation:

$$\mathbf{h}_{\text{fused}} = [\mathbf{h}_T; \mathbf{h}_I], \tag{9}$$

where $[;]$ denotes concatenation, and $\mathbf{h}_{\text{fused}} \in \mathbb{R}^{d_T+d_I}$.

A fully connected layer maps the fused representation to a binary classification output:

$$\hat{y} = \sigma(\mathbf{W}\mathbf{h}_{\text{fused}} + \mathbf{b}), \tag{10}$$

where σ is the sigmoid activation function, $\mathbf{W}$ is the weight matrix, and $\mathbf{b}$ is the bias term. To classify the meme, a threshold $\tau = 0.5$ is applied and predicted label is given by:

$$y = \begin{cases} 1, & \hat{y} > 0.5, \\ 0, & \hat{y} \leq 0.5. \end{cases} \tag{11}$$

Binary cross-entropy loss is used to optimize the model:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{12}$$

3.3 M3 (ResNet-50 + ai4bharat/indic-bert)

In this model (M3), our approach leverages multimodal learning by combining textual and visual features using indic-bert and ResNet-50 respectively.

Each meme consists of an image and an associated text transcription. Given a dataset of N memes, each sample can be represented as:

$$M_i = (I_i, T_i, y_i) \tag{13}$$

where:

- $I_i \in \mathbb{R}^{H \times W \times C}$ is the meme image with height H, width W, and C color channels.
- T_i is the textual transcription.
- $y_i \in \{0, 1\}$ is the binary label (0: non-misogynistic, 1: misogynistic).

We use IndicBERT to encode text representations. Given a transcription T_i, the tokenized input is:

$$\mathbf{x}_i = \text{Tokenizer}(T_i) \tag{14}$$

The text encoder outputs contextual embeddings:

$$\mathbf{h}_i = f_{\text{BERT}}(\mathbf{x}_i) \tag{15}$$

We extract the CLS token representation as the text feature:

$$\mathbf{t}_i = \mathbf{h}_i^{[CLS]} \in \mathbb{R}^{d_t} \tag{16}$$

where d_t is the hidden dimension of the BERT model.

We use a pretrained ResNet-50 to extract image features. Given an input image I_i, the image embedding is obtained as:

$$\mathbf{v}_i = f_{\text{ResNet}}(I_i) \in \mathbb{R}^{d_v} \tag{17}$$

where d_v is the feature dimension of the ResNet-50 output.

The extracted textual and visual features are projected into a common latent space:

$$\mathbf{t}'_i = W_t \mathbf{t}_i + b_t \in \mathbb{R}^{d_f} \tag{18}$$

$$\mathbf{v}'_i = W_v \mathbf{v}_i + b_v \in \mathbb{R}^{d_f} \tag{19}$$

where d_f is the fusion feature dimension, and W_t, W_v are learnable projection matrices.

The final multimodal feature vector is obtained by concatenation:

$$\mathbf{z}_i = [\mathbf{t}'_i; \mathbf{v}'_i] \in \mathbb{R}^{2d_f} \tag{20}$$

A binary classifier maps the fused representation to a scalar output:

$$\hat{y}_i = \sigma(W_o \mathbf{z}_i + b_o) \tag{21}$$

where W_o and b_o are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid activation. To classify the meme, a threshold $\tau = 0.5$ is applied:

$$\hat{y} = \begin{cases} 1, & \text{if } y \geq \tau \text{ (Misogyny)}, \\ 0, & \text{otherwise (Non-Misogyny)}. \end{cases}$$

We optimize the model using Binary Cross-Entropy (BCE) loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big] \tag{22}$$

The model is trained using AdamW optimizer with learning rate η and weight decay λ:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L} + \lambda\theta \tag{23}$$

.

3.4 Majority Voting

The proposed algorithm determines the final classification label for misogyny detection using predictions from three different models (M1, M2, and M3). It employs a majority voting mechanism, where the final label is assigned based on the agreement of at least two out of three used models using Algorithm 1 (Appendix 7).

4 Experimental Settings and Evaluations Metrics

We implemented our model using PyTorch and Hugging Face Transformers, with training conducted on an Nvidia A30 GPU. The model was trained for 12 epochs using an AdamW optimizer with a learning rate of 1e-5 and batch size of 8. For reproducibility, we set the manual seed to 30 and used a dropout rate of 0.3 to prevent overfitting.

4.1 Results and Analysis

Our proposed framework achieved an F1 score of 0.77 for misogyny meme detection in Tamil. For Malayalam, the framework attained an F1 score of 0.84. Tables 2 and 3 present our system's performance compared to other participating systems on the test dataset.

For misogyny meme detection in Tamil languages [2], our system achieved competitive results, ranking 5th among all participants. The best-performing system achieved an F1 score of 0.83682, demonstrating the challenging nature of misogyny meme detection in Tamil languages. In misogyny meme detection in Malayalam languages [3], our system ranked 4th among all participants. While the top system achieved an F1 score of 0.87631, the relatively small performance gap (0.03631) between the first and fourth positions suggests the complexity of the task and the effectiveness of various approaches.

Table 2: Results comparison of top systems for misogyny meme detection in Tamil languages

System	F1	Rank
DLRG_RR	0.83682	1
CUET-NLP_Big_O	0.81716	2
byteSizedLLM	0.80809	3
CUET-823	0.78120	4
Dll5143 (ours)	0.77591	5

Table 3: Results comparison of top systems for misogyny meme detection in Malayalam languages

System	F1	Rank
CUET_Novice	0.87631	1
HerWILL	0.87483	2
One_by_zero	0.86658	3
Dll5143 (ours)	0.84927	4

Detailed performance analysis of all model variants of Tamil 8.1 and Malayalam 8.2 is presented in Appendix 8.

5 Conclusion

In this study, we explored misogyny meme detection challenges in Dravidian-scripted languages through our participation in Dravidian-LangTech@NAACL 2025. Our experimentation with various models demonstrated that the majority voting mechanism achieved the best performance, with macro F1 scores of 0.77 and 0.84 for Tamil and Malayalam, respectively. The integration of majority mechanisms proved crucial in addressing model biasness in both languages. Despite achieving competitive rankings—5th in Tamil with an F1 score of 0.77 and 4th in Malayalam with an F1 score of 0.84—our analysis revealed persistent challenges as discussed in the section 6.

6 Limitations

Our work advances low-resource language processing by demonstrating how majority voting across models enhances performance. However, the model struggles with unbalanced data, particularly in detecting implicit or body-part-targeted misogyny due to cultural nuances in Tamil (9.1). Despite using robust multilingual models, they may lack script-specific features needed for Dravidian languages, especially for code-mixed or symbolic terms in Malayalam (9.2). Future work could explore specialized pre-training or script-specific

[2] https://drive.google.com/file/d/1yoXmaSYo4ZJT4AaDHzakDvlgnuUem_v6/view

[3] https://drive.google.com/file/d/1zfg30ajYWTCxPuRVaqY-caZafstetCR1/view

models to better distinguish satirical from hateful content, particularly in body-part-related contexts.

Acknowledgments

The authors express their gratitude to the Ministry of Education, the Indian Institute of Information Technology Allahabad, and the Deep Learning Lab at IIITA for providing the resources necessary to complete this work.

References

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, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, Charmathi Rajkumar, et al. 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.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

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.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

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.

H Shaun, Samyuktaa Sivakumar, R Rohan, 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.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Ashok Yadav and Vrijendra Singh. 2024. Hatefusion: Harnessing attention-based techniques for enhanced filtering and detection of implicit hate speech. *IEEE Transactions on Computational Social Systems*.

Ashok Yadav and Vrijendra Singh. 2025. Dll5143a@ nlu of devanagari script languages 2025: Detection of hate speech and targets using hierarchical attention network. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 278–288.

7 Appendix A

7.1 Majority Algorithm

Algorithm 1 Majority-Based Final Label Assignment

```
Require: Predictions from three models: $M1, M2, M3$ where $M1, M2, M3 \in \{0, 1\}$
Ensure: Final classification label $F \in \{0, 1\}$
Initialize: Model predictions $M1, M2, M3$
Compute:
    $count\_0 \leftarrow \sum_{i=1}^{3}(M_i = 0)$    ▷ Count models predicting 0
    $count\_1 \leftarrow \sum_{i=1}^{3}(M_i = 1)$    ▷ Count models predicting 1
if $count\_1 \geq 2$ then
    $F \leftarrow 1$    ▷ Assign Misogyny label
else
    $F \leftarrow 0$    ▷ Assign Non-Misogyny label
end if
return $F$
```

8 Appendix B

8.1 Misogyny in Tamil Language Results

The performance for misogyny in Tamil using model M1 is shown in Table 4. The performance for misogyny in Tamil using model M2 is presented in Table 5. The performance for misogyny in Tamil using model M3 model is shown in Table 6.

Table 4: Results for misogyny in Tamil using model M1

	prec.	rec.	f1	supp.
0	0.82	0.90	0.86	210
1	0.62	0.47	0.53	74
acc.			0.78	284
macro	0.72	0.68	0.70	284
weighted	0.77	0.78	0.77	284

Table 5: Results for misogyny in Tamil using model M2

	prec.	rec.	f1	supp.
0	0.80	0.87	0.83	210
1	0.52	0.40	0.45	74
acc.			0.75	284
macro	0.66	0.63	0.64	284
weighted	0.73	0.75	0.73	284

Table 6: Results for misogyny in Tamil using model M3

	prec.	rec.	f1	supp.
0	0.83	0.91	0.87	210
1	0.68	0.5	0.57	74
acc.			0.80	284
macro	0.76	0.70	0.72	284
weighted	0.79	0.80	0.79	284

8.2 Misogyny Detection in Malayalam Results

The performance of our proposed framework with different models in Malayalam language is discussed in this section. The performance of model M1 for Malayalam is presented in Table 7.

Table 7: Results for misogyny in Malayalam using model M1

	Prec.	Rec.	F1	Supp.
0	0.91	0.90	0.91	97
1	0.85	0.87	0.86	63
acc.			0.89	160
macro	0.88	0.89	0.88	160
weighted	0.89	0.89	0.89	160

The M2 model in Malayalam achieved an accuracy of 70% on a val set of 160 samples, as shown in Table 8.

Table 8: Results for misogyny in Malayalam using model M2

	Prec.	Rec.	F1	Supp.
0	0.83	0.63	0.72	97
1	0.59	0.80	0.68	63
acc.			0.70	160
macro	0.71	0.72	0.70	160
weighted	0.74	0.70	0.70	160

The M3 model in Malayalam achieved an accuracy of 85% on val set of 160 samples, as shown in Table 9.

Table 9: Results for misogyny in Malayalam using model M3

	Prec.	Rec.	F1	Supp.
0	0.83	0.93	0.88	97
1	0.88	0.71	0.78	63
acc.			0.85	160
macro	0.85	0.82	0.83	160
weighted	0.85	0.85	0.84	160

9 Appendix C (Error Analysis)

In the error analysis, challenges like the diglossic nature of Tamil, Sandhi (joining of words), and code-mixing add significant complexity to error detection. Understanding these linguistic features is essential for an overall analysis. Tamil is highly diglossic, meaning it has two forms of the same language, used for different purposes, and classical elements and complex morphology make its further interpretation difficult. We can take the case of Sandhi Rules, which means rules for joining two words. In Malayalam, the joining is straightforward without much alteration of the internal letters. In Tamil, words are merged smoothly, but there are letter changes. So we can highlight that it will be easier for tokenizers to handle the case of Malayalam, and as a result, we obtain a better interpretation of the context, whereas the same in the case of Tamil is difficult because it poses a challenge for transformers to split and interpret accurately. Furthermore, there are classical elements and alternatives to Tamil that make interpretation difficult; for example, most memes contain modern Tamil and might miss references to classical alternatives of words that mean the same.

Furthermore, in the Tamil dataset, most of the memes were of a code-mixed nature. And it is a known challenge in hate speech identification in code-mixed languages such as Tamil-English or Malayalam-English. It is much more challenging because of inconsistent language patterns that include vocabulary and grammar shifts that make context interpretation challenging here, as there is a switch between scripts, making tokenization and understanding difficult, and further, like we highlighted, the vocabulary confusion and grammar differences. Whereas most of the Malayalam memes were in Malayalam script only, with limited code-mixed content, explaining the performance gap between Tamil and Malayalam. We also observed that the Tamil dataset was more skewed compared to the Malayalam dataset. In the Tamil dataset, the percentage of misogynistic memes is only 25% while that of the Malayalam data set is 40%, so the model is biased towards the majority class. Such class imbalance makes the model biased towards the majority class. Some of the misclassified examples are discussed in subsequent subsections in detail.

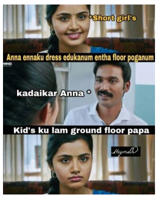
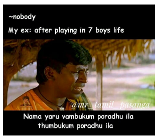
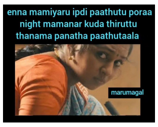
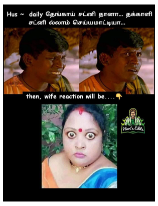
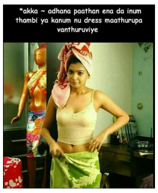
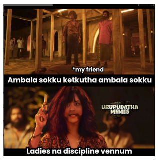
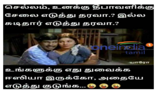
9.1 Appendix C -I

The model exhibited specific challenges in classifying memes containing symbolic language, named entities, and code-mixed expressions, particularly in understanding nuanced cultural and script-specific references in Tamil. Table 10 shows representative examples that illustrate these limitations.

The meme (ID: 1097) contains text that includes words like ’dress’ and ’floor’ alongside Tamil words, scripted in English. This is a classic example of code-mixing, demonstrating how non-standard vocabulary may not be recognized by the pretrained model, making comprehension more challenging and adding complexity. As a result, M1 misclassifies it, failing to detect the sarcastic tone and non-standard vocabulary, which imply ’women are short.’ However, M2 and M3 effectively detect misogyny, highlighting the advantage of using multiple models to capture such nuances. The meme (ID: 1163) is flagged as misogynistic by both models M2 and M3, while model M1 views the content as non-misogynistic, contrary to the original misogynistic label, as it captures the reinforcing negative stereotypes about women as manipulative and deceitful. It uses the notion of "playing" with emotions, which can be seen as trivializing the sincerity of relationships and portraying women negatively. The meme (ID: 1329) with the original label as misogynistic is captured wrongly by M2 due to different sensitivity levels and failure to capture the negative stereotypes properly, especially when playing into stereotypes of intrafamilial female conflict, and hence couldn’t capture the negative views about the behavior of women within family dynamics.

The meme (ID: 1431) reflects domestic stereotypes and cultural critique. It primarily contains Tamil text with some English, representing the wife’s reaction. The text translates to: ’You only make coconut chutney every day. Don’t you ever make tomato chutney?’ This critique of a woman’s cooking habits reflects a common domestic stereotype in Tamil culture, where women are expected to manage household chores. Such cultural nuances influence prediction, as recognizing implicit domestic criticism and stereotypes requires models to understand societal expectations and gender roles prevalent in Tamil culture. Models M1 and M3 identified the meme as misogynistic based on these implicit stereotypes and cultural patterns. However, Model M2 misinterpreted it, viewing the husband’s complaint as marital humor rather than a targeted critique of women.

Table 10: Error Analysis of Samples for Misogyny in Tamil

Image ID	M1	M2	M3	Actual Label	Image
1097	0	1	1	1	 *Short girl's Anna ennaku dress edukanum entha floor poganum kadaikar Anna * Kid's ku lam ground floor papa
1163	0	1	1	1	 ~nobody My ex: after playing in 7 boys life @mr_tamil_pasanga Nama yaru vambukum poradhu ila thumbukum poradhu ila
1329	1	0	1	1	 enna mamiyaru ipdi paathutu poraa night mamanar kuda thiruttu thanama panatha paathutaala marumagal
1431	1	0	1	1	 Hus ~ daily தேங்காய் சட்னி தானா... தக்காளி சட்னி ல்லாம் செய்யமாட்டியா... then, wife reaction will be....
1340	1	0	1	1	 *akka ~ adhana paathan ena da inum thambi ya kanum nu dress maathurupa vanthuruviye
1643	0	1	1	0	 *my friend Ambala sokku ketkutha ambala sokku URUPUDATHA MEMES Ladies na discipline vennum
1639	0	1	1	0	 செல்லம், உனக்கு தீபாவளிக்கு சேலை எடுத்து தரவா.? இல்ல சுடிதார் எடுத்து தரவா.? oneindia tamil உங்களுக்கு எது துவைக்க ஈஸியா இருக்கோ, அதையே எடுத்து குடுங்க...😛😛😛

The meme (ID: 1340) is labeled misogynistic for objectifying or inappropriately commenting on a woman’s attire. This highlights differences in how models interpret satire and humor, with M1 and

M3 recognizing misogyny, while M2 fails to do so.

The meme (ID: 1643) is labeled non-misogynistic in the original dataset, but M3 and M2 interpret certain textual or visual elements as misogynistic due to cultural context or direct translations linked to gender stereotypes. In contrast, M1 does not classify it as misogynistic. The meme (ID: 1639), originally labeled as non-misogynistic, presents a challenging scenario for model interpretation. While Models M2 and M3 detect misogyny based on visual or textual signals, Model M1 does not. This suggests that M1 better grasps context or nuances like sarcasm, which the other models overlook. In addressing the challenges posed by the detection of misogynistic content in visual media, the utilization of varying capabilities of these models to interpret nuances and contextual cues—where M1 sometimes outperforms M2 and M3 in recognizing sarcasm and cultural contexts—illustrates the diversity in model training and perspective. This diversity is instrumental in capturing a wider range of interpretations that might be missed by a single model. Even though some inaccuracies still persist in cases of majority voting, this pragmatic blend of accounting insights from multiple models ensures a more consistent and accurate interpretation in the majority of cases.

9.2 Appendix C -II

To evaluate our model's robustness in distinguishing misogyny and non-misogyny, we conducted an error analysis on misclassified instances. This analysis provided insights into common misclassification patterns. Table 11 presents examples of these errors along with interpretations for each case.

The meme (ID: 954) was originally labeled as non-misogynistic. Models M2 and M3 correctly identified that the phrase conveys admiration without any negative or objectifying language toward women. However, Model M1 misinterpreted the context, associating keywords like 'children' or 'young fellows' with biased content. Majority voting helped maintain stability, ensuring the final decision aligned with the correct label.

The meme (ID: 239) and original label as non-misogynistic, even though this was just a general discussion, model M3 predicted it as misogynistic because of the presence of both the gender and model's lack of understanding of context in deeper levels, flagging terms like "he" or "childhood" as signals for gender discussions even when the con-

Table 11: Error Analysis of Samples for Misogyny in Malayalam

Image ID	M1	M2	M3	Actual Label	Image
954	1	0	0	0	
239	0	0	1	0	
545	1	1	0	1	
725	0	1	1	1	
112	0	1	0	1	
649	1	1	0	1	
168	0	0	1	0	
317	0	0	0	0	

tent doesn't support such an interpretation. The absence of objectifying or specific gender references should have indicated a non-related context, while models M1 and M2 demonstrate a good contextual understanding. As a result, the majority voting predicts the correct label; this can again be observed with the meme (ID: 545), which translates to "Asking the neighbor's sister for a game," which directly pertains to objectification and inappropriate propositions. Again, Model M3 fails to understand the context of intent, while Models M1 and M2 correctly identify the content as related by recognizing the objectifying request towards women.

Moving to the meme (ID: 725), which is labeled as misogynistic and makes remarks about women's "backs" and has an image of a woman and the phrase "Back, That's all". Here, model M1 failed to map the woman's image and its phrase and recognize the context in which the word "back" is used, while models M2 and M3 were effective in detecting objectifying language even when it pertains to specific body parts, and so the majority voted.

The meme (ID: 112) labeled as misogynistic is predicted the opposite by majority voting, and models M1 and M3 failed to map the remarks "flower," which actually made remarks about women's body parts. Both models failed to detect the explicit objectifying content, possibly due to focusing on specific keywords without fully understanding the context or the nature of the remarks. While Model M2 was effective in mapping and recognizing the indirect remarks about women's body parts, the suggests we need to further work on these models to make them more robust and make these model's aware to map these indirect references correctly.

The meme (ID: 649) is marked as misogynistic, with the translated keyword "There's an aunty like this everywhere, to stir up the crowd," where model M1 and M2 were correctly able to catch the explicit objectifying language, whereas we can say that model M3 requires better contextual sensitivity to identify nuanced objectifying or critical statements about gender dynamics.

The meme (ID: 168) depicts a lighthearted conversation: 'BF: When are we getting married? GF: When the movie releases. And that movie is not getting released anytime soon.' It is labeled as non-misogynistic as it simply portrays a humorous situation. However, Model M3 incorrectly predicts it as misogynistic, associating it with negative stereotypes related to women's reliability or commitment. In contrast, the other models correctly identify it as unrelated to misogyny. Detecting misogynistic content in Malayalam visual media presents significant challenges, particularly in interpreting nuanced cultural contexts, implicit biases, and subtle forms of objectification. While leveraging the strengths of different models and majority voting helps mitigate individual model biases, it occasionally propagates errors when multiple models misclassify similar content. However, majority voting generally improves performance in ambiguous cases. For example, in the case of meme (ID: 317), the models remain accurate despite the presence of women in the image, as the context is political rather than misogynistic.

Nevertheless, the complexity of Malayalam's linguistic and societal nuances, subtle misogyny, and indirect references remains a challenge. Addressing these issues requires integrating culturally specific and contextually rich datasets, enhancing models' ability to recognize implicit biases, and incorporating advanced techniques such as weighted ensemble methods and fine-grained contextual embeddings in future research.