CUET-NLP_Big_O@DravidianLangTech 2025: A Multimodal Fusion-based Approach for Identifying Misogyny Memes

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904007, u1904086, u1904102, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Memes have become one of the main mediums for expressing ideas, humor, and opinions through visual-textual content on social media. The same medium has been used to propagate harmful ideologies, such as misogyny, that undermine gender equality and perpetuate harmful stereotypes. Identifying misogynistic memes is particularly challenging in low-resource languages (LRLs), such as Tamil and Malayalam, due to the scarcity of annotated datasets and sophisticated tools. Therefore, DravidianLangTech@NAACL 2025 launched a Shared Task on Misogyny Meme Detection to identify misogyny memes. For this task, this work exploited an extensive array of models, including machine learning (LR, RF, SVM, and XGBoost), and deep learning (CNN, BiLSTM+CNN, CNN+GRU, and LSTM) are explored to extract textual features, while CNN, BiLSTM + CNN, ResNet50, and DenseNet121 are utilized for visual features. Furthermore, we have explored feature-level and decision-level fusion techniques with several model combinations like MuRIL with ResNet50, MuRIL with BiLSTM+CNN, T5+MuRIL with ResNet50, and mBERT with ResNet50. The evaluation results demonstrated that BERT + ResNet50 performed best, obtaining an F1 score of 0.81716 (Tamil) and were ranked 2nd in the task. The early fusion of MuRIL+ResNet50 showed the highest F1 score of 0.82531 and received a 9th rank in Malayalam.

1 Introduction

The unprecedented proliferation of social media has brought about an exponential increase in meme-based communication, where image and text combine in powerful messages of influence in public opinion (Singh et al., 2024). Memes are primarily vehicles for humor and social commentary (Ponnusamy et al., 2024); however, in multilingual contexts, they have increasingly been used as a conduit for misogynistic content (Suryawanshi et al., 2020b; a P K et al., 2020). Misogynistic memes are digital seeds of negativity disguised as humor that perpetuate harmful stereotypes and normalize disrespect toward women (H et al., 2024; Singh et al., 2024). Hence, determining whether shared content on social media is misogynistic or not is necessary.

Although much recent work has explored the analysis of emotions conveyed in memes (Mishra et al., 2023), the identification of offensive and hate content in memes (Hermida and Santos, 2023; Rizwan et al., 2024), focused on identifying misogynistic content in memes in high-resource languages such as English (Farinango Cuervo and Parde, 2022). The challenge remains in LRLs, such as Tamil and Malayalam, due to a lack of resources (Magueresse et al., 2020), linguistic complexity, and cultural quirks (Kumari et al., 2023). To address these challenges, the DravidianLangTech@NAACL 2025 Shared Task on Misogyny Meme Detection[1] focused on Tamil and Malayalam languages (Ponnusamy et al., 2024; Chakravarthi et al., 2025), two widely spoken Dravidian languages with distinct linguistic characteristics. The challenge aims to develop a robust approach to identify misogynistic content in memes in these languages. It requires processing visual and textual information and grasping their combined meaning in cultural and linguistic contexts. In this paper, we take an integrated approach to this challenge, leveraging the power of advanced models in both visual and textual modalities. Hence, the contributions of the work are as follows:

- Developed a multimodal architecture that effectively combines visual and textual features for misogynistic content detection in Tamil and Malayalam languages.

[1] https://codalab.lisn.upsaclay.fr/competitions/20856

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 505–512

- Investigated various ML, DL, and transformer-based models with different fusion techniques to identify misogynistic memes while evaluating performance metrics and conducting error analysis to determine the best strategy for detecting misogynistic content in both languages.

2 Related Work

While the meme culture is going strong, there has been significant research into detecting trolling, hostility, offensive, and abusive language from social media data, with several studies conducted by researchers (Suryawanshi et al., 2020b,a; Kumari et al., 2023; H et al., 2024) in recent years. Many studies have focused solely on textual features to detect harmful content (Sreelakshmi et al., 2020; Baruah et al., 2020). However, many researchers have investigated memes' textual and visual features to classify trolls, offenses, and aggression. For instance, Suryawanshi et al. (2020a) introduced the MultiOFF[2] meme dataset, containing 743 memes for detecting offensive content and proposed a stacked LSTM with VGG16 approach for multi-modal analysis, which outperformed other models using a single feature, achieving an F1-score of 0.50. Another study by Sultan et al. (2024) introduced MemesViTa, a multi-modal fusion model combining Vision Transformer (ViT) and DeBERTa, which achieved 94.29% accuracy and 95.82% F1 score, surpassing both visual and textual models in troll meme detection. However, in the context of misogyny meme identification, several works on misogyny meme identification focus on a linguistic perspective (Anzovino et al., 2018), proposing a corpus of misogynistic tweets and exploring machine learning models for detection. Butt et al. (2021) tackled sexism detection in multilingual social media text, achieving an F1 score of 0.78 for sexism identification and 0.49 for categorization using data augmentation. A data augmentation approach using song lyrics was introduced to improve misogyny detection, outperforming conventional transfer learning techniques on English and Spanish datasets (Calderón-Suarez et al., 2023).

Several studies have revealed multi-modal approaches for misogyny content detection. Ponnusamy et al. (2024) developed the MDMD (Misogyny Detection Meme Dataset) to analyze misogyny, gender bias, and stereotypes in Tamil and Malayalam-speaking communities through memes. Rizzi et al. (2023) evaluated four uni-modal and three multi-modal approaches for detecting misogynistic memes, introducing a bias estimation technique and Bayesian Optimization to improve accuracy by 61.43%. Another work done by Singh et al. (2024) achieved a 0.73 F1-score on 5,054 Hindi-English memes using BiT+MuRIL, outperforming unimodal models. H et al. (2024) used MNB for text and ResNet50 for images, achieving F1-scores of 0.69 (Tamil) and 0.82 (Malayalam) in LT–EDI 2024[3] shared task. Although hateful and offensive memes are well-studied, misogynistic meme detection in Tamil and Malayalam remains unexplored. This work improves previous efforts by integrating visual and textual modalities for better performance.

3 Task and Dataset Description

This study aims to detect misogynistic content in memes using textual and visual features, framing it as a binary classification task in Tamil and Malayalam languages. The given dataset (Ponnusamy et al., 2024; Chakravarthi et al., 2024) contains a total of 1776 data points in the Tamil language and 1000 data points in the Malayalam language, combining train, validation, and test set. Table 1 shows the class-wise distribution of samples for the Tamil dataset, highlighting an imbalance with more *Non-misogyny* samples (851 train, 210 valid, 267 test) compared to *Misogyny* samples (285 train, 74 valid, 89 test).

Classes	**Train**	**Valid**	**Test**	$\mathbf{W}_T$	$\mathbf{UW}_T$
Non-misogyny	851	210	267	26770	15194
Misogyny	285	74	89	9243	6749
Total	1136	284	356	36013	21943

Table 1: Class-wise distribution of train, validation, and test set for the Tamil language, where W_T and UW_T denote total words and total unique words in three datasets.

Table 2 presents the class-wise distribution for the Malayalam dataset, showing an imbalance with more *Non-misogyny samples* (381 train, 97 valid, 122 test) compared to *Misogyny* samples (259 train, 63 valid, 78 test). This imbalance, coupled with the smaller size and vocabulary of the *Misogyny* class,

[2] https://shorturl.at/DEyxx

[3] https://codalab.lisn.upsaclay.fr/competitions/16097

poses challenges for model performance.

Classes	Train	Valid	Test	W_T	UW_T
Non-misogyny	381	97	122	11004	7574
Misogyny	259	63	78	6398	4378
Total	640	160	200	17402	11952

Table 2: Class-wise distribution of train, validation, and test set for the Malayalam language, where W_T and UW_T denote total words and total unique words in three datasets.

The implementation details of the tasks will be found in the GitHub repository[4].

4 Methodology

Several ML, DL, and transformer-based models were explored to develop a framework for Misogyny meme detection (Figure 1).

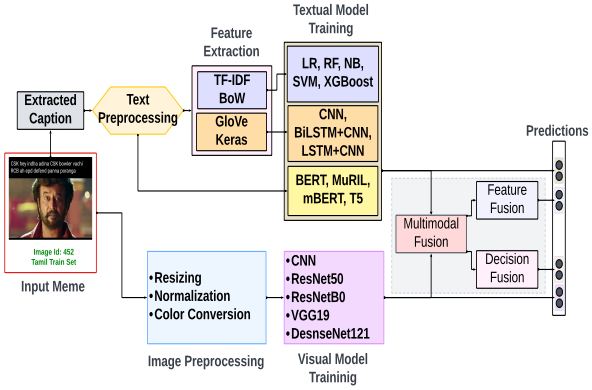

Figure 1: Schematic process of Misogyny meme detection.

4.1 Data Preprocessing

Text preprocessing included language-specific tokenization, such as MuRIL tokenizer for Malayalam and Tamil, with a maximum sequence length of 128 tokens, handling special characters, preserving language-specific Unicode characters, and eliminating extra whitespace. Image preprocessing involved the most common transformations, such as scaling images to 224x224 pixels, converting to RGB format, and normalizing using ImageNet statistics *(mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])*. The preprocessing pipeline maintains different configurations for the training and inference phases.

[4] https://github.com/RJ-Hossan/MMD-NAACL-2025

4.2 Feature Extraction

TF-IDF and Bag of Words (BoW) were used to extract textual features for ML models. In TF-IDF, up to 5000 features were extracted, and stop words were removed. DL models used 100-dimensional pre-trained GloVe embeddings. An embedding matrix was built to map vocabulary words to vectors. For out-of-vocabulary terms, zero vectors were used. Text tokenization and vocabulary creation were performed using a *CountVectorizer* with 7000 features, while contextual embeddings were reduced using a fully connected layer. For images, pre-trained models extracted visual features aligned with the pre-processing pipeline.

4.3 Baselines

Several unimodal (visual or textual) and multimodal (visual and textual) models were explored and fused using early and late fusion approaches, with the necessary hyperparameter tuning, to perform the tasks.

4.3.1 Unimodal Baselines

Various ML and DL models were utilized to develop the unimodal approach. For textual features, traditional models such as LR, SVM, RF, and Gradient Boosting and deep learning models such as BiLSTM, CNN, and BiLSTM+CNN were employed, using GloVe and Keras-based embedding. CNN-based architectures, such as DenseNet-121, EfficientNet-B0, ResNet-50, and VGG19, were implemented for visual features.

Table 3 outlines the hyperparameters for the LR, RF, SVM, and XGBoost models trained on textual features. LR uses *max_iter=1000* and *random_state=27* whereas Random Forest sets *n_estimators=100* and *random_state=27*. The SVM used a specified kernel. XGBoost adopted *eval_metric=mlogloss*. Each model has different parameter configurations.

Parameter	LR	RF	SVM	XGBoost
max_iter	1000	-	-	-
random_state	27	27	15	15
n_estimators	-	100	-	-
kernel	-	-	linear	-
probability	-	-	True	-
use_label_encoder	-	-	-	False
eval_metric	-	-	-	mlogloss

Table 3: Parameters used for ML models (textual features only).

Table 4 presents the tuned hyperparameters for deep learning (DL) models using textual features, including BiLSTM+CNN, Text-CNN, LSTM+CNN, and BiLSTM. It outlines parameters such as *embedding size*, *max sequence length, hidden dimension, filter sizes, number of layers, dropout rate, learning rate*, and *optimizer (Adam)*. The loss function varies between *CrossEntropyLoss* and *BinaryCrossEntropyLoss*.

Hyperparameter	BiLSTM+CNN	Text-CNN	LSTM+CNN	BiLSTM
embedding_size	100 (GloVe vectors)	100 (GloVe vectors)	100 (GloVe vectors)	100 (GloVe vectors)
max_sequence_length	100	100	100	100
hidden_dimension	256	-	256	256
number_of_filters	-	128	128	-
filter_sizes	-	[3, 4, 5]	[3]	-
number_of_layers	2	-	-	2
batch_size	32	32	32	32
learning_rate	0.001 (Adam)	0.001 (Adam)	0.001 (Adam)	0.001 (Adam)
epochs	45	45	45	45
dropout_rate	-	0.5	-	-
loss_function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
optimizer	Adam	Adam	Adam	Adam

Table 4: Tuned hyperparameters for DL models (textual features only).

Table 5 outlines the tuned hyperparameters for DL models using visual features, including CNN, VGG19, EfficientNetB0, and DenseNet121. It specifies *input size, preprocessing techniques, learning rate, loss function, batch size, activation function (ReLU, Sigmoid),* and *fine-tuning strategy* (whether the layers are frozen or not).

Hyperparameter	CNN	VGG19	EfficientNetB0	DenseNet121
Input Size	(224, 224, 3)	(224, 224, 3)	(224, 224, 3)	(224, 224, 3)
Base Model	-	VGG19	EfficientNetB0	DenseNet121
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	0.001	0.001	0.001	0.001
Loss Function	Binary Crossentropy	Binary Crossentropy	Binary Crossentropy	Binary Crossentropy
Epochs	20	20	20	20
Batch Size	32	32	32	32
Activation Function	ReLU, Sigmoid	ReLU, Sigmoid	ReLU, Sigmoid	ReLU, Sigmoid
Dropout Rate	0.5	0.5	0.5	0.5
Fine-Tuning	No (frozen layers)	No (frozen layers)	No (frozen layers)	No (frozen layers)

Table 5: Tuned hyperparameters for DL models (visual features only).

4.3.2 Multimodal Baselines

The multimodal baseline models combined textual and visual features using early fusion (EF) and late fusion (LF) techniques. Models like T5+MuRIL+ResNet-50, BiLSTM+CNN+MuRIL, MuRIL+ResNet-50, and BERT+ResNet-50 utilized pre-trained models for feature extraction. For instance, the text is tokenized in Tamil using BERT with padding and truncation to a fixed length of 128 tokens and processed through a fully connected layer to reduce dimensionality to 256. The images were resized to 224x224 pixels and processed through ResNet-50, with the final fully connected layer replaced to output 256 features. Textual and visual features were concatenated into a single vector of size 512, which was passed through a classifier. The training used the AdamW optimizer with a learning rate of 2e-5, a batch size of 16, and a learning rate scheduler that reduced the rate when validation loss plateaus. Despite differences in dataset size, class distribution, and language complexity between Tamil and Malayalam, the tuned hyperparameters for both tasks were kept similar and are presented in Table 6. Moreover, similar models, i.e., BiLSTM+CNN+MuRIL, MuRIL+ResNet-50, and T5+MuRIL+ResNet-50, were also employed with late fusion, where fusion happened only at the decision level after both modalities had been independently processed.

Hyperparameter	BERT+ResNet50 (Tamil)	MuRIL+ResNet50 (Malayalam)
Learning Rate	2e-5	2e-5
Batch Size	16	16
Number of Epochs	45	45
Max Sequence Length	128	128
Optimizer	AdamW	AdamW
Dropout Rate	0.3	0.3
Image Model	ResNet50	ResNet50
Text Model	BERT (base-uncased)	MuRIL (base)
Scheduler	ReduceLROnPlateau	ReduceLROnPlateau

Table 6: Tuned hyperparameters used in multimodal fusion for Tamil and Malayalam languages.

4.4 System Requirements

Most models, specifically fusion models, were trained on a dual GPU setup (NVIDIA Tesla T4x2), utilizing parallel processing for textual and visual features. The BERT+ResNet50 model used 7-8 GB of GPU memory, while MuRIL+ResNet50 required approximately 8-10 GB of GPU memory. Training for 45 epochs for both approaches took 120-150 minutes, depending on the size of the data set and the calculation of the class weight.

5 Result Analysis

Table 7 provides a comparative analysis of the performance of different approaches used to detect misogynistic memes in Tamil and Malayalam datasets. In textual-only methodologies, conventional models such as SVM and Gradient Boosting performed well, with F1 scores of 0.6507 and 0.6848 (Tamil) and 0.6693 and 0.7058 (Malayalam). As for deep learning models, BiLSTM (Glove) had the lowest F1 scores (0.4826 Tamil, 0.5187 Malayalam), while BiLSTM+CNN (Glove) performed best for Tamil (0.6703) and LSTM+CNN (Glove) for Malayalam (0.6448).

Regarding visual-only models, EfficientNet-B0 performed better in both languages, achieving an F1 score of 0.6546 for Tamil and 0.7640 for Malayalam. The VGG19 model had shown excellent performance, especially for the Malayalam language,

Approaches	Classifiers	Tamil				Malayalam			
		P	R	F1	G	P	R	F1	G
Textual Only	LR	0.7871	0.5618	0.5496	0.6650	0.7128	0.6803	0.6858	0.6964
	SVM	0.7027	0.6348	0.6507	0.6679	0.6727	0.6673	0.6693	0.6700
	RF	0.7720	0.5805	0.5807	0.6694	0.6895	0.6963	0.6911	0.6929
	Gradient Boosting	0.7197	0.6635	0.6848	0.6910	0.7058	0.7058	0.7058	0.7058
	BiLSTM (Glove)	0.4750	0.5000	0.4826	0.4873	0.5750	0.5167	0.5187	0.5451
	CNN (Glove)	0.6347	0.6367	0.6357	0.6357	0.5922	0.5632	0.5514	0.5775
	LSTM+CNN (Glove)	0.6420	0.6592	0.6480	0.6505	0.6884	0.6836	0.6448	0.6860
	BiLSTM+CNN (Glove)	0.6774	0.6648	0.6703	0.6711	0.6426	0.6318	0.6305	0.6372
Visual Only	CNN	0.6146	0.5787	0.5844	0.5964	0.6619	0.6568	0.6587	0.6593
	DenseNet-121	0.8072	0.5787	0.5786	0.6835	0.7120	0.6693	0.6739	0.6903
	EfficientNet-B0	0.8332	0.6330	0.6546	0.7262	0.7616	0.7722	0.7640	0.7669
	ResNet-50	0.6782	0.5843	0.5899	0.6295	0.8134	0.8058	0.8007	0.8096
	VGG19	0.7448	0.6854	0.7044	0.7145	0.8134	0.8058	0.8215	0.8096
Multi-modal Fusion (EF)	T5+MuRIL+ResNet-50	0.8170	0.8146	0.8158	0.8158	-	-	-	-
	BiLSTM+CNN+MuRIL	0.8365	0.7865	0.8065	0.8111	0.8270	0.8255	0.8262	0.8262
	MuRIL+ResNet-50	0.8281	0.7981	0.8013	0.8130	**0.8451**	**0.8157**	**0.8253**	**0.8303**
	BERT+ResNet-50	**0.8160**	**0.8184**	**0.8172**	**0.8172**	-	-	-	-
Multi-modal Fusion (LF)	T5+MuRIL+ResNet-50	0.8315	0.7940	0.8099	0.8125	-	-	-	-
	BiLSTM+CNN+MuRIL	0.8160	0.7884	0.8005	0.8021	0.8374	0.8226	0.8284	0.8299
	MuRIL+ResNet-50	0.8178	0.8031	0.8017	0.8104	0.8044	0.7911	0.7962	0.7977
	mBERT+EfficientNet-B0	-	-	-	-	0.7844	0.7857	0.7851	0.7850

Table 7: Result comparison on test data, where EF, LF, P, R, F1, and G denote early fusion, late fusion, precision, recall, F1-score, and geometric mean score of precision and recall, respectively.

achieving an F1 score of 0.8215. Concerning multimodal fusion, EF models like BERT+ResNet-50 performed better than others in Tamil with an F1 score of 0.8172, which helped us rank 2nd in this task. During our observation after the competition, we noted that the BiLSTM+CNN+MuRIL model outperformed the MuRIL+ResNet-50 model in Malayalam with a higher F1 score of 0.8262 but had a lower G score[5] of 0.8262 compared to 0.8303. In addition to this, the LF models performed well, as BiLSTM+CNN+MuRIL reached an F1-score of 0.8284 for the Malayalam language. Appendix A demonstrates a comprehensive error analysis of the employed models.

6 Conclusion

This paper demonstrated a shared task solution to detect misogynistic memes in Tamil and Malayalam that exploited textual and visual characteristics. The results showed that the multimodal approach BERT+ResNet-50 with early fusion achieved the highest F1 score of 0.8172 in Tamil. However, MuRIL with ResNet50 outperformed all models and obtained the highest F1 score (0.8253) through early fusion in Malayalam. Although the results are promising, the current approach has room for further improvement. Advanced preprocessing techniques, such as data augmentation, would enrich the dataset and improve the generalization of the model. Future work aims to explore vision-based transformer models and advanced multimodal techniques (i.e., CLIP) for enhanced performance.

7 Limitations

The study on misogynistic content identification in multimodal memes has several drawbacks, influenced by following factors:

- Fine-tuned DL and transformer models may fail when meme contexts differ from training data.
- Since the dataset used for training the models was imbalanced and no advanced augmentation was employed, it could have led to biased predictions in another set of memes.
- Although multimodal fusion approaches showed strong results, the complexity of combining multiple models and managing text-image interactions may have caused computational inefficiencies and overfitting, limiting scalability.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

[5]The G score of precision and recall is the square root of the product of precision and recall.

References

Abdul Rasheed a P K, Carmel Jose, and Anju Michael. 2020. Social media and meme culture: A study on the impact of internet memes in reference with 'kudathai murder case'.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. *Automatic Identification and Classification of Misogynistic Language on Twitter*, pages 57–64.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 76–82, Marseille, France. European Language Resources Association (ELRA).

Sabur Butt, Noman Ashraf, Alexander Gelbukh, and Grigori Sidorov. 2021. Sexism identification using bert and data augmentation–exist2021.

Ricardo Calderón-Suarez, Rosa M. Ortega-Mendoza, Manuel Montes-Y-Gómez, Carina Toxqui-Quitl, and Marco A. Márquez-Vera. 2023. Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases. *IEEE Access*, 11:13179–13190.

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.

Charic Farinango Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 785–792, Seattle, United States. Association for Computational Linguistics.

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.

Paulo Cezar de Q. Hermida and Eulanda M. dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, 82:1–36.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *ArXiv*, abs/2006.07264.

Shreyash Mishra, Suryavardan S, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes.

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.

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, and Animesh Mukherjee. 2024. Zero shot vlms for hate meme detection: Are we there yet? *Preprint*, arXiv:2402.12198.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

K. Sreelakshmi, Premjith B., and Soman Kp. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Tipu Sultan, Mohammad Abu Tareq Rony, Mohammad Shariful Islam, Saad Aldosary, and Walid El-Shafai. 2024. Memesvita: A novel multimodal fusion technique for troll memes identification. *IEEE Access*, 12:177811–177828.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

A Error Analysis

We conducted quantitative and qualitative error analyses to gain a deeper understanding of the performance of the best-performed model.

Quantitative Analysis: The best-performing models were used for a quantitative error analysis, utilizing confusion matrices for Tamil and Malayalam to identify misogynistic memes. Figure A.1 demonstrated that the proposed BERT+ResNet50 model using late fusion revealed strong overall performance with an accuracy of 86.2%.

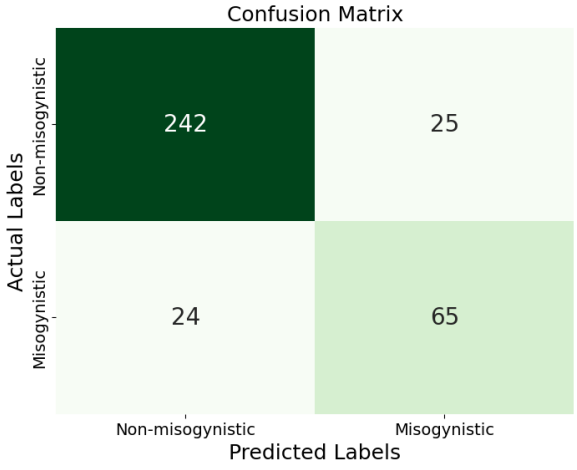

Figure A.1: Confusion matrix of the proposed approach (BERT+ResNet50 by early fusion) for Tamil language.

The model correctly identifies 242 *non-misogynistic* memes and 65 *misogynistic* memes while misclassifying 25 *non-misogynistic* memes as *misogynistic* and missing 24 *misogynistic* memes. The misclassifications arise from visually complex memes with overlapping misogynistic and non-misogynistic elements and subtle text based on sarcasm or cultural context. We have found that the model struggles with imbalanced data, leading to the misclassification of most non-misogynistic memes as misogynistic due to the subtlety in textual complexity.

Figure A.2 shows the confusion matrix to identify misogyny memes for the Malayalam language, using the fusion of MuRIL (for text) and ResNet50 (for images). It shows that 113 *non-misogynistic* memes are correctly identified, while 55 *misogynistic* memes are accurately identified as *misogynistic* from 78 misogynistic memes. However, 9 *non-misogynistic* memes are incorrectly classified as *misogynistic*, and 23 *misogynistic* memes are missed. These outcomes indicate that the model performs relatively well but still has some drawbacks due to poor handling of data imbalance and the dynamic nature of the contextual meanings of memes.

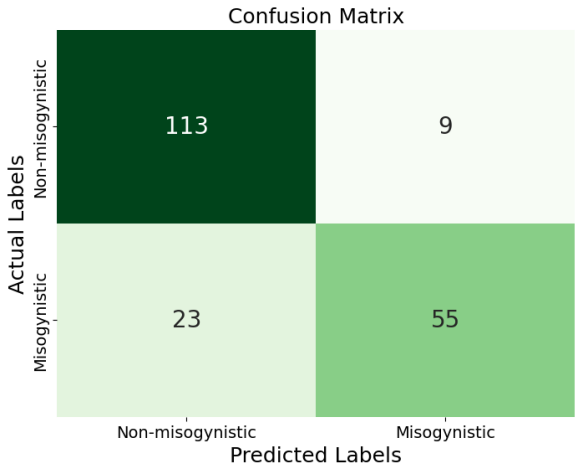

Figure A.2: Confusion matrix of the proposed approach (MuRIL+ResNet50 by early fusion) for Malayalam language.

Qualitative Analysis: Figure A.3 and A.4 highlight the best-performed model's predicted outputs for sample inputs in identifying misogynistic memes for both Tamil and Malayalam datasets. In Figure A.3, the proposed model (BERT+ResNet50) accurately predicted samples 2 and 3 but incorrectly predicted samples 1 and 4 in Tamil, indicating some prediction inconsistencies. Similarly, Figure A.4 illustrates the model's performance in Malayalam, where it correctly identified samples 2 and 3 but incorrectly predicted samples 1 and 4. These errors may be due to class imbalance, as the *misogyny* class in the Malayalam dataset contains only 78 instances, which likely impacts the model's generalization capability.

Image Id: 1064.jpg
True Label: Misogyny (1)
Predicted Label: Non-misogyny (0)

Sample 1

Image Id: 1258.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 2

Image Id: 1396.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 3

Image Id: 1563.jpg
True Label: Non-misogyny (0)
Predicted Label: Misogyny (1)

Sample 4

Image Id: 545.jpg
True Label: Misogyny (1)
Predicted Label: Non-misogyny (0)

Sample 1

Image Id: 488.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 2

Image Id: 431.jpg
True Label: Misogyny (1)
Predicted Label: Misogyny (1)

Sample 3

Image Id: 954.jpg
True Label: Non-misogyny (0)
Predicted Label: Misogyny (1)

Sample 4

Figure A.3: Few sample predictions by the BERT+ResNet50 for the Tamil language.

Figure A.4: Some predicted outputs by the MuRIL+ResNet50 for the Malayalam language.