LinguAIsts@DravidianLangTech 2025: Misogyny Meme Detection using multimodel Approach

ARTHI R[1], **Pavithra J** [2], **G Manikandan** [3], **Lekhashree A**[4]
Dhanyashree G[5], **Bommineni Sahitya**[6], **Arivuchudar K**[7], **Kalpana K**[8]
R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{arth22004, pavi22039, mgk, lekh22026}.ad@rmkec.ac.in
{dhan22012, bomm22009, ariv22002, kalp22020}.ad@rmkec.ac.in

Abstract

Memes often disseminate misogynistic material, which nurtures gender discrimination and stereotyping. While it is an effective tool of communication, social media has also provided a fertile ground for online abuse. This vital issue in the multilingual and multimodal setting is tackled by the Misogyny Meme Detection Shared Task. Our method employs advanced NLP techniques and machine learning models to classify memes in Malayalam and Tamil, two low-resource languages. Preprocessing of text includes tokenization, lemmatization, and stop word removal. Features are then extracted using TF-IDF. With the best achievable hyperparameters, along with the SVM model, our system provided very promising outcomes and ranked 9th among the systems competing in the Tamil task with a 0.71259 F1-score, and ranked 15th with an F1-score of 0.68186 in the Malayalam taks. With this research work, it would be established how important AI-based solutions are toward stopping online harassment and developing secure online spaces.

1 Introduction

Social media has enabled international artistic exchange but is also a platform for the circulation of harmful content, especially gender-based abuse. There is growing concern about the proliferation of misogynistic memes, which are a group of images and text that support anti-woman discourses Ponnusamy et al., 2024. Since they reinforce negative gender norms and stereotypes, their identification becomes crucial in widespread terms and Chakravarthi, 2021.

Earlier research dealt with all dimensions of on- line hate speech detection. Transformer models such as BERT has made tremendous progress in language understanding (Devlin et al., 2019). It is evident that deep learning techniques have been used to identify offensive language . The detection of hate speech in multimodal content has also been studied, emphasizing how such models need to be able to process both text and images Gomez et al., 2020. Further, multilingual NLP approaches, like MuRIL, have advanced abusive content detection in Indian lan- guages (Khanuja et al., 2021). But for misogynistic memes, there is a specific approach that is needed, one that combines both linguistic and visual fea- tures Suryawanshi et al., 2021.

Figure 1: Sample for misogynistic meme in Tamil content.

The situation is particularly challenging for low- resource languages like Tamil and Malayalam, where content moderation technologies are rela- tively weak compared to high-resource languages (Thavareesan et al.,2019). Most languages glob- ally have inadequate datasets and tools necessary for filtering harmful content, which facilitates the propagations of misogynistic material unabatedly (Bishop, 2014). The issue is multilingual and mul- timodal; therefore, automated detection systems need to be built in order to counter gender-based abuse effectively.

Therefore, DravidianLangTech@NAACL 2025 introduces this Shared Task on Misogyny Meme Detection, focusing on low-resource languages to determine the distinction between misogynistic and non-misogynistic memes. This initiative not only addresses an urgent social issue but also fills a gap in research by developing AI-based systems for ethical content moderation in underrepresented lan- guage communities

We present an approach that relies on sophis- ticated Natural Language Processing (NLP) tech- niques to

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 387–392

Figure 2: Sample for non-misogynistic meme in Tamil con- tent.

analyze the textual aspect of memes. To- kenization splits the text into manageable units, lemmatization reduces words to their root forms while maintaining meaning, and Term Frequency- Inverse Document Frequency (TF-IDF) points out the most important terms in the dataset. For classifi- cation, we rely on Support Vector Machines (SVM), a machine learning algorithm famous for its ef- ficiency in high-dimensional data environments (Suryawanshi et al., 2021). Our work, through the incorporation of low- resource languages and multimodal analysis, offers a scalable and inclusive solution to gender-based abuse across various online platforms. Our study, titled Towards Protecting Marginalized Communi- ties: Mitigating Inferences from Large Language Models, aims to establish a robust system for de- tecting sexism in online content while laying the foundation for future applications in multilingual and multimodal AI research.

2 Related Work

Training a misogyny detectory feature on memes has had a lot of traction lately, especially with re- cent studies done on low resource language like Tamil and Malayalam, but we just scratched the surface. Datasets such as HASOC (Kumar et al., 2021) and Mandl et al. (2020) are oriented towards hate speech in Hindi and Bengali while the Hate- ful Memes Challenge (Kiela et al., 2020) aims to multimodal hate speech detection in English, trans- ferring these approaches to detect gender-based hate in multilingual settings is still a challenge. Some noteworthy examples include research such as Fersini et al. (2018)—while there is a multi-level sexism detection, English and Spanish based, such multilingual multimodal datasets targeting misogy- nistic memes in languages like Tamil and Malay- alam remain insufficient. Ponnusamy et al., 2024 does an excellent job at bridging this gap by pro- viding an annotated dataset for misogyny detection in Tamil and Malayalam memes. Overall, njihova dataset and framework provide a necessary step in the right direction for being able to witnes and identify sexist abuse in local social media mate- rial. We extend their work by introducing a SVM classifiers with TF-IDF feature extraction and text pre-processing for challenging low-resource languages. With these techniques, we hope to have a model that does better than others in the task of misogynistic detection in memes, thus vastly im proving the study of gender bias in multilingual platforms.

3 System Description

In this section, we give a detailed description of the dataset and offer more information about the experiments that were carried out for the study. In Figure 3, the system architecture for misogynistic and non-misogynistic meme classification utilising machine learning (ML) approaches, including Grid- SearchCV for hyperparameter tuning, is depicted. The whole classification process flow is depicted in the diagram, which highlights the crucial steps in identifying misogynistic content in low-resource 2 languages like Tamil and Malayalam.

3.1 Dataset

The dataset used in this research consists of text memes of Tamil and Malayalam. The dataset for all languages is split into train, validate and test parts. The dataset is annotated into 2 classes: Misogy- nistic and Non-Misogynistic.Table 1 gives distribu- tion details over the dataset and number of samples available in each subset for both languages.

Language	Train	Validate	Test
Tamil	1137	285	357
Tamil	641	161	201

Table 1: Dataset distribution for Tamil and Malayalam.

The datasets were adapted and modified from publicly accessible datasets originally published as part the DravidianLangTech@NAACL program to suit the specific context of this study. This program focuses on building language technology resources and tools for low-resource Dravidian languages.

4 Methodology

In general, there are four processes involved in Identification of Misogynistic and Non- Misogynistic memes in Tamil and Malayalam such as EDA, Preprocessing, Modelling and Evaluation. In EDA, we explored the dataset and analysed ob- servation of meme linguistics. Pre-processing To- kenisation and lemmatisation

were performed to clean and standardise the raw text. During the Modelling phase, Support Vector Machines (SVM) were used to classify the memes. In the Evaluation phase, we used metrics (Tuning phase) such as accuracy, F1 score, precision, and recall to evaluate the performance of the train model. F1 score of a balanced classification was the measure of per- formance. This comprehensive strategy ensured the accurate classification of memes while consid- ering the subtleties of the Tamil and Malayalam languages.

4.1 Exploratory Data Analysis

Through exploratory data analysis (EDA) it is found that there are some important linguistic patterns present in Misogynistic and Non-Misogynistic memes in Tamil and Malayalam. The common occurence of word pari/triples were determinated with N-gram Analysis while common associated words to sexist material were identified using Term Frequency Analysis. These attempts made informed additions to the set of features for training the models, such as highlighting abusive phrases and identifying certain linguistic patterns that significantly improved the meme classification accuracy.

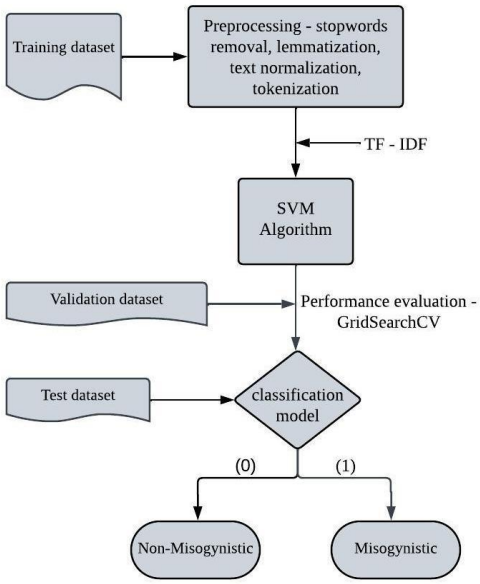

Figure 3: System Architecture for classify misogyny meme using ML models.

4.2 Preprocessing

We simplified the language by removing punctu- ation, converting all characters to lowercase, and removing numeric characters. In order to retain meaning when a variety of words were included, lemmatisation was used to reduce words to their lowest forms. The revised text was subsequently converted into numerical features leveraging TF- IDF vectorisation, using bigrams and unigrams to encapsulate both isolated and contextually corre- lated word associations.

4.3 Machine Learning Model

For the classfication of Misogynistic and Non-Misogynistic memes in Tamil and Malayalam we used Support Vector Machines(SVM) as the core machine learning based model. SVM was chosen for being effective with textual data and for its power to deal with the intricacies of classifying memes. The model was selected due to its capability of performing binary classification which was a task given to identify a meme as misogynistic or non-misogynistic.

- **Support Vector Machine (SVM)** The best hyperplane for separating the Misogynistic and Non-Misogynistic memes in the high- dimensional TF-IDF vector/matrix was found by using a linear kernel in the SVM model. This was indicative of strong text-based data handling, as well as generalization. Accura- cies were found to be 66fying Tamil memes and Malayalam memes re- spectively indicative of the difficulty faced in classifying memes in low resource languages.

4.4 Model Evaluation

Important metrics (accuracy, F1 score, precision, recall) were used for the assessing the model for balanced classification of Misogynistic and Non- Misogynistic memes. The accuracy for Tamil memes was 78resents a good performance, but reveals that there is work to do in translating between precision and recall. For Malayalam memes, the same model's F1 score was 71were somewhat better and showed that the model had handled the characteristics of the Malayalam language better than others. Cross-validation was used to ensure robustness, while the F1 score was prioritized to maintain the balance between preci- sion and recall.

5 Results

We evaluated our model effectiveness at Misogy- nistic and Non-Misogynistic meme detection us- ing the macro-average f1-score as our performance metric. Since there is a class imbalance in our data, we compute the macro F1-score: it calculates the F1-score for each class (Misogynistic and Non Misogynistic) and takes the mean of these scores to not let the imbalance influence the evaluation and to treat both classes equally. Using this method gives a more even gauge of performance across categories.

These results indicate that the model was able to distinguish between the two categories successfully, being able to detect Misogynistic memes with high recall but accuracy in Non-Misogynistic recognition.

5.1 AUC and ROC Curve

The ROC curve entails plotting the True Positive Rate (TPR, alternatively known as recall) against the False Positive Rate (FPR) at varying thresholds for classification. This approach is a visual way of examining

Labels	Accuracy	F1-Score	Precision	Recall
Non-Misogynistic	0.78	0.87	0.80	0.95
Misogynistic	0.78	0.44	0.71	0.32
Macro average	0.78	0.66	0.75	0.64
Weighted average	0.78	0.76	0.78	0.79

Table 2: Tamil memes misogynistic content classification report

Labels	Accuracy	F1-Score	Precision	Recall
Non-Misogynistic	0.73	0.78	0.77	0.78
Misogynistic	0.73	0.65	0.66	0.63
Macro average	0.78	0.66	0.75	0.64
Weighted average	0.73	0.72	0.72	0.72

Table 3: Tamil memes misogynistic content classification report

the trade-offs between sensitivity and specificity for a model. A model AUC (goods under the curve) of 0.88 indicates a pretty good performance, meaning it is able to discriminate between Misogynistic memes and Non-Misogynistic memes very efficiently. The closer the AUC is to 1, the model has a higher discriminative power, which means it accurately classifies positive and negative instances. A higher AUC denotes better model performance, while an AUC of 0.5 simply means a random guess. By looking at the ROC curve, we will gain more knowledge about how different thresholds could affect classification performance, thereby ensuring better threshold selection. This curve will allow one to balance between detecting misogynistic memes and lowering false positives.

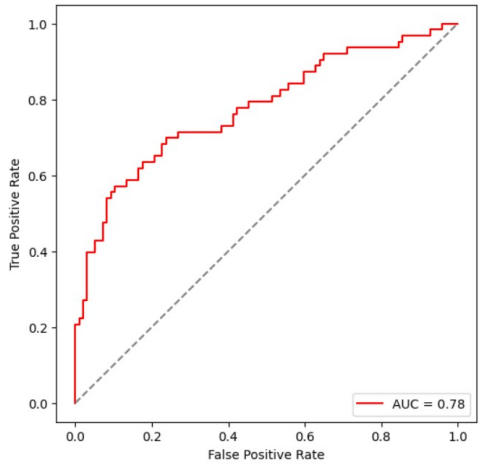

Figure 4: AUC and ROC curve for Malayalam memes, indicating the model's classification performance.

5.2 Confusion Matrix

Our misogyny meme classification model is tested on the basis of the Confusion Matrix, which shows the count of True Positives (TP), True Neg- atives (TN), False Positives (FP), and False Nega- tives (FN). These are the counts needed to calcu- late significant performance metrics such as ac- curacy, precision, recall, and F1 score. These metrics help in identifying how effectively the model distinguishes between misogynistic and non- misogynistic memes, giving an accurate view of its classification performance.

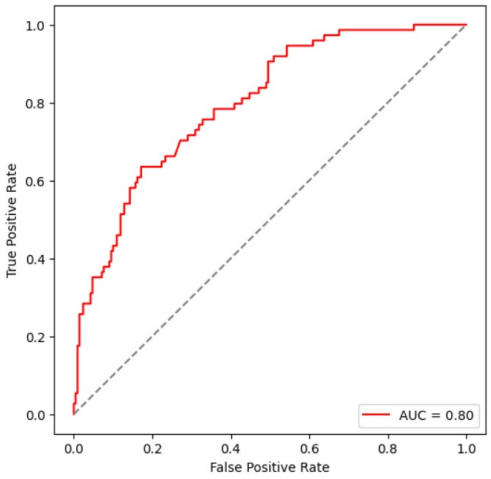

Figure 5: Measures the model's performance in distinguishing misogynistic content in Tamil memes.

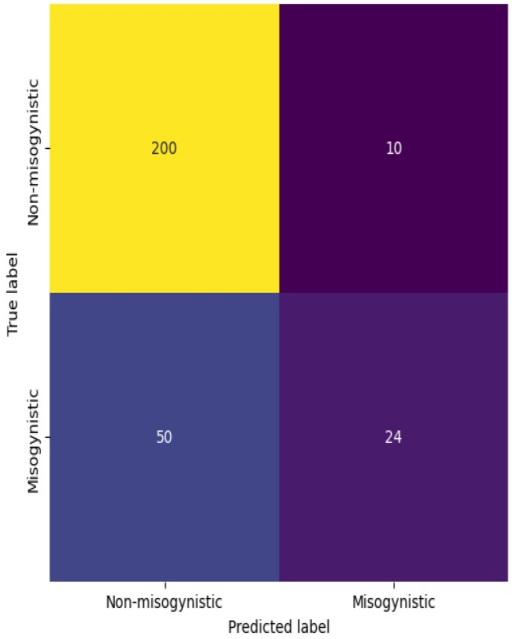

Figure 6: Displays correctly and incorrectly classified misogynistic and non-misogynistic memes in Tamil.

6 Future work

We plan to broaden the scope of our misogyny meme-detecting system to improve its performance and flexibility with different languages and differ- ent types of memes in future work. The idea can be applied to multimodal analysis, where text- and image-based content within a meme for classifica- tion are taken into account. This can be done by us- ing advanced transformer-based models such as Vi- sualBERT or CLIP, which is quite

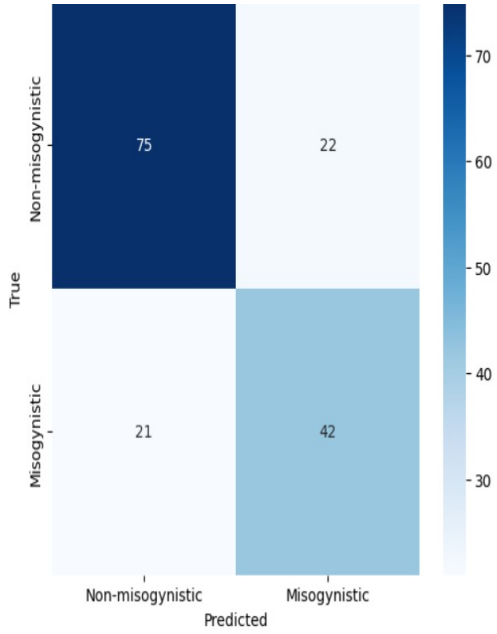

Figure 7: Displays correctly and incorrectly classified misogynistic and non-misogynistic memes in Tamil.

promising in han- dling multimodal data. We would like to improve the model by using more diverse datasets and fine- tune the model with domain-specific datasets to bet- ter capture regional flavor in languages like Tamil and Malayalam. An another way for the improve- ment is by using active learning approaches that reduce manual effort in interpreting new memes so that the system can continue to improve as more data is present. Finally, investigating real-time meme detection could open up possibilities for the practical application of the system, For example, in social media moderation tools that could help fight gender-based abuse online more effectively.

7 Conclusion

This study addresses the very important issue of detection of misogynistic content in memes, especially in two low resource languages, Malayalam and Tamil which are generally neglected by the present-day content moderation systems. Using a combination of SVM classification, text preparation and data EDA the method successfully detects dangerous language hidden in the meme content. It shows the problems and intricacies of addressing abusive content in linguistically and culturally diverse environments. This work aids in the identification of gender-based discrimination and emphasizes the demand for continued re- search and development of more precise and context aware meme categorization models. Additional datasets, language-related components, and multi-modal analysis.

8 Limitations

While advances have occurred, several limita- tions persist. Low-resource systems face chal- lenges when adapting themselves to multiple low- resource languages due to the lack of large anno- tated datasets and linguistic variance in classifica- tion accuracy. Hence, high computational capabil- ity and resource training are involved in handling multimodal data, raising concerns over scalability. Biases in training data will often produce inconsis- tency in the detection of misogyny in differing cul- tural contexts, restricting the system's generaliza- tion, even when trained on heterogeneous datasets. However, aiding the process with adaptive learning techniques will still require heavy human interven- tion to double-check and amend model predictions, which will be quite tedious. Real-time detection re- sults in latency, which inhibits the efficient process- ing of large amounts of data. Another challenge is that the content is updated frequently with new words, symbols, and hidden connotations, which also need to be updated frequently to maintain their accuracy. For in-text citation, add Wu et al. (2006) after "huge ramifications". Include a few sentences about the possible ramifications of wrongful classi- fication. Wrongly characterizing presumably gen- erous content can jeopardize its trustworthiness. Erroneously flagging harmful content as benign puts an additional dent in the trustworthiness of the system. Any kind of building of a content modera- tion system needs to implicate ethics and privacy directly. It must weigh censorship against free ex- pression properly.

Acknowledgment

We thank DravidianLangTech-2025 at NAACL 2025 shared task organizers for providing data sets and guidance.https://sites.google.com/view/dravidianlangtech-2025/shared-tasks-2025

References

Ponnusamy, Rahul and Kathiravan Pannersel- vam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, An- shid 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 Evalua- tion (LREC-COLING 2024)*, pages 7480-7488.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. *Multi- model Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Textf.*, pages 32-41.

Jonathan Bishop. 2014. Dealing with internet trolling in political online communities: To- wards the this is why we can't have nice things scale *International Journal of E-Politics (IJEP), 5(4):1-20.*

Sajeetha Thavereesan and Sinnathamby Mahe- san. 2019.Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation.*In 2019 14th Confer- ence on Industrial and Information Systems (ICIIS)*, pages 320-325.

Thomas Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *In Proceedings of the International AAAI Con- ference on Web and Social Media (ICWSM).*

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Sajeetha Thavareesan and Sinnathamby Mahasen. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th Inter- national Conference on Industrial and Informa- tion Systems(ICIIS),*, pages 478– 482.

Suryawanshi2021 Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. Association for Computational Linguistics.*

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Ra- jiv Teja Nagipogu, Shachi Dave.In *Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730.*

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE NLP@DravidianLangTech- EACL2021: Offensive language identification 7 on multilingual code-mixing text. In*Proceed- ings of the First Workshop on Speech and Lan- guage Technologies for Dravidian Languages, pages 313–318, Kyiv.*

E. S. Smitha, S. Sendhilkumar, and G. S. Maha- laksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing, Cham. Springer International Publishing*, pages 1015–1031.

Suryawanshi2021 Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A Sentiment Analysis Dataset for code-Mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under- resourced languages (SLTU) and Collabora- tion and Computing for Under-Resourced Lan- guages (CCURL),* pages 177–184, Marseille, France.

Chakravarthi, Bharathi Raja, Ponnnusamy, Rahul and Rajiakodi, Saranya and Muthusamy, Sivagnanam, Bhuvaneswari and Kizhakkeparambil, Anshid. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Preceedings of the Fifth Workshop on Speech, Vision and Language Technologies for Dravidian Languages* Association for Computational Linguistics, 2025.