KEC_AI_VSS_run2@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media

Kogilavani Shanmugavadivel[1],**Malliga Subramanian**[2],
Sathiya Seelan S[1], **Suresh Babu K**[1], **Vasikaran S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{sathiyaseelans.22,sureshbabuk.22,vasikarans.22aid}@kongu.edu

Abstract

The increasing instances of abusive language against women on social media platforms have brought to the fore the need for effective content moderation systems, especially in low-resource languages like Tamil and Malayalam. This paper addresses the challenge of detecting gender-based abuse in YouTube comments using annotated datasets in these languages. Comments are classified into abusive and non-abusive categories. We applied the following machine learning algorithms, namely Random Forest, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting and AdaBoost for classification. Micro F1 score of 0.95 was achieved by SVM for Tamil and 0.72 by Random Forest for Malayalam. Our system participated in the shared task on abusive comment detection,out of 160 teams achieving the rank of 13th for Malayalam and rank 34 for Tamil, and both indicate both the challenges and potential of our approach in low-resource language processing. Our findings have highlighted the significance of tailored approaches to language-specific abuse detection.

1 Introduction

Social media has revolutionized communication, interaction, information sharing, and expression. However, its misuse has led to serious issues, including gender-based abuse targeting women. Such abusive language is mostly derogatory and threatening, which reflects societal biases and has serious psychological, social, and professional consequences for the victims.Tamil and Malayalam are low-resource languages that lack robust automated systems to address this challenge. Identifying abusive content in these languages is important for effective moderation. This paper focuses on the detection of gender-based abuse in Tamil and Malayalam YouTube comments.This can help in creating safer and more inclusive online environments.

Detecting abusive content in low-resource languages is challenging due to the scarcity of annotated datasets and the complexity of linguistic nuances. The implicit bias, coded expressions, and slangs used in abusive language make it difficult for the automated system to detect Tamil and Malayalam. We approach these challenges by curating annotated datasets of YouTube comments in Tamil and Malayalam. Each comment is labeled as either abusive or non-abusive, and examples reflect explicit and implicit abuse. Developing accurate classification models for these datasets is important for enhancing content moderation systems.

We, therefore, use machine learning algorithms such as Random Forest, SVM, KNN, Gradient Boosting, and AdaBoost on the comments dataset to classify them as either abusive or non-abusive. The experiments show that SVM gives the highest micro F1 score at 0.95 for Tamil, but Random Forest gives the best score of 0.72 for Malayalam. These findings reflect the efficiency of language-specific models in detecting abusive content. The empirical findings from this research study will help to solve the imperative case of gender-based abuse issues in social media and make cyberspace safer for women.

2 Literature Survey

The Multimodal Tamil Hate (MATH) dataset has been introduced for detecting hate speech in Tamil across text, audio, and video modalities Mohan et al. (2023). A combination of BERT for text, TimeSformer for video, and Wav2vec2 for audio was used, achieving 81.82% accuracy with a multimodal fusion approach. Tamil abusive comment classification has been improved through multilingual transformers and data augmentation, leading to a 15-unit increase in the macro F1-score using the MURIL model Sheik et al. (2023). Additionally, Tamil and code-mixed Tamil-English datasets for

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 287–291

abusive comment detection on YouTube have been created, showing classical models as more effective due to limited data Chakravarthi et al. (2023).

A transformer-based approach has been proposed for detecting abusive content in 13 Indic code-mixed languages, outperforming classical models. The combination of XLM-RoBERTa with BiGRU and emoji embeddings achieved an F1 score of 0.88 and an AUC of 0.94 Bansal et al. (2022). Multilingual embeddings like IndicBERT and MuRIL have been utilized for Tamil and Telugu, demonstrating superior performance over classical models on YouTube datasets Vegupatti et al. (2023).

A toxic comment detection system for Assamese has been developed using SVM with TF-IDF, achieving 94% accuracy and F1-score Dutta et al. (2024). An overview of a shared task on detecting abusive and hope speech in Tamil and Tamil-English social media comments has been provided, employing various machine learning and deep learning methods Priyadharshini et al. (2022).

Studies from the Third Workshop on Speech and Language Technologies for Dravidian Languages have presented insights into abusive comment detection Priyadharshini et al.. Logistic regression with embeddings has been explored for Tamil and Telugu datasets Bala and Krishnamurthy (2023). A study has achieved 99% accuracy in abusive comment detection for code-mixed Tamil-English text Pannerselvam et al. (2023). Machine learning, deep learning, and BERT have been employed for Tamil, achieving notable rankings in a shared task Shanmugavadivel et al. (2023).

3 Task Description

This task concentrated on finding abusive comments toward women in the language of Tamil and Malayalam from social media sites, such as YouTube. We annotated the data carefully with proper labels in both abusive and non-abusive categories to have a high-quality set of labels for supervised learning. We utilized several machine learning models, namely SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost to classify comments against this challenge. SVM achieved the highest micro F1 score of 0.95 for Tamil, while Random Forest performed best for Malayalam with a score of 0.72. The models were trained and fine-tuned using various feature extraction techniques, including TF-IDF and word embeddings, to capture linguistic nuances.Our system participated in the shared task Rajiakodi et al. (2025) on abusive comment detection, ranking 13th for Malayalam and 34th for Tamil out of 160 teams. This therefore shows the problems of abusive language detection in low-resource languages and the need to develop robust language-specific content moderation systems.

4 Dataset Description

The Tamil dataset has a total of 2,790 records split between the classes evenly. Therefore, it has 1,366 abusive comments and 1,402 non-abusive comments, which makes the dataset balanced for unbiased training and model testing. A balanced dataset means that none of the models gets biased to prefer any of the classes, one being abusive and the other being not.

The Malayalam dataset has 2,933 records with a little imbalance in its distribution. This dataset contains 1,531 abusive comments and 1,402 non-abusive comments; this is true to real life, where most of the instances of abusive language are dominant in certain contexts.This dataset is a bit more challenging as it also is imbalanced and includes the complexity of the Malayalam language.

Language	Abusive(N)	Non Abusive
Malayalam	1531	1402
Tamil	1366	1424

Table 1: Dataset Description

5 Methodology

The methodology for abusive comment detection in Tamil and Malayalam consists of data preprocessing, feature extraction, and model development. Each step ensures efficient processing by transforming raw data into meaningful representations using machine learning-based classification.

5.1 Data Preprocessing

Data preprocessing involves cleaning raw YouTube comments for classification. This includes removing special characters, emojis, punctuation, and URLs. Tokenization splits each comment into words or phrases for analysis, followed by stopword removal to enhance computational efficiency. Next, text normalization standardizes spelling and slang variations in Tamil and Malayalam. Labels are encoded into numbers for machine learning

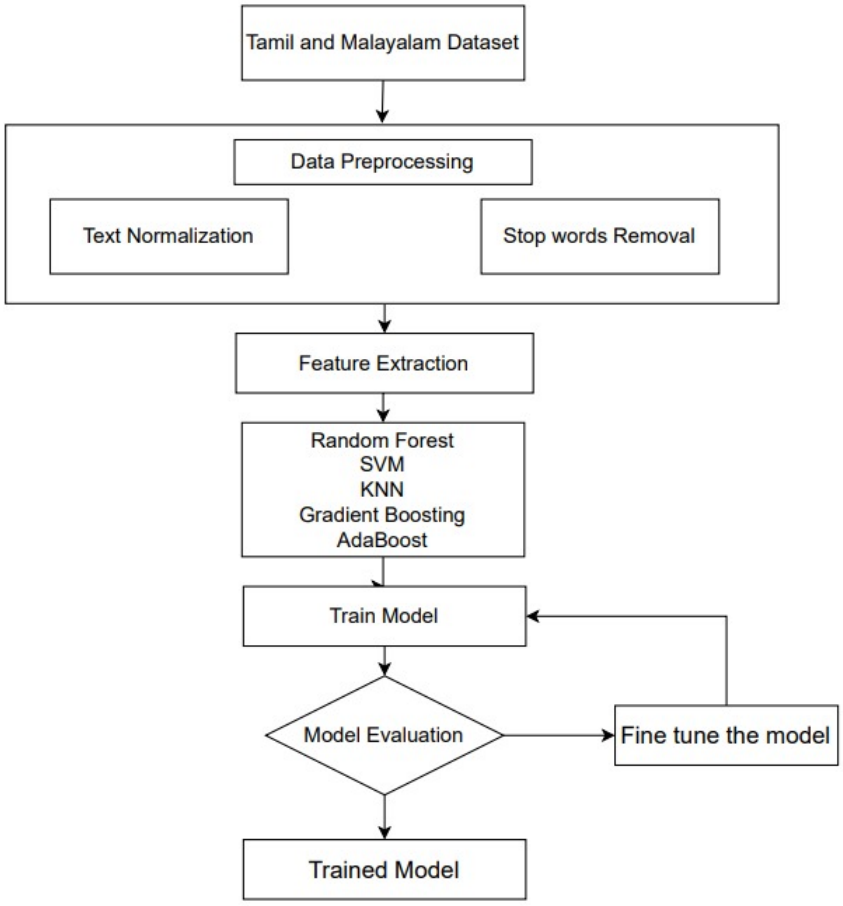

Figure 1: Proposed System Workflow

compatibility. To address class imbalance, methods like over-sampling the majority class and under-sampling the minority class are applied, ensuring better model performance by preventing bias toward one class.

5.2 Feature Extraction

After preprocessing, features are extracted using methods like TF-IDF (Term Frequency-Inverse Document Frequency), which weighs words based on their frequency in abusive comments while down-weighting common terms. Additionally, word embeddings like Word2Vec, FastText, and BERT capture contextual meaning, enabling the model to understand relationships between words. These features are crucial for enhancing model precision in distinguishing abusive from non-abusive language.

5.3 Model Development

The final step involves selecting, training, and testing machine learning models for abusive comment classification. The dataset is split into training, validation, and test sets for effective generalization. Several algorithms, including SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost, are fine-tuned for optimal performance. Evaluation metrics such as accuracy, precision, recall, and micro F1 score are calculated. SVM performs best for Tamil with a micro F1 score of 0.95, while Random Forest achieves the highest micro F1 score of 0.72 for Malayalam, indicating its effectiveness in handling linguistic variations in abusive comments. The workflow diagram in Figure 1 shows the entire process, from preprocessing to model evaluation.

6 Experimental Analysis

To check the performance of the models, Macro-F1 score is employed that is widely used in classification problems, especially on imbalanced datasets. The experiments are carried out on text data for both the Tamil and Malayalam languages. Accordingly, the corresponding performances are explained in the subsections below.

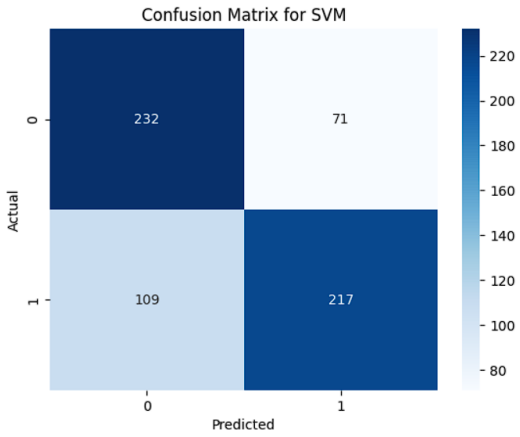

Figure 2: Confusion Matrix of Tamil Data

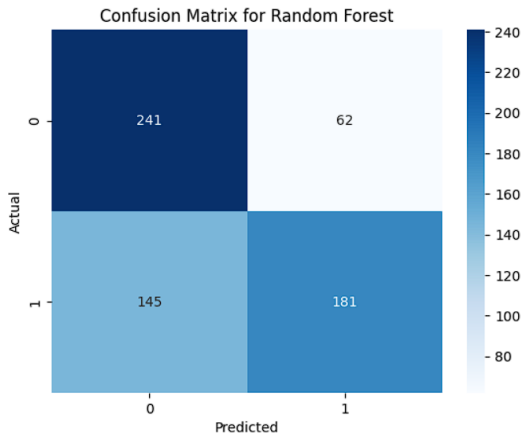

Figure 3: Confusion Matrix of Malayalam Data

6.1 Tamil

For this, the model performance was analyzed with accuracy, precision, recall, and micro F1 score metrics to compare their ability in classifying the comments as abusive or non-abusive. For the Tamil dataset, Support Vector Machine has been found out to be performing better than others with a micro F1 score of 0.95, which is really a high score. This is due to the reason that SVM handles the linguistic variations in Tamil such as colloquial terms and context-dependent variations. Figure 2 Depicts the best-performing model's confusion matrix on Tamil data. The best model was able to strongly capture the explicit and implicit abusive language patterns and, therefore could be relied on for Tamil text classification.

6.2 Malayalam

For the Malayalam dataset, Random Forest was the best model with a micro F1 score of 0.72. Although the score is lower than that of Tamil, it shows that Random Forest generalizes well despite the complexity of Malayalam grammar and vocabulary. The model was able to capture patterns of abuse, including slang and implicit bias in Malayalam comments. This performance depicts robustness and applicability for moderate abusive content in the Malayalam language effectively. Figure 3: Confusion matrix of the best performing model on Malayalam data.

7 Limitations

This study has several limitations. First, the dataset size is limited, which may affect the generalizability of the models. Second, the imbalance in the Malayalam dataset could lead to biased predictions. Third, the approach relies on traditional machine learning models, which might not capture complex linguistic patterns effectively. Lastly, the absence of deep learning techniques limits the potential for higher accuracy.

8 Conclusion

The research throws light on the increasing phenomenon of gender-based abuse on social media and the growing need for automatic content moderation, especially in low-resource languages such as Tamil and Malayalam. A classification system has been developed for effectively detecting abusive comments by applying SVM, Random Forest, KNN, Gradient Boosting, and AdaBoost machine learning models. Our results indicate that SVM obtained the best micro F1 score of 0.95 for Tamil, while the best-performing model for Malayalam was Random Forest with a score of 0.72. Our system has also participated in the shared task on abusive comment detection and ranked 13th in Malayalam and 34th in Tamil out of 160 teams. Deep models and contextualized embeddings are much more beneficial for the advancement of this problem, and future work will be directed towards enlarging datasets and fine-tuning classification techniques to make them more accurate and generalizable for real-world applications. The code for this shared task can be accessed at Github

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@ DravidianLangTech: Abusive comment detection in Tamil and Telugu using logistic regression. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian*

Languages, pages 231–234, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Vibhuti Bansal, Mrinal Tyagi, Rajesh Sharma, Vedika Gupta, and Qin Xin. 2022. A transformer based approach for abuse detection in code mixed indic languages. *ACM transactions on Asian and low-resource language information processing*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Surajit Dutta, Mandira Neog, and Nomi Baruah. 2024. Assamese toxic comment detection on social media using machine learning methods. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–8. IEEE.

Jayanth Mohan, Spandana Reddy Mekapati, and Bharathi Raja Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Kathiravan Pannerselvam, Saranya Rajiakodi, Rahul Ponnusamy, and Sajeetha Thavareesan. 2023. CSS-CUTN@DravidianLangTech:abusive comments detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 306–312, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Shri Durga R, Srigha S, Sree Harene J S, and Yasvanth Bala P. 2023. KEC_AI_NLP@DravidianLangTech: Abusive comment detection in Tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Reshma Sheik, Raghavan Balanathan, and Jaya Nirmala S. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465, Goa University, Goa, India. NLP Association of India (NLPAI).

Mani Vegupatti, Prasanna Kumar Kumaresan, Swetha Valli, Kishore Kumar Ponnusamy, Ruba Priyadharshini, and Sajeetha Thavaresan. 2023. Abusive social media comments detection for tamil and telugu. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 174–187. Springer.