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

Malliga Subramanian[1], Kogilavani Shanmugavadivel[2], Indhuja V S[1], Kowshik P[1], Jayasurya S[1]
[1]Department of CSE, Kongu Engineering College, Perundurai, Erode.
[2]Department of AI, Kongu Engineering College, Perundurai, Erode.
{mallinishanth72, kogilavani.sv}@gmail.com
{indhujavs.23cse, kowshikp.23cse}@kongu.edu
jayasuryas.23cse@kongu.edu

Abstract

The detection of abusive text targeting women, especially in Dravidian languages like Tamil and Malayalam, presents a unique challenge due to linguistic complexities and code-mixing on social media. This paper evaluates machine learning models such as Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest Classifiers (RFC) for identifying abusive content. Code-mixed datasets sourced from platforms like YouTube are used to train and test the models. Performance is evaluated using accuracy, precision, recall, and F1-score metrics. Our findings show that SVM outperforms the other classifiers in accuracy and recall. However, challenges persist in detecting implicit abuse and addressing informal, culturally nuanced language. Future work will explore transformer-based models like BERT for better context understanding, along with data augmentation techniques to enhance model performance. Additionally, efforts will focus on expanding labeled datasets to improve abuse detection in these low-resource languages.

1 Introduction

The rise of social media has led to an increase in online abuse, particularly gender-based harassment targeting women. Detecting such abusive content in languages like Tamil and Malayalam presents unique challenges for Natural Language Processing (NLP). Despite growing concerns, there is limited research on abusive language detection in Tamil and Malayalam languages. This study explores the effectiveness of machine learning models, including Support Vector Machines (SVM), Logistic Regression, and Random Forest, for identifying abusive content in Tamil and Malayalam social media posts.This paper contributes to advancing content moderation techniques for multilingual social media platforms.

2 Literature Survey

Recent studies on abusive language detection have predominantly concentrated on English, yielding promising results with advanced machine learning models. However, research on low-resource languages, particularly Dravidian languages such as Tamil and Malayalam, remains limited. These languages often feature code-mixing, informal expressions, and context-dependent nuances, posing unique challenges for accurate detection Priyadharshini et al., 2022. The survey provides an overview of models submitted for abusive text identification in DravidianLangTech@NAACL 2025. Additionally, the absence of large annotated datasets and the difficulty of detecting implicit and subtle abuse further complicate the task Chen et al., 2018. Future research should focus on developing robust, language-specific models and expanding annotated datasets to enhance detection accuracy.

2.1 Abusive Detection in English and Major Language

Early abusive content detection relied on blacklists and regular expressions but struggled with subtle expressions, sarcasm, and context-dependent abuse Jiangbin et al., 2021. Machine learning models improved detection using features like n-grams and sentiment analysis Akhter et al., 2022, yet they struggled with nuanced language. Transformer-based models like BERT, RoBERTa, and ALBERT enhanced accuracy by capturing context through self-attention mechanisms.

2.2 Abusive Detection in Dravidian Languages

Abusive language detection in Tamil and Malayalam, especially gender-targeted abuse on social media, remains under-explored despite its importance. These Dravidian languages pose challenges due to complex syntax, rich morphology, and

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 265–268

limited annotated datasets. Traditional models like SVM and Random Forest struggle with implicit, code-mixed, and culturally nuanced abuse Eshan and Hasan, 2017. While deep learning and transformer models show promise, they still face difficulties in capturing subtle abuse. There is a pressing need for culturally aware models and expanded datasets to improve detection in these languages.

2.3 Deep Learning and Transformers

Recent advancements in deep learning, especially with transformer-based models like BERT, RoBERTa, and ALBERT, have significantly enhanced the accuracy of abusive language detection in social media Rajiakodi et al., 2025. These models are particularly powerful in handling the complexity of informal, code-mixed, and context-dependent expressions in Tamil and Malayalam. Unlike traditional methods, which rely on handcrafted features, transformers utilize self-attention mechanisms to understand the relationship between words in a sentence. This enables them to capture subtle forms of abuse such as implicit and gender-targeted language. However, the lack of large annotated datasets in these languages remains a challenge. Fine-tuning transformer models on Tamil and Malayalam data is crucial to improving performance Priyadharshini et al., 2023.

3 Materials and Methods

3.1 Taskset Description

This study focuses on detecting abusive language in Tamil-English and Malayalam text, particularly targeting women, sourced from social media platforms such as YouTube. The dataset was collected by scraping YouTube comments using specific queries related to controversial and sensitive topics, ensuring a diverse representation of abusive and non-abusive content. Figures 1 and 2 provide sample datasets for Tamil and Malayalam, respectively. The dataset consists of 5,723 texts in the training set, 1,229 texts in the development set, and 1,227 texts in the test set, each labeled as "Abusive" or "Non-Abusive." These texts exhibit significant linguistic challenges, including code-mixing, informal expressions, and culturally nuanced content, which are common in Dravidian languages. The data was processed and annotated to facilitate robust classification, leveraging machine learning techniques. By employing models like SVM, Random Forest, and Logistic Regression, this study evaluates classification performance across diverse linguistic patterns.

Text	Class
You tube ல இப்படி எல்லாம் முன்னேற்றம் நம் சேனலையும வந்து முன்னேற்றுங்கள்!!	Non-Abusive
ஆமா பா கட்டி வச்சி தோல உரிக்கணும	Abusive
சிறியவர்கள், இதயநோயாளிகள் , கர்ப்பிணி பெண்கள் யாரும் இதை பார்க்க வேண்டாம்	Non-Abusive
உன் பல்ல பாத்தாலே பயமா இருக்கு. ஆளும் பல்லும் உருவமும் பேச்சும் பாரு. இதுல 2 kg டூ வேற	Abusive
இவயா இதற்கெல்லாம் அர்த்தங்கள் என்ன கேக்க வேண்டும் உன்னை காலம் கை கூடினால்.....	Non-Abusive

Figure 1: Sample training texts from Tamil dataset

Text	Class
തൊലിക്കട്ടി ഉണ്ടോ എന്നാൽ എന്ത് പിഴവ് വേണമെങ്കിലും ഉണ്ടാക്കാൻ പറ്റും	Abusive
ആ കുലയുടെ രണ്ട് വിത്ത് കിട്ടുമോ ?	Abusive
അടി തെറ്റിയാൽ .... ജിമിക്കിയും കമ്മലും വിഴും	Non-Abusive
"എന്ത് നാണക്കേട്.ആല് കിളിത്തലും അത് തണലായി കരുതുന്ന ടീംസ് ആണ്."	Abusive
"നോട്ടപ്പിശക് പോലും കേൾക്കുമ്പോൾ കലി വരുന്നു. ഒടുക്കത്തെ നോട്ടപ്പിശക്"	Non-Abusive

Figure 2: Sample training texts from Malayalam dataset

3.2 Preprocessing and Feature Extraction

Preprocessing plays a crucial role in transforming raw text data into a format suitable for machine learning. The dataset was first cleaned by handling missing values and removing incomplete rows. The labels for "Abusive" and "Non-Abusive" content were mapped to binary values (1 and 0) for binary classification. For feature extraction, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was applied, converting text into numerical features while capturing the significance of words. Stop words were eliminated to remove common, non-informative words that would contribute little to classification. This process reduced noise and enhanced the model's ability to focus on relevant terms Sai and Sharma, 2021. Additionally, the text data was standardized and formatted to ensure uniformity. While the current approach focuses on TF-IDF-based vectorization, future improvements could integrate transliteration normalization through mapping dictionaries and custom tokenization strategies to better handle transliterated words and mixed scripts. These preprocessing steps ensured that the models received high-quality input, enabling efficient training and accurate predictions for abusive language detection in social media.

3.3 Models and Methodology

This study explores the use of machine learning models to detect abusive language in Tamil and Malayalam text. We employed three models: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest Classifier (RFC). SVM, with its ability to find optimal decision

boundaries, was used for precise classification. Logistic Regression, a linear model, was utilized for its interpretability and effectiveness in binary classification tasks. Random Forest, an ensemble method, leveraged multiple decision trees to enhance predictive performance Mahmud et al., 2024. Text data was preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, which converts text into numerical features. The TF-IDF vectorizer was set to a maximum of 5000 features, with stop words removal enabled. The selected hyperparameters included a linear kernel for SVM, 100 estimators for Random Forest, and a random state of 42. The models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

4 Results and Discussion

The study on abusive content detection in Tamil and Malayalam revealed that while traditional machine learning models like Logistic Regression and Random Forest performed effectively, Support Vector Machine (SVM) demonstrated superior performance, making it the most suitable model for this task.

4.1 Performance Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures overall correctness, while Precision indicates correctly predicted abusive texts. Recall reflects the proportion of actual abusive texts identified, and F1-Score balances Precision and Recall, crucial for imbalanced datasets. These metrics help assess real-world effectiveness and optimize abuse detection systems. Given linguistic complexity of Tamil and Malayalam, they provide insights into model adaptability across diverse text patterns. Table 1 illustrates classification performance for Tamil, while Table 2 presents results for Malayalam, ensuring robust and reliable abuse detection models.

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	Non-Abusive	68	68	70	69
Logistic Regression	Abusive	68	70	68	69
Random Forest	Non-Abusive	68	67	69	68
Random Forest	Abusive	68	69	67	68
SVM	Non-Abusive	69	69	67	68
SVM	Abusive	69	69	71	70

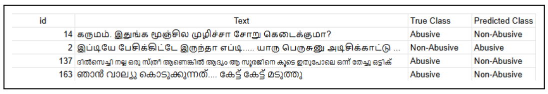
Table 1: Performance of Classifiers for Abusive and Non-Abusive Text Detection in Tamil

Classifiers	Class Labels	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	Non-Abusive	64	63	67	65
Logistic Regression	Abusive	64	66	62	64
Random Forest	Non-Abusive	61	59	67	63
Random Forest	Abusive	61	64	56	60
SVM	Non-Abusive	64	63	67	65
SVM	Abusive	64	67	62	64

Table 2: Performance of Classifiers for Abusive and Non-Abusive Text Detection in Malayalam

4.2 Error Analysis

The SVM model performed well but struggled with indirect and implicit abusive language, often misclassifying it as non-abusive due to limited context understanding. It also faced challenges with Tamil-Malayalam code-mixed sentences, as TF-IDF failed to capture nuanced language patterns, leading to false negatives. Figure 5 provides an example of such misclassification. These limitations highlight the need for more context-aware models like BERT. Future work will focus on expanding the dataset, improving tokenization, and addressing ethical considerations for unbiased predictions.



id	Text	True Class	Predicted Class
14	கருமம். இதுங்க மூஞ்சில முழிச்சா சோறு கெடைக்குமா?	Abusive	Non-Abusive
2	இப்டியே பேசிக்கிட்டே இருந்தா எப்டி..... யாரு பெருசுனு அடிசிக்காட்டு ...	Non-Abusive	Abusive
137	ദിൽസെച്ചി നല്ല ഒരു സ്ത്രീ ആണെങ്കിൽ ആദ്യം ആ സൂരജിനെ കൂടെ ഇതുപോലെ ഒന്ന് തേച്ചു ഒട്ടിക്	Abusive	Non-Abusive
163	ഞാൻ വാല്യൂ കൊടുക്കുന്നത്.... കേട്ട് കേട്ട് മടുത്തു	Abusive	Non-Abusive

Figure 5: Example of a misclassified Tamil-Malayalam code-mixed text by the SVM model.

5 Limitations

Abusive language detection in Tamil and Malayalam using traditional machine learning models such as Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) encounters several hurdles. The scarcity of labeled datasets in these languages limits the models' effectiveness. Additionally, the intricate grammatical structures and diverse vocabulary present in Tamil and Malayalam complicate the accurate identification of offensive content. Existing models often rely on simple feature extraction techniques, which struggle to grasp the subtleties of code-mixing and contextual variations in online discourse. Lastly, recognizing gender-based abuse requires a deeper understanding of cultural context, which current models fail to fully address.

6 Conclusion

In this study, we explored the detection of abusive content targeting women in Tamil and

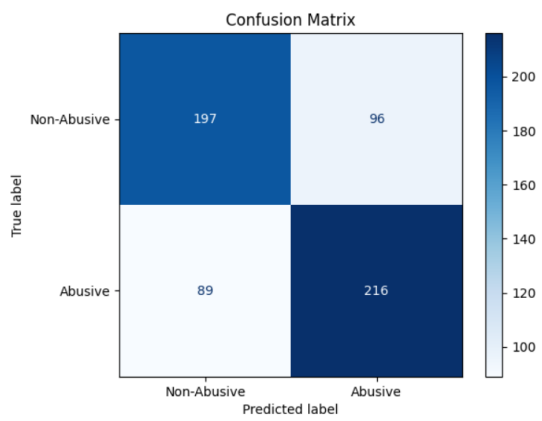


Figure 3: Confusion matrix for the Tamil dataset

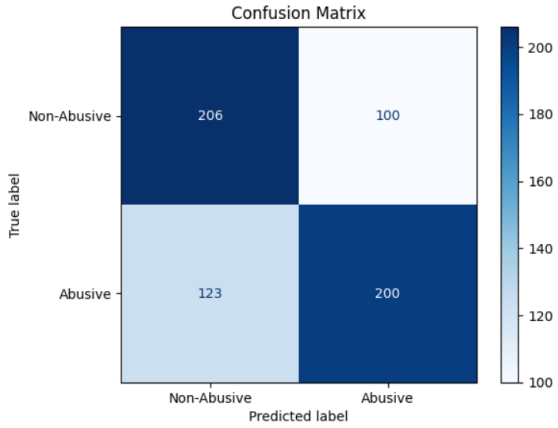


Figure 4: Confusion matrix for the Malayalam dataset

Malayalam social media posts using machine learning. Support Vector Machine (SVM) demonstrated superior performance in handling informal, context-dependent, and code-mixed language. The findings highlight challenges in detecting abuse in morphologically rich, low-resource languages and the need for scalable solutions. While results are promising, future research should focus on larger annotated datasets and hybrid models combining traditional methods with transformer-based architectures. The dataset and implementation code utilized in this study are publicly available at GitHub Repository to support reproducibility and further research.

References

M.P. Akhter, Z. Jiangbin, I.R. Naqvi, and et al. 2022. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, 28:1925–1940.

H. Chen, S. McKeever, and S.J. Delany. 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In *Social Informatics. SocInfo 2018*, volume 11185, pages 1–10. Springer, Cham.

S. C. Eshan and M. S. Hasan. 2017. An application of machine learning to detect abusive bengali text. In *Proceedings of the 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6, Dhaka, Bangladesh.

Zheng Jiangbin, Syed Irfan Naqvi, Mohammed Abdelmajeed, and Tehseen Zia. 2021. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, 28.

T. Mahmud, T. Akter, M. K. Uddin, M. T. Aziz, M. S. Hossain, and K. Andersson. 2024. Machine learning techniques for identifying child abusive texts in online platforms. In *Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, Kamand, India.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*. Recent Advances in Natural Language Processing.

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.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv. Association for Computational Linguistics.