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

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

Abstract

Social networks are becoming crucial sites for communication and interaction, yet are increasingly being utilized to commit gender-based abuse, with horrific, harassing, and degrading comments targeted at women. This paper tries to solve the common issue of women being subjected to abusive language in two South Indian languages, Malayalam and Tamil. To find explicit abuse, implicit bias, preconceptions, and coded language, we were given a set of YouTube comments labeled Abusive and Non-Abusive. To solve this problem, we applied and compared different machine learning models, i.e., Support Vector Machines (SVM), Logistic Regression (LR) and Naive Bayes classifiers, to classify comments into the given categories. The models were trained and validated using the given dataset, achieving the best performance with an accuracy of 89.89% and a macro F1 score of 90% using the best-performing model. The proposed solutions aim to develop robust content moderation systems that can detect and prevent abusive language, ensuring safer online environments for women.

1 Introduction

Over the years, social networks have become an overwhelmingly popular channel for entertainment, communication, and distribution of information. But despite this advantage, it has also become a platform in which cyberbullying and harassment occur predominantly. Cyberbullying occurs in a major way among women, a reflection of deep-seated cultural prejudice or gender inequality, and it also often manifests itself in the form of nasty, vilifying, and threatening speech. Given the strong psychological, social, and professional consequences of this type of focused harassment, creating strong protections against such speech is absolutely necessary.

Malayalam and Tamil are two prominent languages used on social media platforms in South India. However, the resource-scarce nature adds to the challenges of effective content-moderation tools in these two languages. Inappropriate comments with low-resource languages usually include explicit language, implicit bias, stereotypes, and coded language that makes them more difficult to spot.

This research aligns with the shared task on the detection of abusive comments in Tamil and Telugu proposed by Priyadharshini et.al (2023). Their analysis provides a benchmark dataset and evaluations used to contribute to the advancement of abusive comment detection. Also, Priyadharshini et.al, in the DravidianLangTech@ACL (2022) workshop, discussed the impact of abusive language on social media and highlighted the challenges posed by code-mixed Tamil and English text. By incorporating these insights, our study contributes to ongoing efforts in low-resource language processing. It improves the accuracy of abuse detection systems and reinforces the need for multilingual AI-driven moderation tools.

The present research will identify gender-related abusive content in comments posted on the Malayalam and Tamil YouTube streams, with a focus on solving the concern. The goals of this project are to implement machine learning algorithms to classify comment categories as abusive and non-abusive using datasets that have received binary labels applied. The current data set used contains diverse abusive content, both explicit and implicit.We have used the support vector machine (SVM), logistic regression (LR) and Naive Bayes machine learning models to perform the classification task. For implementation, please refer to this GitHub repository(Dhanyashree-G).

2 Related Work

The Abusive Comment Detection in Tamil-ACL 2022 shared task consisted of an experiment by Balouchzahi et al.(2022) on detecting abusive comments in Tamil. To address challenges such as code mixing, context dependence, and data imbalance, their experiment considered abusive language in native Tamil script, as well as code-mixed Tamil texts. They proposed two models for the task: (i) a 1D Convolutional LSTM (1D Conv-LSTM) model and (ii) an n-gram Multilayer Perceptron model (n-gram MLP) utilizing char n-grams and performed well for Tamil mixed code with a weighted F1 score of 0.56. For detecting abusive content, the n-gram MLP model outperformed the 1D Conv-LSTM model. This paper illustrates how feature engineering and classical machine learning can be used to detect abusive content in low-resource, code-mixed languages.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 371–376

The shared task of Chakravarthi et al. was discussed in his presentation shortly after the third publication. The (2021) project of offensive language identification in Tamil, Malayalam, and Kannada languages was conducted through the (2021), addressing the challenges of detecting abuse in under-resourced Dravidian languages. This task emphasized the importance of identifying offensive language in multilingual and code-mixed texts prevalent in user-generated content on social media platforms. The dataset for this task included six categories of annotations, such as Not offensive, offensive untargeted, and offensive. Participants used a wide variety of methodologies, including traditional machine learning algorithms, deep learning architectures, and transformer-based models. Pre-trained multilingual transformers such as mBERT, XLM-R, and IndicBERT have been carefully evaluated to classify offensive content. The model that performs best have achieved F1 scores of up to 0.97% for Malayalam and 0.78% for Tamil, highlighting the potential utility of transformer-based models in offensive language detection.

Rajalakshmi et al. solved the problem of detecting abusive comments in Tamil and Tamil-English datasets under the shared task DravidianLangTech@ACL 2022. The primary goal of the study was to detect abusive content categories such as homophobia, transphobia, xenophobia, and counter-speech that often form within the community. Three approaches were used by the authors: transformer-based models, deep learning (DL), and machine learning. Random Forest outperformed other algorithms with a weighted F1 score of 0.78% on its Tamil and English dataset. Pre-trained word embeddings with BiLSTM models performed better among deep learning models for Tamil data. mBERT was the best-performing transformer-based model with an F1 score of 0.70% for Tamil comments. Issues such as class imbalance and the dominance of code-mixed and code-switched data that make detection tasks more difficult were also addressed in the study. The authors published a paper using advanced techniques such as balanced class weights and fine-tuning transformer models to identify abusive content.

Pannerselvam et al. (2023) addressed the issue of identifying offensive remarks in code-mixed Tamil-English and Telugu-English text. They concentrated on developing a multiclass classification model that could distinguish between eight types of offensive remarks. The study used the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and two text representation techniques, Bag of Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF), to solve the problems caused by code-mixed data. Machine learning algorithms such as Support Vector Machine (SVM), Random Forest, and Logistic Regression were used to perform the categorization. It performed best among them with an F1 score of 0.99% TF-IDF representation and SMOTE-balanced data to achieve the highest performance. The study shows how mixing SMOTE and TF-IDF works well to handle unbalanced datasets and catch the subtle differences in mean speech across languages. Their method proved strong and looked good for real-world use in managing angry comments on online platforms. This was clear from how they came in ninth place in the shared task, even when dealing with issues like language gaps or changes in what people think is offensive.

The author, Zichao Li, has combined classes, adjusted course weights with respect to the reciprocal of log frequencies, and used focal loss to put more focus on the minority classes during training to tackle challenges such as class imbalance in the dataset. We applied additional adversarial training to improve the robustness and generalization ability of the model. This resulted in one of the top-performing systems with a weighted average F1 score of 0.75%, 0.94%, and 0.72%, individually placing it at fourth, third, and fourth in Tamil-English, Malayalam-English, and Kannada-English tasks, respectively. This work emphasizes the usefulness of transformer approaches for dealing with code-mixed texts in low-resourced languages such as Tamil, Malayalam, and Kannada through novel use of multilingual transformers, applicable preprocessing methods, and specialized loss functions.

3 Methodology

The dataset and the experiments we carried out for the study are described in depth in this section. The system architecture for classifying abusive comments into binary classes using machine learning (ML) methods, such as GridSearchCV, The general flow of the categorization process is shown in the figure 1.

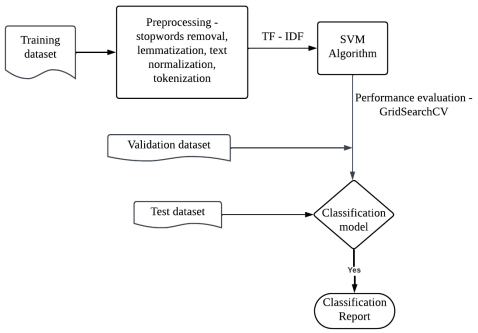

Figure 1: System Architecture for Detecting Abusive Comments Using ML Models.

3.1 Dataset

The dataset for this study consists of YouTube comments written in Tamil and Malayalam. Each language has its datasets divided into three subsets: train, validate, and test. The data sets are divided into two classes: Abusive and Non-Abusive. The detailed distribution of each dataset is showed in Table 1.

Language	Train	Validate	Test
Malayalam	2933	629	500
Tamil	2790	598	450

Table 1: The dataset distribution for Tamil and Malayalam, including the number of samples for each language.

The datasets were adapted and modified from publicly accessible datasets originally published as part of the DravidianLangTech@NAACL2025 program to suit the specific context of this study.

3.2 Proposed Solution

The proposed solution for detecting abusive and non-abusive comments in Tamil and Malayalam employs a combination of preprocessing techniques, feature engineering, and machine learning models to achieve precise and interpretable classification. Similarly, exploratory data analysis was integrated using WordCloud to visualize the most common abusive and non-abusive terms in the dataset, providing insight into language patterns.

3.3 Exploratory Data Analysis

Unbiased comments were generated in WordCloud for abusive and non-abusive comments compared to standard datasets. These visualizations highlighted frequently used terms in each category, allowing better interpretation of patterns in abusive language specific to Tamil and Malayalam.

3.4 Preprocessing

The text becomes more uniform after being converted into lowercase, having all the punctuation signs stripped off, and numbers excluded. Words were rearranged to cut them down to their root forms or lemmatized but with meaning in various forms of the word.

To convert text to numerical features, we employed vectorization using Term Frequency-Inverse Document Frequency (TF-IDF), which is a method that analyzes the relevance of a certain word within a certain document in relation to the entire data collection available. Term Frequency (TF) is the frequency of how often a word is found in a document, and Inverse Document Frequency (IDF) scales down the value of frequently occurring words so that highly used but less informative words do not dominate the model.

We used TF-IDF with unigrams and bigrams, with unigrams as single words and bigrams as two-word sequences, preserving the contextual relationship between words. This representation is more effective for the model to detect patterns of abusive language.

3.5 Machine Learning Models

To classify Tamil and Malayalam comments as abusive or non-abusive, we examined and evaluated three different machine learning models in order to compare them. Each model has been selected on the basis of suitability for efficient text data processing and ability to handle the nuances of classification tasks.

- **Logistic Regression:** A probabilistic model that predicts the probability that comments are abusive using the logistic function. Vectored features representing the TF-IDF were used as input, and hyperparameters such as regularization strength C and solver optimization were tuned through grid searches. It provides a strong baseline performance with an accuracy of 87.48%, offering simple and interpretable results.
- **Support Vector Machine (SVM):** SVM uses the optimal hyperplane to separate abusive and non-abusive comments in the high-dimensional TF-IDF space. With a linear kernel, it identifies the optimal hyperplane. SVM achieved the highest accuracy (89.89%), demonstrating robust handling of text data and effective generalization.

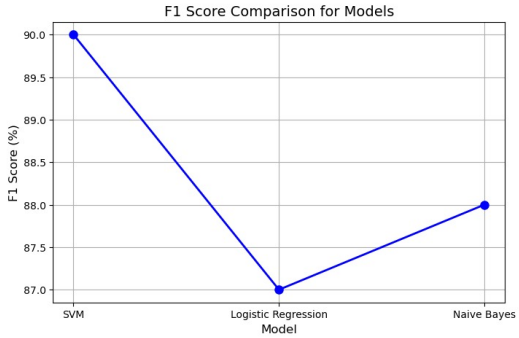

Figure 2: F1 score comparison for 3 models: Support Vector Machine, Logistic Regression, Naive Bayes.

- **Naive Bayes:** This probabilistic classifier uses Multinomial Naive Bayes to evaluate the likelihood of words contributing to each class. Its simplicity and effectiveness make it ideal for quick training and testing, achieving an accuracy of 87.51%.

3.6 Model Evaluation

The models were evaluated using metrics such as accuracy, F1 score, precision, and recall to ensure a balanced classification of offensive and non-abusive comments. Logistic regression achieved an accuracy of 87.48% and an F1 score of 87%, providing good baseline performance. SVM outperformed other models with the highest accuracy of 89.89% and an F1 score of 90%, showing its robustness in handling high-dimensional text data. As a result, Naive Bayes achieved an accuracy of 87.51% and an F1 score of 87%, known for its efficiency. Cross-validation was used to ensure robustness; the F1 score was prioritized to balance precision and recall. These evaluations demonstrate that while SVM achieved the best overall performance, logistic regression and Naive Bayes provide reliable and efficient

alternatives for deployment. As illustrated in Figure 3., enhanced model performance Improvement will be guided by an analysis of the 174 false positives, maybe using methods such as weighted loss functions for class imbalance. Plots like precision-recall curves and ROC will also be useful. Our system will be better able to trust people and make wise decisions when identifying abusive content if we reduce false positives.

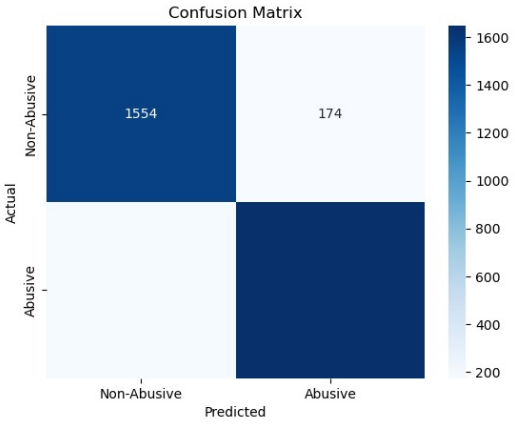

Figure 3: Comprehensive Evaluation and Improvement for the Malayalam dataset. Visualizing the model's performance in terms of false positives and false negatives.

4 Results

The project results indicate the effectiveness of different machine learning models for classifying abusive and non-abusive comments in Tamil and Malayalam. The models were used for evaluation based on accuracy, F1 score, precision, and recall so that the models exhibit almost balanced performance in both class categories (abusive and non-abusive). Here, a detailed summary of results is presented in Table 2.

The results provide strong evidence for the detection of abusive content in Tamil and Malayalam. It is advised that SVM be used for deployment due to its efficient performance on all fronts. Logistic regression and Naïve Bayes can act as simpler workarounds subject to resources. This system promises much greater potential use on real-time social media content management and user protection.

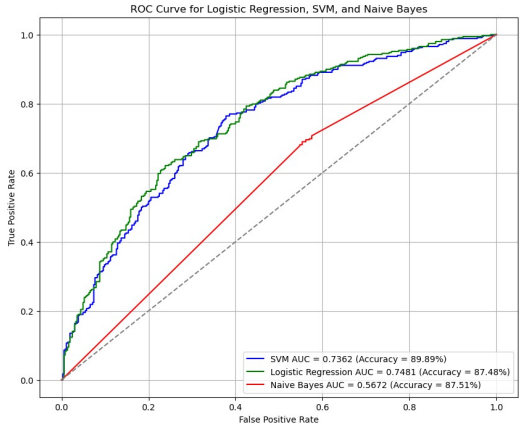

Figure 4: The ROC curve shows that SVM performs the best, followed by Logistic Regression, with Naive Bayes having the lowest AUC, indicating relatively poorer performance.

5 Conclusion

The widespread existence of gender-based abuse on social media platforms requires effective mechanisms to detect and mitigate abusive content targeted at women. In this work we have attempted the challenge of classifying comments in Tamil and Malayalam as abusive or non-abusive by applying robust preprocessing, exploratory analysis, and machine learning techniques. By using TF-IDF vectorization and fine-tuning hyperparameters, three models are Logistic Regression, Support Vector Machine (SVM), and Naive Bayes.

Among them, SVM was found to be the best model with an overall accuracy of 89. 89% and balanced F1 scores of 0. 90% for both the Abusive and Non-Abusive classes. This goes to show how SVM handles high-dimensional text features effectively, ensuring fair detection across categories. Among the other statistical procedures evaluated were Naive Bayes and Logistic Regression. The proposed solution combines high accuracy with interpretability by using WordCloud visualization to gain insight into language patterns. The study shows that machine learning can be used for automated moderation of abusive comments in Tamil and Malayalam languages. The results show the feasibility of using machine learning for automated moderation of abusive comments in Tamil and Malayalam-speaking language environments.

6 Limitation

The proposed solution was able to effectively detect abusive comments in both Tamil and Malayalam; still, a couple of limitations arose that would eventually further improve the model. It is not endowed with the deep contextual understanding that traditional models usually rely on, the basis of SVM, logistic regression, and naïve Bayes, as those models tend to use TF-IDF features and fail to incorporate implicit abuse, sarcasm, and

Model	Accuracy (%)	F1-Score (%)	Precision (Class 0, 1)	Recall (Class 0, 1)	AUC-ROC
SVM	89.89	90	(0.89, 0.90)	(0.90, 0.90)	0.7362
Logistic Regression	87.48	87	(0.88, 0.87)	(0.86, 0.89)	0.7481
Naive Bayes	87.51	88	(0.90, 0.86)	(0.84, 0.91)	0.5672

Table 2: Comparison of Logistic Regression, SVM, and Naive Bayes models on accuracy, F1-score, precision, recall and AUC-ROC curve for classifying abusive and non-abusive comments.

code-mixed language. This is also true because it only serves to process texts, and, most of the time, social media abuse on the actual platforms would encompass multimodal media like images, memes, or videos. It simply cannot sense the visual characteristics of the input, which, in turn, leads to non-recognition of abusive content in image formats or any other multimedia types. Another such limitation is in the context-based elements of any conversation, as this treats comments to be treated and looked into entirely independently without seeking prior interaction among them, leading it to be unable to identify indirect and unfolding abuse within several messages.

The training set is linguistically and culturally so limited in breadth that the current model does not allow it to better generalize to dialectic and other variant forms of spoken Tamil and Malayalam. Another challenge with biased training data is the problem of unfair classification, which hurts specific user groups. Finally, deep learning models for real-time deployment pose challenges in terms of computation: very fast inference with little loss in accuracy remains an open question.

6.1 Future work

To overcome these limitations, the next wave of future work focuses on some areas of improvement. The upgraded transformer-based models with BERT, mBERT, and IndicBERT will be widely used to improve contextual understanding and classification accuracy. The multimodal abuse detection module will be integrated by looking into the textual and visual aspects, where the system can find the harmfulness of content in a meme or other multimedia formats. This way, it can see more subtle, context-dependent abuse once it has gained the capacity for processing many threads and interactions into its historical analysis. This increases the size of the dataset over variations in combinations of linguistic or cultural differences that further augment the methods of data augmentation. Back-translations and replacement of synonyms, among others, may generalize a model for many Dravidian languages. Model optimization techniques like quantization, pruning, and knowledge distillation will be used to reduce the computational overhead while keeping the accuracy intact. Strategies for bias mitigation using explainable AI will be applied to the system to make it fairer and more interpretable in terms of responsible and ethical AI. The final aspect is cross-lingual transfer learning, which would enable the system to support multiple Dravidian languages, thereby making it applicable to a large extent. User feedback with mechanisms involving adaptive learning, where the system keeps on improving continuously. Adaptability toward the newly emerging patterns of online abuse would enable it to track these events correctly. So, with such updates and improvements, this proposed system could be a little more robust in terms of efficiency, as well as just fair enough regarding the detection of the right abusive content.

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

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

Bharathi Raja Chakravarthi, et al. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam and Kannada. In *DRAVIDIANLANGTECH*, pages 1–10.

Rajalakshmi, Ratnavel, Duraphe, Ankita, Shibani, Antonette. 2022. Abusive comment detection in Tamil using multilingual transformer models. *DLRG@DravidianLangTech-ACL2022*, 207–213.

Kathiravan Pannerselvam, et al. 2023. CSSCUTN@DravidianLangTech: Abusive comments detection in Tamil and Telugu. *DRAVIDIANLANGTECH*, pages 1–15.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multilingual transformers for offensive language identification on code-mixing text. *DRAVIDIANLANGTECH*, pages 100–110.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

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.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTechEACL2021: Offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Pradeep Kumar Roy and Abhinav Kumar. 2021. Sentiment analysis on Tamil code-mixed text using BiLSTM. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation*, pages 100–110, Online. CEUR.

Fazlourrahman Balouchzahi, Anusha Gowda, Hosa Halli Shashirekha, and Grigori Sidorov. 2022. MUCIC@TamilNLP-ACL2022: Detection of abusive comments in Tamil using 1D Conv-LSTM. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213.

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

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 DravidianLangTech 2023*. Recent Advances in Natural Language Processing, September.

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. https://aclanthology.org/2022.dravidianlangtech-1.44, DOI: 10.18653/v1/2022.dravidianlangtech-1.44.

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.