KEC'S CODE CRAFTERS@DravidianLangTech 2026 : Abusive Tamil Text Detection Targeting Women on Social Media

Nethrasri S [1] **Nivetha V**[1] **Malliga Subramanian**[2]

[1]Department of CSE, Kongu Engineering College, Perundurai, Erode, India.
{nethrasris.24cse@kongu.edu, nivethav.24cse@kongu.edu}
mallisenthil.cse@kongu.edu}

Abstract

Social media is gaining popularity; however, it has become a platform where digital toxicity abounds. Linguistically diverse women find themselves highly vulnerable to cyberbullying. As a result, there is the necessity to develop a solution capable of tackling region-specific languages. Our paper takes steps in solving this issue by presenting a classification model for "Abusive Tamil Text Detection Targeting Women on Social Media" at the DravidianLangTech-2026 challenge. For developing our model, we used the dataset of 25,948 comments for training and 915 comments for testing. Our main task was the ability to classify content into two classes "Abusive" or "Non-Abusive" for YouTube videos. There were specific challenges associated with using the Tamil language, such as its complicated construction and the fact that the language is often mixed with English within the same sentence. In order to resolve these issues, we designed a pipeline for cleaning these informal scripts. Finally, we ran four traditional ML models: SVM, Logistic Regression, Random Forest, and Multinomial Naive Bayes using TF-IDF as a method of extracting features from texts. Our model achieved the accuracy and F1 scores of 0.86 using Logistic Regression.

1 Introduction

This has created the phenomenon of digital public space, where the level of interaction has touched new heights. Such digital violence is a serious issue that is manifested by silencing of female voices, inflicting psychological harm, and fostering gender disparity in both the physical world and the online world. Tamil is a Classical Dravidian language spoken by more than 70 million speakers globally. Detecting abuse in the Tamil language has been found to be a challenging exercise owing to the complicated nature of Tamil language and code-mixing between Tamil and English in the online world. In addition, the lack of high-quality dataset poses another problem. Thus, there is clearly a significant gap in the research area, which needs to be addressed properly. In order to address this gap, the current research paper attempts to do so by focusing on binary classification problem of Tamil comments found in YouTube video clips. Such a classification would include two categories: either Abusive or Non-Abusive comments with special emphasis on text against women. This research was conducted in relation to the task provided by the shared task at the DravidianLangTech-2026, which motivated this study. A joint effort was made to provide datasets which contain 25,948 comments in Tamil language from YouTube videos labeled for their abuse against women in the training set, and a dataset with 915 comments in the test set. The aim of this study is to develop and evaluate the effectiveness of classical machine learning algorithms for the current task. The specific steps taken in this study include the following: firstly, designing a custom preprocessing pipeline for Tamil text in social media, secondly, obtaining numeric representation through TF-IDF vectorization, and thirdly, evaluating the performance of four different classifiers including SVM, Logistic Regression, Random Forest, and Naive Bayes.

2 Literature Survey

Unfortunately, despite the presence of over 250M speakers of Dravidian language family, very little effort has been put into analyzing the problem of gender-targeted abuse with regards to the Dravidian language family. As a point of reference, the base model proposed for Dravidian language family in DravidianLangTech@ACL2023 scored a meager 0.65 F1 score using SVM and TF-IDF [8]. However, recent developments such as the presentation by Rahman et al. [1] at DravidianLangTech-2025 showed an increase in performance where they got

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 284–288
July 3, 2026 ©2026 Association for Computational Linguistics

0.86-0.87 score using XLM-BERT, while Subramanian et al. [2] managed 0.82-0.84 F1 using ensemble techniques. Further achievements include obtaining 0.84 using TF-IDF-BERT hybrid classifier by Harini et al. [3], while others such as Mohan et al. [4] propose multimodal approaches, among others [5]. In addition to this, papers dealing with the gender-targeted abuse subtask deal with issues related to gender-directed hate speech detection [6] (precision: 0.83, recall: 0.81 with Logistic Regression), misogyny detection using CNN [7] (Tamil, F1: 0.79), SVM-based Gender-directed Hate Speech Detection [8] (Tamil, F1: 0.81) and various transformer based methods [9]. The state-of-the-art for Tamil language abusive language detection is currently at 0.87 using ensemble methods [10].

3 Materials and Methods

3.1 Dataset Description

The dataset used in the current research is that which has been provided by the organizers of "Abusive Tamil Text Detection Targeting Women on Social Media" shared task of DravidianLangTech-2026. The dataset includes 26,863 comments made on YouTube videos in Tamil, out of which, the training set has 25,948 comments and the testing set has 915 comments. YouTube videos were chosen as the sources of the data because it offers an active and diverse Tamil speaking population along with an engaging comments section. A few YouTube videos were randomly chosen based on the criterion that those included comments targeting women. There are abusive as well as non-abusive categories of comments in the dataset, in similar numbers in order to have an efficient training of the model. All of the comments were manually labeled by fluent Tamil speakers and experts in the field.

Non-Abusive (Label 0) The non-abusive comments consist of any benign, positive, or neutral comments, discussions, and questions towards other users or compliments regarding women. It does not include abusive comments towards women. Abusive (Label 1) Comments that contain derogatory, sexist, threatening statements, name-calling towards women, and other terms targeting women (such as looks and sexual harassment), sexually abusive comments, and threatening remarks are considered abusive.

This work focuses on detecting abusive and non abusive text in Tamil language. In this particular work, we have added the feature to detect abuse toward women. It has an important difference when compared to typical English language data sets; as it focuses on a sensitive topic group and a low resource language.

Table 1: Dataset Statistics and Splits

Dataset Split	Total Comments	Language
Training Set	25,948	Tamil
Test Set	915	Tamil

3.2 Pre-processing and Feature Extraction

To get rid of the noise in the social media texts, the following preprocessing techniques were applied to the comments. Firstly, the use of regular expressions helped us to strip the URLs, mentions, hashtags, and special characters like $ % # etc. Secondly, the emojis found in the comments were converted to text-based emojis, preserving their context through the emojis python library. Thirdly, whitespace normalization reduced consecutive spaces, tabs, and new lines into one space character. Finally, all texts were converted into lowercase letters.

Next, the data was vectorized with TF-IDF vectorization technique. Term Frequency is a measure of how many times a word appears in a document. Inverse Document Frequency decreases the weights of commonly used words across the documents. Multiplying the two metrics provides an efficient representation of the document because it helps determine the importance of the term in the document corpus.

The choice of ngrams to be considered by TfidfVectorizer includes unigrams and bigrams, which capture individual words and phrases related to women-centric abuse (such as gender-specific slangs). The minimum document frequency is 2, removing uncommon words from the document corpus. Finally, we limit the feature size to 10000 dimensions.

4 Proposed Classifier

We have developed four classical machine learning models for abusive Tamil text detection for women on social media.

We picked four classic machine learning classifiers, each with its own approach to the problem. All came from scikit-learn. We split our training

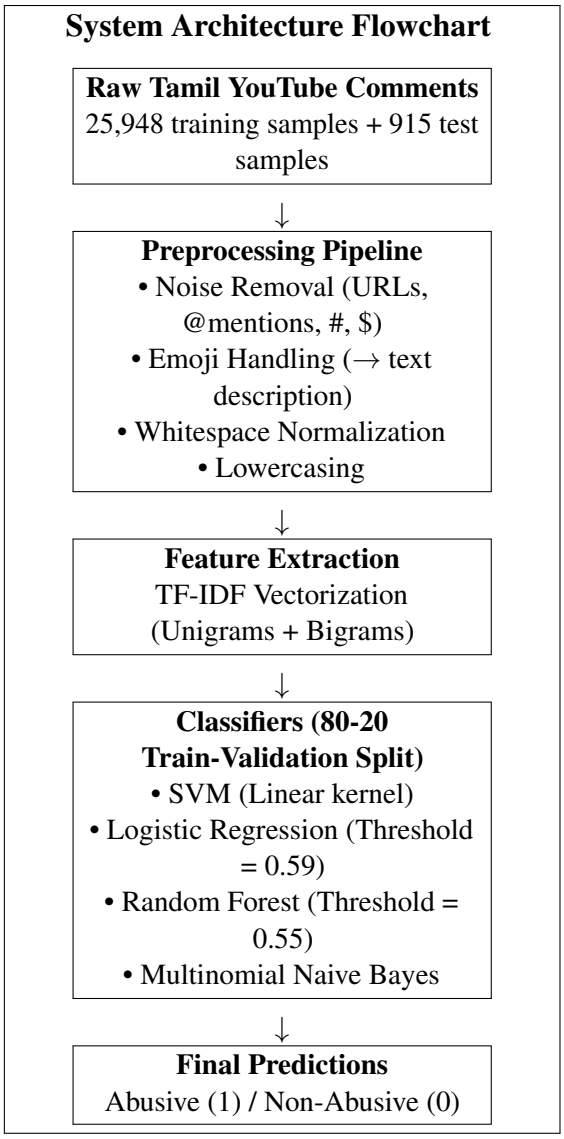

Figure 1: Complete system architecture for abusive Tamil text detection showing the preprocessing pipeline, feature extraction, and model training workflow with optimized parameters

data 80-20 into training and validation sets so we could tune hyperparameters and find the best thresholds for models that output probabilities.

4.1 Support Vector Machine (SVM)

We decided to implement linear SVM. Essentially, it looks for an optimal hyperplane that splits up abusive and non-abusive comments with maximal margin in that high-dimensional TF-IDF vector space. It is known to perform well with sparse, high-dimensional datasets such as text data. The default hyperparameters of LinearSVC was used here.

4.2 Logistic Regression (LR)

In Logistic Regression, probability of a comment being abusive is calculated by feeding an input of weighted features into what is called a sigmoid or a logistic function. The good thing about LR algorithm is that it is an interpretable model, meaning one could see which words contribute how to the final probability output. Moreover, the probabilities provided by LR are accurate. Some threshold tuning was performed for LR. By default, comments having probability of class ¿ 0.5 are classified as abusive. However, it turns out that we can tune the threshold to have a more accurate model. Our searches revealed that an optimal threshold is 0.59 on our validation dataset.

4.3 Random Forest (RF)

It is a machine learning algorithm that builds a large number of decision trees during training and outputs the class that is the most common among the classes output by individual trees. RF is able to model the complex nonlinear nature of the data set that linear classifiers cannot model. Similarly, like logistic regression, it can also predict probabilities. Key hyperparameters that will influence the prediction result of LR were selected: number of estimators (50, 100, 200) and maximum depth (10, 20, or None) using a grid search, and the optimal number of estimators (100) and unlimited depth were selected that provided the best validation accuracy with probability 0.55.

4.4 Naive Bayes (NB)

It is a probabilistic classifier based on the Bayes' theorem with the assumption that all predictors are independent of each other. That seems too simplistic, and realistically not possible because there are clear correlations between words in a sentence. Interestingly, Naive Bayes performs impressively well when predicting text-related problems such as frequency counts or TF-IDF features. Also, it is computationally efficient in terms of both training and predictions. The MultinomialNB method was utilized for features that are distributed according to a multinomial distribution.

5 Results

5.1 Performance Analysis

Logistic Regression had the best accuracy and F1 score of 0.86 which indicates the linear relationship between the TF-IDF features and the abusive

messages. SVM came second just like logistic regression having a F1 score of 0.84 and indicating that SVM works effectively on high-dimensional features. Random forest had low recall of 0.75 for the abusive label indicating that the model was overfitting the data. Naive bayes had the least F1 score of 0.80 since the features are not independent.

5.2 Why Logistic Regression Performed Best

In all the models analyzed, Logistic Regression gave the most optimal results with an accuracy and F1 score of 0.86 since TF-IDF features generate a sparse and high dimensional feature space which is highly appropriate for linear classifiers. The model could learn discriminative abusive terms while still generalizing well. Additionally, Logistic Regression generated calibrated probabilities which made it possible to optimize the threshold value to 0.59, thereby achieving a balance between precision and recall. SVM gave competitive results with an F1 score of 0.84, proving the efficiency of linear models when working with high dimensional text data. Random Forest had lower recall compared to the rest of the models as it was prone to overfitting with sparse lexical features. Naive Bayes yielded poor results since it could not model the context dependency of abusive terms in the Tamil social media text dataset.

5.3 Error Analysis

While showing good performance, the models had issues recognizing the abusive terms when they were indirect or context-based. False negatives were seen in cases where sarcasm was involved in the comment or the abusive sentiment was not explicitly mentioned but was implied in the comment. In many cases of false positives, comments with slang terms and emotional expressions, which were used in non-abusive settings, were wrongly labeled. The transliteration and code-mixing of Tamil-English comments made them harder to recognize due to inconsistencies in features caused by the use of phonetic spellings.

Table 2: Comparative Performance of All Models

Model	Prec%	Rec%	F1%	Acc%
LR+TF-IDF	86	86	86	86.0
SVM+TF-IDF	84	84	84	84.0
RF+TF-IDF	83	83	83	83.0
NB+TF-IDF	80	80	80	80.0

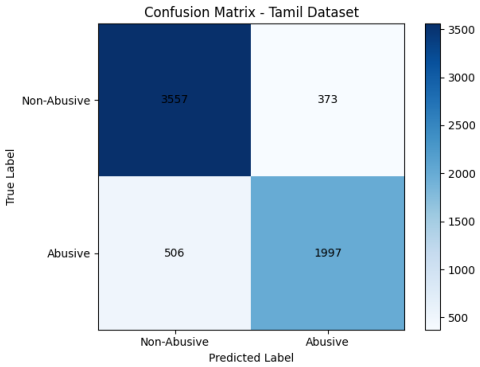

Figure 2: Confusion Matrix - Logistic Regression Model (Tamil Dataset)

6 Conclusion

This research work presented our approach towards the DravidianLangTech-2026 Shared Task for detecting abusive comments against women in the Tamil language. Our team designed four classical machine learning models utilizing TF-IDF features. Our proposed data preprocessing strategy proved to be highly effective for noisy Tamil social media data containing emojis, transliteration, code-switching text, and special characters.

From experimental analysis, Logistic Regression with parameter tuning turned out to be the best performing model among others like SVM, Random Forest, and Naive Bayes with an F1-score and accuracy of 0.86. The classifiers were still bound by their inability to deal with sarcasm, implied harassment and heavily mixed code language. Transliterated spellings of Tamil language and casual writing styles used on social media networks would further lower the stability of feature representations. Lastly, the current approach limited itself to textual data only, ignoring possible use of multimodal features, such as pictures and user profiles.

Future directions will include investigating transformer-based architectures like MuRIL and XLM-RoBERTa, along with ensemble modeling, contextual embeddings, and data augmentation approaches to enhance the effectiveness of semantic comprehension and abusive Tamil text identification tasks.

References

[1] Rahman, M.M., Dhar, S., Hasan, M.M., & Murad, H. (2025). MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL Based Transformer Models for Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media. In *Proceedings*

of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 243-247). Association for Computational Linguistics.

[2] Subramanian, M., Shanmugavadivel, K., Indhuja, V.S., Kowshik, P., & Jayasurya, S. (2025). KECEmpower@DravidianLangTech 2025: Abusive Tamil and Malayalam Text targeting Women on Social Media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. 178-181). Association for Computational Linguistics.

[3] Harini, B.N.S., Meghana, K.V., Supriya, K., Samiksha, T., & Premjith, B. (2025). HTMS@DravidianLangTech 2025: Fusing TF-IDF and BERT with Dimensionality Reduction for Abusive Language Detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. 152-156). Association for Computational Linguistics.

[4] Mohan, J., Mekapati, S.R., Premjith, B., Jyothish Lal, G., & Chakravarthi, B.R. (2025). 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*. DOI: 10.1145/3712260

[5] G, D., et al. (2025). Abusive Text Detection Targeting Women in Tamil and Malayalam using Machine Learning Classifiers. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages* (pp. XXX-XXX). Association for Computational Linguistics.

[6] Bade, G.Y., et al. (2025). GS_DravidianLangTech@2025: Women Targeted Abusive Texts Detection on Social Media. *arXiv preprint*, arXiv:2504.02863. https://arxiv.org/abs/2504.02863

[7] Shanmugavadivel, K., Subramanian, M., Pooja Sree, M., Palanimurugan, V., & Roshini Priya, K. (2025). InnovationEngineers@DravidianLangTech 2025: Enhanced CNN Models for Detecting Misogyny in Tamil Memes Using Image and Text Classification. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

[8] Vaidyanathan, V.K., Srihari, V.K., & Durairaj, T. (2025). NLP_goats@DravidianLangTech 2025: Towards Safer Social Media: Detecting Abusive Language Directed at Women in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

[9] Thirumoorthy, S., Durairaj, T., & Rajalakshmi, R. (2025). Hydrangea@DravidianLanTech2025: Abusive language Identification from Tamil and Malayalam Text using Transformer Models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

[10] Chakravarthi, B.R., Priyadharshini, R., Madasamy, A.K., Thavareesan, S., Sherly, E., Rajiakodi, S., Palani, B., Subramanian, M., Cn, S., & Chinnappa, D. (Eds.). (2025). *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Albuquerque, New Mexico: Association for Computational Linguistics.

code Availability: The implementation of our proposed model is publicly available at: https://github.com/NivethaVallarasu/dravidian