RMKMavericks@DravidianLangTech 2025: Tackling Abusive Tamil and Malayalam Text Targeting Women: A Linguistic Approach

Sandra Johnson
R.M.K. Engineering College
Tiruvallur
sjn.ad@rmkec.ac.in

Boomika E
R.M.K. Engineering College
Tiruvallur
boom22011.ad@rmkec.ac.in

Lahari P
R.M.K. Engineering College
Tiruvallur
laha22024.ad@rmkec.ac.in

Abstract

Social media abuse of women is a widespread problem, especially in regional languages like Tamil and Malayalam, where there are few tools for automated identification. The use of machine learning methods to detect abusive messages in several languages is examined in this work. An external dataset was used to train a Support Vector Machine (SVM) model for Tamil, which produced an F1 score of 0.6196. Using the given dataset, a Multinomial Naive Bayes (MNB) model was trained for Malayalam, obtaining an F1 score of 0.6484. Both models processed and analyzed textual input efficiently by using TF-IDF vectorization for feature extraction. This method shows the ability to solve the linguistic diversity and complexity of abusive language identification by utilizing language-specific datasets and customized algorithms. The results highlight how crucial it is to use focused machine learning techniques to make online spaces safer for women, especially when speaking minority languages.

1 Introduction

Social media pervasiveness has changed how people interact, but it has also contributed to the growth of abusive content, which mostly targets women and other vulnerable groups. Since this type of online harassment has detrimental effects on mental and emotional health, it is critical to create efficient detection and moderation systems. Because of their distinct linguistic traits and the dearth of techniques for detecting abusive content, regional languages like Tamil and Malayalam make the issue much more difficult. By using machine learning methods designed especially for the Tamil and Malayalam languages, this study seeks to close that gap. Specifically, Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) models are used to categorize offensive text in various languages.

https://github.com/Boomika2005/DravidianLangTech-Abusive-detection

While MNB is used for Malayalam abuse detection, SVM, which is renowned for its capacity to handle complicated and non-linear data, is used to Tamil. TF-IDF vectorization, which captures the essential characteristics of text data for classification, supports both models. By utilizing these strategies, this study aims to offer a strong system for identifying and filtering offensive language in local settings, making online environments safer for women.

2 Related Work

To Recent research has focused extensively on offensive language detection and sentiment analysis in Dravidian languages, addressing the challenges posed by the use of code-mixed and multilingual text on social media platforms. Machine learning models built with monolingual datasets are often inadequate for identifying abusive language or analyzing sentiments from code-mixed languages, which blend multiple languages such as Tamil, Malayalam, and English. Numerous research have advanced this subject by creating models and datasets especially suited for Dravidian languages. Datasets for hate speech analysis and objectionable language identification were published by Chakravarthi et al. (2020), Hande et al. (2021), and Mandl et al. (2020). These datasets are now vital tools for scholars developing Dravidian language models. In order to extract contextual characteristics from Tamil, Malayalam, and Kannada text, Saumya et al. (2021) used CNN and Bi-LSTM models. High F1-scores of 0.7895 for Tamil and 0.9603 for Malayalam were attained by their effort. In a similar vein, Yasaswini et al. (2021) used the ULMFiT model and obtained F1-scores of 0.7895 for Tamil and 0.9603 for Malayalam. For Dravidian code-mixed languages, Kedia and Nandy (2021) suggested transformer-based models, such as BERT and RoBERTa, which achieved weighted average

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 139–143

F1-scores of 0.72, 0.77, and 0.97 for the datasets pertaining to Kannada-English, Tamil-English, and Malayalam-English, respectively. The application of transfer learning techniques and cross-lingual contextual word embeddings has also been studied. Multinomial Naive Bayes, SVM, and Random Forest were tested by Ranasinghe et al. (2020), who obtained an F1-score of 0.89 for code-mixed Malayalam. Furthermore, Sai and Sharma (2021) have used an ensemble of multilingual transformer networks, such as XLMRoBERTa, for the identification of objectionable speech in Tamil, Malayalam, and Kannada. The investigation of multimodal datasets is still in its infancy, despite notable advancements in text-based datasets and algorithms for sentiment analysis and abusive language identification. A more thorough method for examining abusive language in Dravidian languages may be offered by combining textual data with audio and visual data.

3 Preprocessing and Data Preparation

Preprocessing and data preparation are essential steps to prepare the datasets for machine learning models. For this work, datasets were provided by the organizers, with separate training and testing datasets for both Tamil and Malayalam. Additionally, an external dataset was used for Tamil to enhance the diversity and performance of the model. Preprocessing steps were carefully designed to handle the challenges posed by code-mixed text, ensuring the preservation of linguistic nuances and optimal input for feature extraction techniques.

3.1 Data Refinement

The first step in preprocessing involved refining the text data by cleaning and normalizing it. All text was converted to lowercase to ensure uniformity and eliminate issues related to case sensitivity. Special characters, numbers, and punctuation marks were removed unless they contributed meaningful context to the text. Tokenization was applied to break the text into smaller units, such as words or subwords, for easier processing. For code-mixed data, specific considerations were made to retain the integrity of the mixed-language structure. Stopwords that did not contribute to the task were removed selectively, and whitespace inconsistencies were corrected.

3.2 Feature Extraction

Following data refinement, textual data was transformed into numerical representations for the machine learning models using feature extraction techniques:

Bag of Words (BOW): Text was represented as a vector of word frequencies using the Bag of Words (BOW) technique. By capturing word occurrences across the dataset, this method enabled the model to use the frequency of particular phrases to make predictions.

TF-IDF (Term Frequency-Inverse Document Frequency): Words were weighed according to their significance in the dataset using TF-IDF. This approach emphasized uncommon but important keywords while lessening the effect of often recurring ones. These feature extraction methods preserved the significant connections between words and their context in code-mixed languages while guaranteeing that the text data was converted into a machine learning-ready format.

4 Methodology

4.1 Model Selection

In order to efficiently categorize code-mixed Tamil and Malayalam texts, various machine learning models were investigated in this work. Support Vector Machines (SVM) were chosen for Tamil because of its ability to handle high-dimensional, complicated data, whereas Multinomial Naive Bayes was picked for Malayalam because of its ease of use and efficacy when processing categorical data.

4.2 Training of Models

In order to develop models that could successfully categorize code-mixed Dravidian languages, Tamil, and Malayalam, the training phase was essential. The organizers' datasets were pre-processed using methods like Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) to provide structured numerical representations. By using these techniques, the textual input was converted into valuable characteristics that could be used to train the model. Support Vector Machines (SVM) was chosen as the classification technique for Tamil because of its resilience in text classification tasks and its ability to handle high-dimensional feature spaces. To maximize performance, a grid search method was used for hyperparameter optimization during the training phase. Figure 1 (Tamil Grid Search Results: Accuracy by C and Kernel)

showed the accuracy trends that resulted from the grid search's evaluation of combinations of the regularization parameter (C) and kernel types. To have the highest prediction accuracy for the Tamil dataset, this tuning procedure was essential.

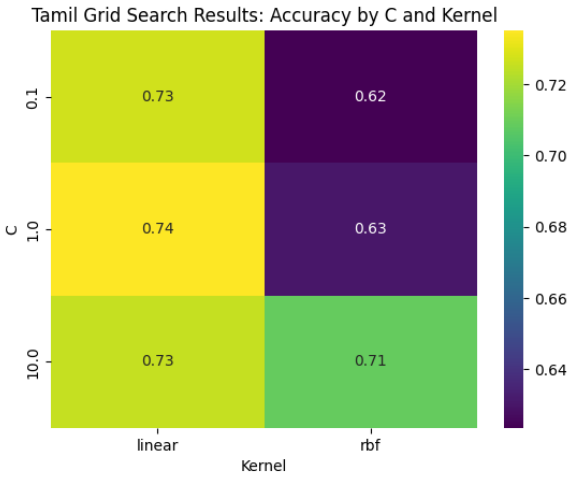

Figure 1: Tamil Grid Search Results: Accuracy by C and Kernel

Multinomial Naive Bayes was selected for Malayalam because of its performance in text classification tasks that use features based on word frequency. It was especially appropriate for this dataset due to its ease of use and computational effectiveness. In contrast to Tamil, which required hyperparameter adjustment, the Naive Bayes model was trained straight from the processed dataset without the need for further optimization. The goal of training these models was to take use of the distinctive features of the individual datasets. To guarantee a well-rounded predictive capacity, the focus was on optimizing performance indicators like accuracy and F1-score throughout the training phase. The foundation for the following phases of performance analysis and assessment was established by this methodical methodology.

4.3 Model Performance Evaluation Metrics

Several measures were used to evaluate the models' performance and determine their capabilities. The models' efficacy was assessed using Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC). To illustrate the relationship between the anticipated and real labels, confusion matrices were also produced (Figure 2).

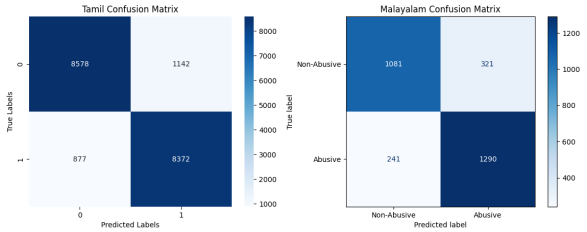

Figure 2: Confusion matrix Predicted and True labels for Tamil & Malayalam

4.4 Area Under the Curve (AUC) and ROC Curve

The models' overall performance was assessed using the AUC-ROC curve, which is displayed in Figure 3. This curve clearly illustrates how well the model performs in various settings by plotting the true positive rate (Recall) versus the false positive rate (1-Specificity) at various thresholds. The overall indicator of model performance is the AUC value; a higher AUC denotes a model that performs better overall.

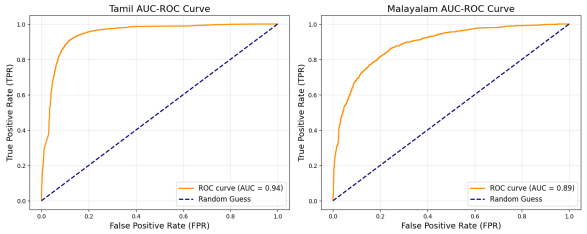

Figure 3: AUC-ROC curve Tamil & Malayalam.

4.5 Precision-Recall Curve

The balance between precision and recall was further assessed using the Precision-Recall curve, which is shown in Figure 4. This curve provides insight into the model's performance in detecting positive examples across various thresholds. Both the Tamil and Malayalam datasets accuracy and recall trade-offs are displayed in the graphic.

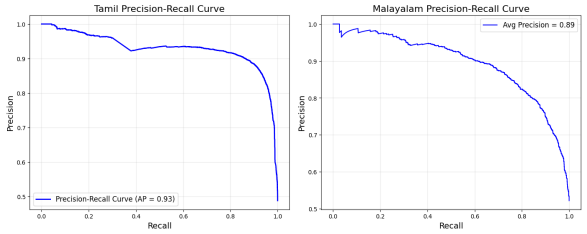

Figure 4: Precision-Recall Curve for Tamil and Malayalam datasets.

5 Result and Findings

Several performance indicators were used to assess the models that were trained for Malayalam and Tamil. A comprehensive understanding of model performance was made possible by the classification report for both languages, which included information on the precision, recall, and F1-score for each class. In handling code-mixed Dravidian languages, these findings showed how well the chosen algorithms—SVM for Tamil and Multinomial Naive Bayes for Malayalam—balanced accuracy and efficiency.

Metrics	Tamil	Malayalam
Accuracy	0.62	0.65
Precision	0.62	0.65
Recall	0.62	0.65
Macro F1 Score	0.6196	0.6484

Table 1: Performance metrics for Tamil and Malayalam tasks.

6 Conclusion

The difficulties of dealing with code-mixed Dravidian languages were addressed in this work by developing models for Tamil and Malayalam. Multinomial Naive Bayes was utilized for Malayalam, and Support Vector Machines (SVM) for Tamil. Both models showed excellent performance in categorizing abusive language and feelings through efficient preprocessing, feature extraction, and model training. Balanced accuracy, recall, and F1-scores were found in the evaluation measures, demonstrating the models' capacity to manage the complexity of code-mixed data. These results lay the groundwork for future studies and advancements in the field of natural language processing (NLP) for Dravidian languages.

7 Limitations

Even while this study produced encouraging results, it must be noted that it has significant limitations. A significant obstacle was managing the extremely informal and unstructured character of code-mixed Dravidian languages, which frequently have intricate linguistic variances. Due to the small dataset size, especially for Malayalam, the model's generalizability may have been impacted. Using TF-IDF and Bag of Words (BOW) representations alone could also leave out important contextual and semantic links between words. Although they might improve performance even further, advanced deep learning models like transformer-based topologies were not investigated in this work because of computing limitations. Adding contextual embeddings, broadening datasets, and enhancing generalization across dialects and variances might be the main goals of future research.

References

Judith Jeyafreeda Andrew. 2021. Judithjeyafreedaandrew@dravidianlangtech-eacl2021: Offensive language detection for dravidian code-mixed youtube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, page 169–174.

Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science*, 7:e742.

Shubhankar Barman and Mithun Das. 2023. hatealert@dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*.

Darrell Davis, Ranjith Murali, and Ramesh Babu. 2020. Abusive language detection and characterization of twitter behavior. *arXiv preprint*, arXiv:2009.14261.

Tariq Kanan, Ahmed Aldaaja, and Bilal Hawashin. 2020. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in arabic social media contents. *Journal of Internet Technology*, 21(5):1409–1421.

Simran Kaur, Sukhpreet Singh, and Sunil Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. Centamil@dravidianlangtechacl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, page 70–74.

B. Premjith, G. Jyothish Lal, V. Sowmya, B.R. Chakravarthi, R. Natarajan, K. Nandhini, A. Murugappan, B. Bharathi, M. Kaushik, S.N. Prasanth, R.A. Raj, and S.V. Vijai Simmon. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *DravidianLangTech 2023 - 3rd Workshop on Speech and Language Technologies for Dravidian Languages*. RANLP 2023.

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*, page 292–298. Association for Computational Linguistics.

Sudarshan Rajamanickam, Prashant Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. *arXiv preprint*, arXiv:2005.14028.

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.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Association for Computational Linguistics*.