CoreFour_IIITK@DravidianLangTech 2025: Abusive Content Detection Against Women Using Machine Learning And Deep Learning Models

Varun Balaji S$_1$, Bojja Revanth Reddy$_1$, Vyshnavi Reddy Battula$_1$, Suraj Nagunuri$_1$,Balasubramanian Palani$_2$
$_1$Department of Computer Science and Engineering, IIIT Kottayam, Kerala, India
$_2$Assistant Professor, Indian Institute of Information Technology Kottayam
{varun22bcs152, revanth22bcs210, vyshnavi22bcs133, suraj22bcy35, pbala}@iiitkottayam.ac.in

Abstract

The rise in utilizing social media platforms increased user-generated content significantly, including negative comments about women in Tamil and Malayalam. While these platforms encourage communication and engagement, they also become a medium for the spread of abusive language, which poses challenges to maintaining a safe online environment for women. Prevention of usage of abusive content against women as much as possible is the main issue focused in the research. This research focuses on detecting abusive language against women in Tamil and Malayalam social media comments using computational models, such as Logistic regression model, Support vector machines (SVM) model, Random forest model, multilingual BERT model, XLM-Roberta model, and IndicBERT (Rajiakodi et al., 2025). These models were trained and tested on a specifically curated dataset containing labeled comments in both languages. Among all the approaches, IndicBERT achieved a highest macro F1-score of 0.75. The findings emphasize the significance of employing a combination of traditional and advanced computational techniques to address challenges in Abusive Content Detection (ACD) specific to regional languages.

1 Introduction

Detecting abusive content targeting women on Online Social Networks (OSNs) presents a significant challenge in the current digital era. Rising instances of online harassment against women in Dravidian language communities have become a growing concern. The increase of anonymous accounts on social media platforms enables harmful behavior to spread, emphasizing the urgent need for robust detection systems to address this issue and protect vulnerable users. The research began with the implementation of a comprehensive preprocessing framework, incorporating thorough data cleaning, normalization, and tokenization processes. For feature engineering, we employed IndicBERT's embedding techniques to effectively capture essential linguistic patterns. While contemporary research typically gravitates towards transformer architectures for their perceived advantages, our findings highlight the efficiency of well-optimized IndicBERT-based methods.

The systematic evaluations of traditional machine learning algorithms were conducted, including Logistic Regression, SVM, and Random Forest classifiers (Priyadharshini et al., 2022). To provide a comprehensive analysis, we extended our investigation to include cutting-edge transformer-driven models such as the mBERT model, XLM-RoBERTa model, and IndicBERT model (Hariharan et al., 2024), each fine-tuned specifically for binary classification tasks. In recent times, transformer models like IndicBERT have demonstrated superior performance. Notably, in our case, the IndicBERT-based model achieved the highest accuracy, surpassing even mBERT. This highlights the effectiveness of language-specific pretraining in enhancing classification performance.

The major contributions of the paper are as follows:

- To predict abusive content against women, classical text encoding techniques are first used in the embedding layer and then Machine Learning (ML) algorithms are utilized.
- To explore transformer-based embedding model with Deep Learning (DL) model, we used mBERT, XLM-Roberta and IndicBERT for contextual feature extraction and then FFN is used for abusive content detection against women.
- To test the working of the proposed model, we utilized the benchmark dataset of abusive content against women.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 707–712
May 3, 2025 ©2025 Association for Computational Linguistics

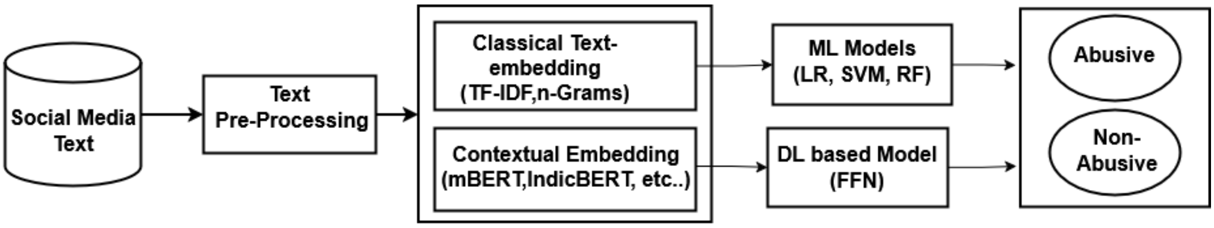

Figure 1: Architecture of the proposed model for ACD.

2 Literature Survey

Research in abusive language detection against women has shown varied approaches. Chakravarthi and Priyadharshini (Chakravarthi et al., 2023) found classical Machine Learning models like Logistic Regression outperforming deep learning on Fine-Grained Abusive Comment Detection (FGACD) due to limited data. Gupta and Roychowdhury (Gupta et al., 2022) improved performance using Term Frequency-Inverse Document Frequency (TF-IDF) for extracting features. Vegupatti et al. (Vegupatti et al., 2023) , Premjith et al. (Premjith et al., 2023) , and Hariharan et al (Hariharan and Anand Kumar, 2022) has focused on leveraging advanced BERT-based models like IndicBERT, MuRIL, and mBERT for multilingual content analysis. Their studies primarily target detecting abusive content against women and fake news across Indian languages, demonstrating the capability of deep learning methods to comprehend and classify text across diverse linguistic contexts.

3 Methodology

The architecture of the proposed model for abusive content detection against women is shown in Figure 1. Throughout the work, IndicBERT, a pretrained language model, is utilized in the embedding layer alongside TF-IDF. Notably, IndicBERT achieved the highest Macro F1-score, demonstrating its effectiveness in capturing linguistic patterns and enhancing classification performance.

3.1 Problem Definition

Abusive language detection against women is framed as a binary classification problem. Considering a dataset $C = \{c_1, c_2, \dots, c_k\}$ consisting of k social media comments, each comment $c_i \in C$ is linked to a class label $y \in \{\text{non-abusive}(0), \text{abusive}(1)\}$. Each comment c_i contains p sentences $\{s_1, s_2, \dots, s_p\}$, where each sentence $s_p \in c_i$ consists of q words $\{w_{j1}, w_{j2}, \dots, w_{jq}\}$. A classification model $f : C \to y$ is defined and trained to predict the class label $y_j \in \{0, 1\}$, where $y_j = 0$ represents non-abusive content, and $y_j = 1$ represents abusive content.

3.2 Data Preprocessing

Data preprocessing prepares the dataset for training an Machine Learning model to detect abusive language in Tamil and Malayalam. The dataset is loaded, class labels are normalized, and missing values are replaced with empty strings. Text is cleaned by removing special characters, punctuation, and numbers. Removing noisy data by stabilizing the dataset for further usage.

Handling low-resource languages presents challenges due to limited annotated data and complex morphology. Key preprocessing includes Unicode normalization, subword tokenization, and noise filtering. Stopword removal eliminates Tamil and Malayalam stopwords, and English words are removed for consistency. The dataset is equalized across classes to prevent overfitting, ensuring a balanced dataset for effective abusive content detection.

3.3 Embedding Layer

The embedding layer converts text into numerical representations after removing Tamil and Malayalam stopwords to enhance relevance.

3.3.1 Traditional Text Encoding Techniques

TF-IDF is used where every word is given a weight based on its occurrence in a comment and its scarcity across the dataset (Shanmugavadivel et al., 2022). This method ensures the model prioritizing words that are most significant for distinguishing between abusive and non-abusive content for women.

3.3.2 Transformer-Based Embedding Techniques

The vectorized text data is divided into training and testing sets, enabling both model learning and performance evaluation. To improve the detection of abusive language targeting women, **pre-trained transformer-based embeddings** (such as BERT, XLM-R, or IndicBERT) are fine-tuned using a domain-specific dataset. This fine-tuning process involves updating model parameters through supervised learning, where abusive and non-abusive samples in Tamil and Malayalam for women help the model refine its contextual understanding.

Transformer models employ the **self-attention mechanism (SAM)** to identify relationships between words in a sequence, even over long distances. SAM assigns attention scores to determine the relative importance of different tokens, as described in Eq. (1):

$$\mathrm{A} = \mathrm{softmax}\left(\frac{qk^{\top}}{\sqrt{d_k}}\right)v \tag{1}$$

where **A** denotes the attention matrix, **q** represents the query matrix, **k** is the key matrix, **v** corresponds to the value matrix, and $\mathbf{d_k}$ is the dimensionality of the key vectors.

Fine-tuning allows the transformer model to adapt to linguistic characteristics and abusive language patterns in Tamil and Malayalam. A classification layer, placed on top of the transformer, further enhances the model's ability to differentiate between abusive and non-abusive text based on Linguistic patterns. The model's performance is optimized by fine-tuning hyperparameters and evaluating key metrics such as precision, recall, and macro F1-score.

3.4 Classification Layer

After preprocessing and vectorization, the data is passed through classification models of ML and DL models to predict if a comment is abusive for women.

3.4.1 ML Models

Machine learning models being used are Logistic Regression, SVM, and Random Forest are used for classifying the text for abusive or non-abusive. The importance of classification is that it becomes easier to predict the input data.

3.4.2 DL Models

Deep learning models being used are multilingual Bert model, XLM-Roberta model and IndicBert model have been used. These models are used in embedding layer to efficiently represent data while capturing semantic relationships.

4 Experiment

This section provides the summary of the benchmark dataset utilized in this study, along with details of the experimental setup, dataset and performance metrics.

4.1 Experimental Setup

The experiment uses ML and DL models, implemented and tested in Jupyter Notebook, which integrates code, visuals, and text for streamlined computation, debugging, and visualization. TensorFlow, with its high-level API Keras, supports model implementation for architectures like mBERT, XLM-RoBERTa, and IndicBERT. The dataset is divided into 75% training, 25% testing, ensuring sufficient data for learning as well as performance evaluation.

Table 1: Summary of Datasets

Language	Class	Number of Samples
Tamil	Abusive	1366
	Non-abusive	1424
Malayalam	Abusive	1531
	Non-abusive	1402

4.2 Dataset

The dataset utilized in this study is well-defined and curated for accurate modeling and evaluation as shown in Table 1 (Premjith et al., 2023; Priyadharshini et al., 2022). It comprises YouTube comments in Tamil and Malayalam, annotated as Abusive or Non-Abusive content, specific to women as shown in Table 2. The data was sourced from publicly available comments and manually labeled based on linguistic and contextual cues.

Abusive comments contain derogatory, offensive, or misogynistic language, while non-abusive ones do not. The dataset also handles code-mixed text (e.g., Tanglish, Manglish). To prevent overfitting, we balanced the dataset by equalizing samples across both classes, ensuring unbiased learning. It was split into training, validation, and test-

Table 2: Example of Abusive and Non-Abusive Content in Tamil and Malayalam dataset

Type	Tamil	Malayalam
Abusive	ஆமா பா கட்டி வச்சி தோல உரிக்கணும்	നിനക്ക് വേണ്ടി വോട്ട് ചെയ്തവരെ തിരിഞ്ഞു കൊത്തുന്നു
Non-Abusive	எப்படி இருக்கிறீர்கள்?	സൂരജന്റെ കല്യാണം എപ്പോൾ നടക്കു അപ്പോൾ അറിയാം......

ing sets for evaluation. Model performance was assessed using precision, recall, and macro F1-score to ensure reliability.

For reproducibility, we use datasets from prior research (Priyadharshini et al., 2022).

4.3 Performance metrics

Metrics on which the models have been evaluated are accuracy, precision, Recall, F1-Score and macro F1-Score from Eq - (2) to (6).

$$\mathrm{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$F1 = 2 \cdot \left(\frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \right) \tag{5}$$

$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \left(\frac{\mathrm{P}_i \cdot \mathrm{R}_i}{\mathrm{P}_i + \mathrm{R}_i} \right) \tag{6}$$

Where, **A**: Accuracy, **P**: Precision, **R**: Recall, **TP**: True Positive, **TN**: True Negative, **FP**: False Positive, **FN**: False Negative, **N**: Number of classes, $\mathbf{P_i}$: precision for class i, and $\mathbf{R_i}$: recall for class i.

5 Results and Analysis

This research investigates the performance of different models on two practical datasets in Tamil and Malayalam, focusing on key evaluation metrics like **Accuracy**, **Precision**, **Recall**, and **macro F1-score**. As illustrated, Table 3 shows the results of various models in terms of Accuracy, Precision, Recall, F1-Score and Macro F1-Score. The evaluation involved comparing traditional ML models, multilingual transformer-driven DL models, and cutting-edge techniques through two distinct experimental setups, aiming to identify the most effective model for detecting abusive content against women in these languages.

The findings reveal that the IndicBERT model outperforms other models across both datasets. This advantage stems from IndicBERT's ability to leverage pre-trained contextual embeddings, making it well-suited for handling code-mixed text and morphologically rich languages like Tamil and Malayalam. By capturing complex linguistic patterns without relying on manual feature engineering, IndicBERT effectively enhances performance on low-resource abusive content detection tasks. These results emphasize the importance of transformer-based models in such challenging scenarios.

5.1 Performance Comparison of Traditional and Transformer-Based Models for Tamil Text Classification

The evaluation of traditional embeddings like **TF-IDF** (Shanmugavadivel et al., 2022) and transformer-based models such as **mBERT**, **XLM-RoBERTa**, and **IndicBERT** (Reshma et al., 2023) demonstrated the superiority of transformer models in detecting abusive content as illustrated in Figure 2, particularly targeting women in Tamil text. Among these, **IndicBERT** emerged as the most effective model, achieving a **Macro F1-score of 0.750** for Tamil dataset. Its multilingual pretraining enabled it to grasp complex linguistic structures and contextual variations in Tamil, making it particularly well-suited for abusive content detection in these low-resource languages.

Traditional models like Random Forest, SVM and Logistic Regression struggled to identify im-

Table 3: Performance Comparison of the Proposed Model on Tamil and Malayalam Datasets

Model		Tamil					Malayalam				
Embedding	Classifier	A	P	R	F1	Macro F1	A	P	R	F1	Macro F1
TF-IDF	LR	0.669	0.670	0.668	0.669	0.669	0.661	0.671	0.648	0.652	0.659
TF-IDF	SVM	0.634	0.633	0.633	0.650	0.633	0.659	0.659	0.659	0.650	0.659
TF-IDF	RF	0.631	0.631	0.630	0.636	0.631	0.623	0.621	0.618	0.600	0.619
mBERT	FFN	0.748	0.748	0.746	0.746	0.745	0.685	0.680	0.680	0.682	0.685
XLM-R	FFN	0.730	0.730	0.730	0.733	0.735	0.700	0.705	0.700	0.703	0.705
IndicBERT	FFN	**0.750**	**0.750**	**0.750**	**0.750**	**0.750**	**0.720**	**0.725**	**0.720**	**0.720**	**0.720**

plicit abuse due to their reliance on manual features. In contrast, **IndicBERT** leveraged deep contextual understanding to detect subtle, context-dependent toxicity, outperforming them in Tamil. This highlights the advantage of transformers in low-resource language abuse detection.

5.2 Performance Comparison of Traditional and Transformer-Based Models for Malayalam Abusive Content Detection

For the Malayalam dataset, a comparative analysis of traditional embeddings (**TF-IDF** and **Hugging Face**) and transformer-based models (**mBERT**, **IndicBERT**, and **XLM-RoBERTa**) (Reshma et al., 2023) demonstrated the effectiveness of transformer-driven approaches in detecting abusive content as illustrated in Figure 2 particularly targeting women. Among these, **IndicBERT** emerged as the top performer, achieving a **Macro F1-score of 0.720** for Malayalam. Its multilingual pretraining allowed it to grasp the contextual intricacies of abusive language, making it highly effective in distinguishing both implicit and explicit forms of abuse.

Traditional models like Random Forest and SVM performed well in classification but struggled with the complexity of abusive language due to their reliance on manual features. In contrast, **IndicBERT** leveraged deep contextual embeddings to detect subtle abuse, excelling in identifying gender-targeted toxicity. This underscores the importance of transformers in addressing online abuse in Malayalam, where traditional methods often fall short.

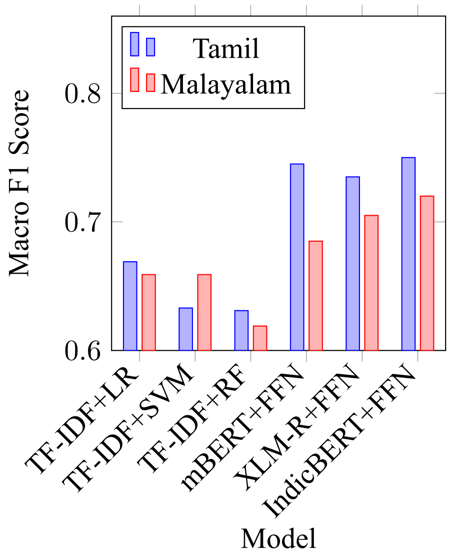

Figure 2: Comparison of Models on Macro F1-Score for Tamil and Malayalam Datasets

6 Conclusion and Future work

This study provides a comprehensive evaluation of Machine Learning and Deep Learning models for abusive content detection in Tamil and Malayalam. Contrary to previous trends favoring traditional machine learning models, transformer-based approaches, particularly **IndicBERT**, demonstrated superior performance in detecting gender-targeted toxicity. **IndicBERT** achieved the highest **Macro F1-score** of **0.750** for Tamil and **0.720** for Malayalam, effectively capturing both explicit and implicit forms of abuse. In contrast, traditional models like Random Forest, SVM, and Logistic Regression, which rely on manually engineered features, struggled to handle the complex and evolving nature of abusive language.

Despite these advancements, challenges remain in understanding model errors and improving robustness. Future research should focus on error analysis, real-world testing, and refining preprocessing to enhance transformer models for low-resource languages like Tamil and Malayalam. Additionally, incorporating external linguistic resources, improving contextual understanding, and mitigating biases within datasets could further enhance model effectiveness. Addressing domain-specific variations and handling code-mixed language more efficiently are also crucial for making these models more adaptable to real-world applications.

References

B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, and J. P. McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

V. Gupta, S. Roychowdhury, M. Das, S. Banerjee, P. Saha, B. Mathew, and A. Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Advances in Neural Information Processing Systems*, volume 35, pages 26176–26191.

R. I. L. Hariharan and M. Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 196–208. Springer.

R. L. Hariharan, M. Jinkathoti, P. S. P. Kumar, and M. A. Kumar. 2024. Fake news detection in telugu language using transformers models. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.

B. Premjith, V. Sowmya, B. R. Chakravarthi, R. Natarajan, K. Nandhini, A. Murugappan, B. Bharathi, M. Kaushik, and P. Sn. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

R. Priyadharshini, B. R. Chakravarthi, S. Cn, T. Durairaj, M. Subramanian, K. Shanmugavadivel, S. Hegde, and P. 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. 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.

S. Reshma, B. Raghavan, and S. J. Nirmala. 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.

K. Shanmugavadivel, S. U. Hegde, and P. K. Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. *DravidianLangTech*, page 292.

M. Vegupatti, P. K. Kumaresan, S. Valli, K. K. Ponnusamy, R. Priyadharshini, and S. 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.