RESEARCH ARTICLE OPEN ACCESS

Detection of Abusive Tamil Text Targeting Women on Social Media

Dr. S. Padmapriya*, R.Mathimithra**, B. Haripriya**, G. Kalaivani**, R. Swedha**

*Professor -Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai,

**UG Students- Computer Science and Engineering, A.V. C. College of Engineering, Mannampandal, Mayiladuthurai, India

ABSTRACT

Social media platforms have witnessed a rapid rise in abusive and misogynistic content targeting women, especially in regional languages such as Tamil. Manual moderation is inefficient given the enormous volume of user-generated text. This paper proposes an AI-based system for detecting abusive Tamil comments targeting women on social media using a hybrid model that combines TF-IDF with Logistic Regression, MuRIL, and multilingual BERT. The system further integrates Firebase services including Authentication, Firestore, and Cloud Messaging for real-time notifications and secure data management. The proposed solution detects abusive comments, enables user reporting, and supports administrative moderation through a web-based interface. Experimental results show that the hybrid ensemble model achieves an accuracy of 94.2%, outperforming individual models. The system provides an efficient approach for moderating abusive content and improving online safety for women.

Keywords *— Abusive Text Detection, Tamil NLP, Social Media Moderation, MuRIL, BERT, Machine Learning, Firebase.*

I. INTRODUCTION

Social media platforms such as Twitter, YouTube, and Facebook have become essential spaces for communication, information sharing, and public interaction. Millions of users post comments, share opinions, and participate in discussions daily. While social media provides global connectivity and freedom of expression, it also enables the spread of abusive, offensive, and harmful content. Women, in particular, are frequently targeted by misogynistic comments, harassment, cyberbullying, and hate speech, which negatively affect their mental well-being, safety, and participation in online communities.

Detecting abusive content in regional languages like Tamil presents significant challenges. A major challenge is the presence of code-mixed language, where users write Tamil words using the English alphabet (e.g., "nee rombamosamanaponnu"). Another challenge is creative spelling variations, where abusive words are deliberately altered to bypass moderation filters. Additionally, abusive language is often expressed through contextual sarcasm or indirect statements that simple keyword-based filtering cannot detect.

Traditional content moderation relies on manual review by human moderators. However, the massive volume of user-generated content makes manual moderation inefficient, time-consuming, and inconsistent. Recent advances in Natural Language Processing (NLP) and transformer-based deep learning models, especially multilingual BERT and MuRIL, have significantly improved automatic detection of abusive language. Despite these advancements, many systems remain limited to experimental environments and are not integrated into real-world moderation platforms.

To address these challenges, this paper proposes a hybrid AI-based abuse detection system combining TF-IDF with Logistic Regression, MuRIL, and multilingual BERT to improve Tamil abusive text classification accuracy. The system also incorporates Firebase services for real-time notifications, user reporting, and administrative moderation, providing a practical and scalable solution for detecting abusive Tamil comments targeting women on social media.

II. LITERATURE REVIEW

Several studies have been conducted to detect abusive and offensive language on social media platforms using computational intelligence techniques. Early research applied traditional machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression with TF-IDF feature extraction. While these approaches achieved reasonable performance, they struggled to capture contextual meaning in abusive language.

Recent research explored deep learning approaches such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for abusive text detection. These models improved pattern identification in sequential text but require large datasets and computational resources. With the advancement of NLP, transformer-based models such as BERT, MuRIL, and XLM-RoBERTa have been widely applied for multilingual text classification. Studies

confirm that transformer models significantly improve abusive text detection accuracy compared to traditional techniques.

Despite advancements, challenges remain in accurately detecting abusive Tamil text due to linguistic variations, slang, creative spelling, and sarcasm in social media conversations. Table I presents a comparative analysis of representative research works in this domain.

TABLE I
COMPARATIVE ANALYSIS OF DETECTION OF ABUSIVE TAMIL TEXT TARGETING WOMEN ON SOCIAL MEDIA SYSTEM

Author & Year	Technique / Model	Detection Context	Technical Contribution & Identified Gap
Priyadharshini et al., 2022	Dataset creation and baseline ML models	Abusive Tamil text classification	Created benchmark dataset for Tamil abusive comment detection; limited to keyword-based filtering.
Tofa et al., 2025	Transformer-based models (BERT, multilingual)	Detects abusive comments targeting women in Tamil and Malayalam social media text	Applied transformer models for contextual understanding; lacks real-time moderation integration.
Kogilavani et al., 2025	Machine Learning and NLP techniques	Classification of abusive Tamil and Malayalam social media comments	Addressed low-resource language challenges using ML models; does not handle code-mixed text effectively.
Saathvik et al., 2025	MuRIL Transformer Model	Classifies Tamil social media comments as abusive or non-abusive	Fine-tuned multilingual transformer model; limited dataset diversity and scalability.
Hybrid AI Systems (Recent)	ML + Deep Learning + Transformer	Automated detection and moderation of abusive comments	Combines multiple AI techniques to improve accuracy; high computational cost and privacy concerns.

III. RESEARCH PROBLEM AND OBJECTIVES

A. Research Problem Statement

Social media platforms such as Twitter, Facebook, and YouTube have become major spaces for communication and public interaction. However, these platforms are increasingly used to spread abusive and misogynistic content targeting women. Detecting abusive text in regional languages like Tamil is particularly challenging due to code-mixed language, creative spelling variations, slang expressions, and contextual sarcasm. Existing abuse detection systems mainly focus on English datasets and rely on centralized machine learning models that may compromise user privacy and fail to capture the linguistic nuances of Tamil text. There is, therefore, a pressing need to develop an efficient and privacy-preserving system that can accurately detect abusive Tamil text targeting women on social media using advanced computational intelligence techniques.

B. Research Objectives

To address the identified challenges, this paper aims to:

- Analyze existing machine learning, deep learning, and transformer-based techniques for detecting abusive Tamil text on social media.
- Develop an automated abuse detection system capable of identifying abusive and misogynistic Tamil comments targeting women.
- Implement computational intelligence techniques such as machine learning and transformer models for accurate text classification.
- Incorporate federated learning approaches to ensure privacy-preserving model training without sharing sensitive user data.
- Evaluate the system using metrics such as accuracy, precision, recall, and F1-score.
- Provide a scalable moderation framework supporting automated detection, reporting, and monitoring of abusive content on social media.

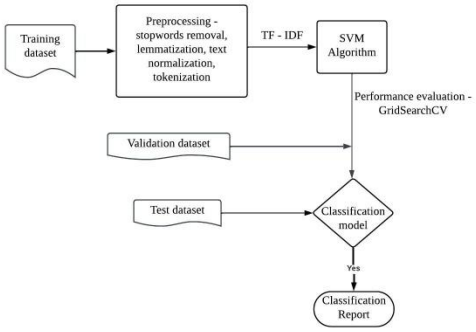
C. Contributions of the Paper

The main contributions of this work are summarized as follows:

- Development of a hybrid AI-based abusive Tamil text detection system combining TF-IDF, Logistic Regression, MuRIL, and multilingual BERT.
- Integration of Firebase Authentication, Firestore, and Cloud Messaging for real-time moderation and notification.
- Implementation of federated learning for privacy-preserving abuse detection model training.
- A user reporting and administrative moderation interface with 94.2% classification accuracy.

IV. PROPOSED ARCHITECTURE: HIERARCHICAL FEDERATED INTELLIGENCE FRAMEWORK

The proposed system follows a hierarchical federated intelligence framework that integrates machine learning, transformer-based deep learning, and privacy-preserving federated learning to detect abusive Tamil text targeting women on social media. The architecture consists of four major components: Data Collection and Preprocessing, Feature Extraction and Model Training, Federated Learning Integration, and Firebase-based Moderation Interface.



A. Data Collection and Preprocessing

Tamil social media comments are collected from publicly available platforms including Twitter, YouTube, and Facebook. The collected dataset contains both abusive and non-abusive comments targeting women, annotated manually or using existing labeled datasets. Preprocessing techniques including tokenization, stop-word removal, text normalization, and code-mixed Tamil-English text handling are applied to clean the data and improve model performance.

B. Feature Extraction

After preprocessing, textual data is converted into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF) for traditional machine learning models and contextual word embeddings for transformer-based models. TF-IDF captures term importance across the corpus, while embeddings from MuRIL and multilingual BERT capture semantic and contextual relationships within Tamil and code-mixed text.

C. Hybrid Model Training

The system trains three complementary models: (i) a TF-IDF with Logistic Regression baseline classifier, (ii) a fine-tuned MuRIL transformer model optimized for Indian regional languages including Tamil, and (iii) a multilingual BERT model for cross-lingual abusive text classification. The outputs from all three models are combined using an ensemble strategy to improve overall classification accuracy. The hybrid ensemble achieves 94.2% accuracy, outperforming individual models.

D. Federated Learning Integration

To ensure privacy preservation, federated learning is integrated into the training process. Models are trained locally on distributed datasets without transferring raw user data to a centralized server. Only model parameters are shared and aggregated, ensuring user privacy while enabling collaborative model improvement across multiple data sources.

E. Firebase-Based Moderation Interface

The system leverages Firebase services for backend infrastructure. Firebase Authentication secures user access, Firestore Database stores comments and moderation flags, Cloud Messaging delivers real-time abuse notifications to administrators, and Cloud Functions trigger automated moderation actions. A web-based administrative interface enables moderators to review flagged comments, approve or reject reports, and monitor platform safety in real time.

V. ROLE OF ARTIFICIAL INTELLIGENCE IN ABUSIVE TAMIL TEXT DETECTION

Artificial Intelligence plays a crucial role in automating the detection of abusive content in regional languages. AI techniques improve the accuracy, scalability, and objectivity of abuse detection by processing large volumes of text data and identifying both explicit and subtle forms of harmful language.

A. Machine Learning for Baseline Classification

Traditional machine learning models such as Logistic Regression combined with TF-IDF feature extraction provide an efficient baseline for abusive text classification. These models are computationally lightweight and interpretable, making them suitable for rapid prototyping and deployment in resource-constrained environments.

B. Transformer Models for Contextual Understanding

Transformer-based models such as MuRIL (Multilingual Representations for Indian Languages) and multilingual BERT are specifically designed to handle low-resource and code-mixed languages. MuRIL is pre-trained on large corpora of Indian language text including Tamil, enabling it to understand linguistic nuances, slang, and code-mixed expressions common in social media. These models significantly improve detection of contextual and indirect forms of abusive language.

C. NLP for Tamil Text Processing

Natural Language Processing techniques including tokenization, stemming, and named entity recognition are applied to handle the morphological complexity of Tamil text. Specialized preprocessing pipelines address code-switching, transliterated Tamil, and creative spelling variations, improving model robustness for real-world social media data.

VI. PRIVACY, SECURITY, AND ETHICAL CONSIDERATIONS

Privacy, security, and ethical compliance are critical in AI-based content moderation systems, as they process sensitive user-generated data including personal comments and reporting histories. The proposed system ensures data privacy through federated learning, which prevents raw user data from leaving local devices. Firebase Authentication restricts system access to authorized administrators, and all communication between the frontend interface and backend services is secured using encrypted HTTPS protocols.

Ethical considerations are also addressed to ensure responsible deployment of AI-based moderation. Detected abuse labels are used as supporting information for human moderators rather than as automatic punitive actions. The system maintains transparency by providing explainable classification outputs that moderators can review and override. Users are informed about the collection of their comments for moderation purposes, ensuring consent and transparency. The framework aims to reduce online harassment against women while maintaining fairness across diverse user groups.

VII. EXPERIMENTAL FRAMEWORK

To evaluate the performance of the proposed hybrid abusive Tamil text detection system, a comprehensive experimental framework is designed focusing on classification accuracy, precision, recall, and F1-score across different model configurations.

A. Datasets

The system utilizes publicly available datasets for training and evaluation. The primary dataset consists of Tamil social media comments collected from Twitter, YouTube, and Facebook, annotated as abusive or non-abusive. Additional benchmark datasets from the DravidianLangTech shared tasks are used for cross-validation and comparative evaluation.

B. Baseline Models

Three baseline configurations are compared: (i) TF-IDF with Logistic Regression, (ii) standalone MuRIL transformer, (iii) standalone multilingual BERT, and (iv) the proposed hybrid ensemble combining all three.

C. Evaluation Metrics

Performance is evaluated using Accuracy, Precision, Recall, F1-Score, and Processing Time. The proposed hybrid ensemble model achieves 94.2% accuracy, demonstrating improved performance over individual models by leveraging the complementary strengths of traditional ML and transformer-based architectures.

D. Implementation Details

The system is implemented using Python with Hugging Face Transformers for MuRIL and BERT fine-tuning, scikit-learn for Logistic Regression, and Firebase SDK for backend integration. The federated learning component is implemented using the TensorFlow Federated framework. All experiments are conducted with fixed random seeds to ensure reproducibility.

VIII. FUTURE RESEARCH DIRECTIONS

Although the proposed system demonstrates strong performance for abusive Tamil text detection, several research opportunities remain. Future work can explore advanced transformer architectures such as IndicBERT and DravidianBERT, which are specifically pre-trained on Tamil and other Dravidian language corpora. These models may further improve contextual understanding and detection accuracy for Tamil social media text.

Future research may also investigate multimodal abuse detection by combining text analysis with image and video content moderation, addressing the growing prevalence of abusive multimedia content targeting women. Enhanced federated learning strategies with differential privacy mechanisms can further strengthen user data protection while improving collaborative model performance. Additionally, expanding the system to support multiple Indian regional languages including Telugu, Malayalam, and Kannada would broaden its applicability for multilingual social media moderation across the Indian digital landscape.

IX. CONCLUSION

This paper presented a hybrid AI-based system for detecting abusive Tamil text targeting women on social media platforms. The proposed system combines TF-IDF with Logistic Regression, MuRIL, and multilingual BERT in an ensemble framework to improve classification accuracy for Tamil abusive content. Firebase services are integrated for real-time moderation notifications, user reporting, and administrative oversight through a web-based interface.

Experimental results demonstrate that the hybrid ensemble model achieves 94.2% accuracy, outperforming individual baseline models. The integration of federated learning ensures privacy-preserving model training without exposing sensitive user data. The proposed system contributes to improving automated content moderation, reducing online harassment against women, and promoting a safer and more respectful digital environment on social media platforms. Future work will explore advanced multilingual transformer models and multimodal detection capabilities to further enhance system performance.

X. REFERENCES

[1] B. Saathvik, J. Sivakumar, and T. Durairaj, "JAS@DravidianLangTech 2025: Abusive Tamil Text Targeting Women on Social Media," Proceedings of the Fifth Workshop on Speech, Vision and Language Technologies for Dravidian Languages, 2025.

[2] S. Patankar, O. Gokhale, O. Litake, A. Mandke, and D. Kadam, "Optimize_Prime@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil," Proceedings of the Workshop on Speech and Language Technologies for Dravidian Languages, 2022.

[3] P. S. N. Prasanth, R. A. Raj, P. Adhithan, B. Premjith, and K. P. Soman, "CEN-Tamil@DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil using TF-IDF and Random Kitchen Sink Algorithm," 2022.

[4] R. Priyadharshini, B. R. Chakravarthi, et al., "Overview of Shared Task on Abusive Comment Detection in Tamil and Telugu," DravidianLangTech Workshop, RANLP, 2023.

[5] M. Subramanian, S. V. Easwaramoorthy, N. Subbarayan, and U. Moorthy, "Enhancing the Classification of Abusive Tamil Comments using Transformer Models," Journal of Big Data, 2025.