Solvers@LT-EDI-2025: Caste and Migration Hate Speech Detection in Tamil-English Code-Mixed Text

Mohanapriya K T[1], Anirudh Sriram K S[1], Devasri A[1], Bharath P[1], Ananthakumar S[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Hate speech detection in low-resource languages such as Tamil presents significant challenges due to linguistic complexity, limited annotated data, and the sociocultural sensitivity of the subject matter. This study focuses on identifying caste- and migration-related hate speech in Tamil social media texts, as part of the LT-EDI@LDK 2025 Shared Task. The dataset used consists of 5,512 training instances and 787 development instances, annotated for binary classification into caste/migration-related and non-caste/migration-related hate speech. We employ a range of models, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based architectures such as BERT and multilingual BERT (mBERT). A central focus of this work is evaluating model performance using macro F1-score, which provides a balanced assessment across this imbalanced dataset. Experimental results demonstrate that transformer-based models, particularly mBERT, significantly outperform traditional approaches by effectively capturing the contextual and implicit nature of hate speech. This research underscores the importance of culturally informed NLP solutions for fostering safer online environments in underrepresented linguistic communities such as Tamil.

1 Introduction

Detecting hate speech in low-resource languages such as Tamil is a complex and crucial task, especially in the context of caste- and migration-related discrimination. Tamil, predominantly spoken in Tamil Nadu (India), Sri Lanka, and among diasporic communities, is rich in cultural and linguistic diversity, which poses unique challenges to Natural Language Processing (NLP). In recent years, the spread of hate speech targeting caste and migrant communities has escalated on social media platforms, demanding robust automatic detection systems that are socially aware and ethically grounded.

The identification of such harmful content is complicated by the nuanced ways in which caste and migration are discussed, often involving implicit language, sarcasm, and regional idioms. Additionally, Tamil's morphological richness, the scarcity of annotated corpora, and the limited availability of linguistic tools make it difficult to develop high-performance hate speech classifiers. To address these issues, the LT-EDI@LDK 2025 Shared Task released a manually annotated dataset in Tamil, categorizing instances as either caste/migration-related hate speech or non-caste/migration-related hate speech (Rajiakodi et al., 2024).

In this study, we evaluate the performance of several machine learning and deep learning models—specifically, Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based architectures such as BERT and multilingual BERT (mBERT)—for the task of hate speech classification (Vaswani et al., 2017; Devlin et al., 2019). Given the class imbalance and the sensitive nature of the content, we use macro F1-score as the primary evaluation metric. Our findings show that transformer-based models, particularly mBERT, perform significantly better at capturing contextual cues and implicit hate. This work contributes to the growing body of research aimed at improving online safety in underrepresented linguistic communities and emphasizes the importance of culturally and ethically grounded NLP approaches to hate speech detection.

2 Literature Survey

Hate speech detection is an increasingly important task in natural language processing (NLP), particularly with the rise of social media platforms where offensive and abusive content is frequently encountered. While substantial progress has been made in hate speech detection for high-resource

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 100–104

languages like English, the task remains underdeveloped for low-resource languages such as Tamil. Early shared tasks like HASOC and OffensEval laid foundational work in this domain (Zampieri et al., 2019; Mandl et al., 2020). Tamil poses unique challenges due to its rich morphology, code-mixed usage with English, and the cultural sensitivity of topics like caste and migration. The scarcity of large, annotated datasets in Tamil further complicates model development for hate speech classification.

Recent shared tasks such as the LT-EDI (Language Technology for Equality, Diversity, and Inclusion) series have helped bring attention to the issue, offering benchmark datasets and encouraging the development of hate speech detection systems specifically for Tamil and other Dravidian languages (Rani et al., 2022). These initiatives have laid the groundwork for evaluating traditional machine learning, deep learning, and transformer-based approaches on nuanced categories such as caste and migration hate speech.

2.1 Hate Speech Detection in Tamil

Initial approaches to Tamil hate speech detection involved traditional machine learning techniques, with Support Vector Machines (SVM) and Naive Bayes being among the earliest models. These methods utilized hand-crafted features like n-grams, part-of-speech tags, and TF-IDF vectors. Although simple and interpretable, these models often struggled to capture the semantic depth and informal variations in Tamil text, especially in code-mixed settings.

Deep learning methods, such as Convolutional Neural Networks (CNNs), were later introduced to overcome the limitations of feature engineering. CNNs have proven effective in capturing local dependencies in text, particularly through their use of convolutional filters across embedding sequences. Their success in image recognition tasks translated well to text classification by modeling syntactic patterns in sentence fragments.

More recently, transformer-based models such as BERT and multilingual BERT (mBERT) have transformed the field of NLP. These models employ self-attention mechanisms to capture contextual semantics across entire sentences, making them well-suited for detecting implicit hate speech and nuanced expressions. For Tamil, mBERT's multilingual training across 104 languages has proven particularly effective in low-resource scenarios by transferring knowledge from related high-resource languages (Devlin et al., 2019).

Vasantharajan and Thayasivam (2021) demonstrated the effectiveness of mBERT in classifying Tamil code-mixed YouTube comments. Similarly, Benhur and Sivanraju (2021) applied mBERT to the Tanglish dataset, achieving competitive results. These studies highlight the advantage of using transformer-based models in multilingual and culturally diverse contexts.

2.2 Caste and Migration Hate Speech

Detecting hate speech related to caste and migration in Tamil presents unique challenges due to the deep cultural and historical roots of these social structures. The language used in such contexts often includes sarcasm, indirect references, and culturally embedded terms, which are difficult to classify using traditional models. Studies have shown that transformer models like mBERT and BERT are better suited to handle such implicit and contextual expressions of hate speech.

For instance, Alam et al. (2024) conducted a comparative analysis of various transformer-based models on caste- and migration-related Tamil hate speech data, concluding that mBERT consistently delivered the best performance, with a macro F1-score of 0.80. Their work validates the use of multilingual transformers for sensitive and domain-specific tasks, particularly in underrepresented languages like Tamil.

2.3 Challenges in Hate Speech Detection for Tamil

Despite progress, several challenges continue to hinder the development of robust hate speech detection systems for Tamil:

Data Scarcity: The lack of large-scale, annotated datasets tailored to caste and migration hate speech remains a major barrier.

Code-Switching: Frequent code-mixing between Tamil and English on social media creates ambiguity for monolingual models.

Cultural Nuance: Tamil expressions of hate often include regional dialects, idioms, and sarcasm that require culturally aware annotation and modeling.

Informality: The informal and noisy nature of social media content makes tokenization, POS tagging, and syntactic parsing more difficult.

These factors collectively call for the use of sophisticated models like BERT and mBERT, which

can encode complex context and benefit from multilingual pretraining. However, even these models are constrained by the quality and quantity of labeled data available for fine-tuning.

2.4 Transformer Models and Advances

Transformer models such as BERT, multilingual BERT (mBERT), and MuRIL have revolutionized NLP, particularly for tasks in low-resource settings. Introduced by (Vaswani et al., 2017), the Transformer architecture enables parallel processing and better captures long-range dependencies, which are essential for understanding context in hate speech.

In the context of Tamil, mBERT and MuRIL have shown superior results due to their multilingual training on large-scale corpora. These models can transfer knowledge from high-resource languages to Tamil, thereby compensating for the lack of labeled data. IndicBERT, which is trained on 12 Indian languages, including Tamil, has also been explored for its lightweight architecture and adaptability to resource-constrained environments (Kakwani et al., 2020).

Despite these innovations, there remains a pressing need for more annotated datasets, domain-specific pretraining, and culturally sensitive modeling approaches (Hendrycks et al., 2021; Touvron et al., 2023) to further improve hate speech detection systems in Tamil.

3 Materials and Methods

3.1 Dataset Description

These texts are code-mixed with English and centered around themes of caste and migration, annotated for hate speech detection (Rajiakodi et al., 2024). The dataset is divided into three subsets: a training set with 8,042 samples, a validation set (dev.csv) with 1,006 samples, and a test set (test.csv) containing 1,001 samples. Each sample is labeled as either "Hate Speech," "Non-Hate Speech," or "Offensive but not Hate," allowing multi-class classification.

The class distribution in the training set is imbalanced, with "Non-Hate Speech" forming the majority, followed by a smaller proportion of "Hate Speech" and "Offensive" samples. To handle this imbalance, oversampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and random duplication were employed during training. This annotation and structure provide a nuanced foundation for detecting subtle and explicit hate expressions related to caste and migration.

3.2 Pre-processing and Feature Extraction

Due to the informal and code-mixed nature of the dataset, preprocessing was a critical step to improve model performance. The following techniques were employed:

Text Normalization: All text was lowercased, and punctuation, special characters, and elongated words were normalized. Hashtags were split when meaningful.

Tokenization: Tokenization was performed using Hugging Face tokenizers for BERT and mBERT, while standard NLP tokenizers were used for SVM and CNN.

Noise Removal: URLs, mentions (@username), emojis, and numbers were removed to reduce noise in social media-style texts.

Code-Mixing Handling: To address Tamil-English mixed content, language identification was applied. When possible, transliteration was used to normalize Tamil content written in Roman script.

Stopword Removal: Tamil and English stopword lists were used to eliminate non-informative tokens, mainly for traditional feature-based models like SVM.

Feature Extraction varied depending on the model type:

SVM: Employed Bag-of-Words (BoW) and TF-IDF vector representations to convert text into numerical features.

CNN: Used pretrained FastText Tamil word embeddings (Bojanowski et al., 2017) to capture local semantic and syntactic patterns.

BERT and mBERT: Fine-tuned transformer models with contextual embeddings that capture deep semantic features. Sentence embeddings generated by these models provided rich context representations, crucial for identifying implicit hate speech. Prior research has shown the utility of such embeddings in low-resource settings (Bouraoui et al., 2020).

3.3 Proposed Classifiers

Four classifiers were developed and evaluated, categorized as follows:

Traditional Model:

Support Vector Machine (SVM): Selected for its robustness in high-dimensional spaces and interpretability. TF-IDF features provided the best

performance among traditional vector representations.

Deep Learning Model:

Convolutional Neural Network (CNN): Designed to learn local n-gram features from FastText embeddings. It effectively captured spatial hierarchies in the text data.

Transformer-Based Models:

BERT: Fine-tuned on the hate speech dataset to leverage deep contextual embeddings.

Multilingual BERT (mBERT): Trained on a large multilingual corpus, mBERT was particularly effective in handling code-mixed Tamil-English text and showed robust generalization across hate speech categories.

4 Results and Discussion

The hate speech detection experiments using Tamil-English code-mixed data demonstrate that transformer-based models, particularly Multilingual BERT (mBERT), achieve the highest performance compared to traditional and shallow learning models. mBERT excels in capturing complex, context-rich semantic structures within code-switched and informal text. Deep learning models such as CNN and BiLSTM performed moderately well, capturing local and sequential patterns, respectively. Traditional models such as SVM, Logistic Regression, and KNN were able to handle basic binary discrimination but struggled with fine-grained category separation.

4.1 Performance Metrics

All models were evaluated on the development set using standard metrics: Accuracy, Precision, Recall, F1-score, and Macro-Averaged scores to ensure balance across class imbalances. The confusion matrices revealed that the major source of error involved misclassifications between the *Hate Speech* and *Offensive* classes, suggesting semantic overlap and subtle contextual differences.

Table 1: Performance Comparison of Top Models

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
CNN	79.0	77.5	74.7	75.9
BERT	80.0	79.0	78.0	78.0
mBERT	**80.0**	**80.0**	**78.0**	**79.0**

Among the tested models, mBERT consistently achieved the best macro-averaged F1-score (79%), indicating reliable performance across both majority and minority classes. Its ability to handle code-mixed inputs effectively distinguishes it in multilingual contexts. Compared to BERT, mBERT displayed marginal but consistent improvements in recall and macro-F1, especially for underrepresented hate-related classes.

4.2 Limitations

Despite promising results, several limitations were encountered:

Class Imbalance: The dataset exhibits a skewed distribution with fewer samples in the "Hate Speech" and "Offensive" categories. This imbalance led to minor bias in model predictions toward the majority class. Though techniques like SMOTE were employed, they couldn't fully replicate the diversity and complexity of real hateful content. Research shows that data augmentation using back-translation or paraphrasing can enhance minority class learning (Barro et al., 2023).

Computational Cost: Transformer-based models like mBERT require significant computational resources and training time. This makes real-time or edge deployment challenging. Future work may focus on model distillation or lighter variants like DistilBERT or ALBERT to balance performance and efficiency.

Code-Mixing Complexity: Many samples contain informal Tamil-English blends or slang that are difficult to tokenize or embed correctly. This causes confusion particularly between offensive and hate speech categories. More robust language identification and translation pipelines may mitigate this issue.

Limited Annotated Data: The lack of large-scale annotated code-mixed Tamil datasets restricts model generalization. Introducing active learning and human-in-the-loop feedback mechanisms could help create a continuously evolving and more representative dataset.

5 Conclusion

In this study, we addressed the challenge of hate speech detection in Tamil-English code-mixed social media text, a complex task due to linguistic diversity, informal language, and limited annotated resources. We implemented and evaluated a classification pipeline using a combination of traditional and modern techniques, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and transformer-based models such

as BERT and multilingual BERT (mBERT).

Our experiments demonstrated that transformer-based models, particularly mBERT, achieved the best overall performance, effectively capturing contextual and semantic nuances in code-mixed text. CNNs provided competitive results by modeling local syntactic patterns, while SVM served as a strong baseline for traditional feature-based learning in low-resource conditions.

Despite the encouraging results, several challenges remain unresolved, including class imbalance, informal code-switching, and the limited size of annotated datasets. Future work may focus on incorporating domain-adaptive pretraining, leveraging data augmentation strategies, and employing multilingual knowledge integration to further improve model robustness. This research contributes to the development of effective NLP solutions for underrepresented South Asian languages and supports broader efforts to ensure safer, more inclusive online spaces.

6 Project Repository

The full source code for this project is available on GitHub: Bharath

References

Sarah Barro, Marcos Zampieri, and Ahmed Abdelali. 2023. Investigating the impact of data augmentation techniques for low-resource hate speech detection. In *Proceedings of ACL 2023*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. In *Proceedings of AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Dan Hendrycks, Collin Burns, Steven Basart, and et al. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Divyanshu Kakwani, Raghav Aggarwal, Siddhant Garg, and et al. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. *Findings of EMNLP*.

Thomas Mandl, Sandip Modha, Pooja Rani, and et al. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. In *Working Notes of FIRE 2020*.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Pooja Rani, Thomas Mandl, Sandip Modha, and et al. 2022. Hasoc 2022: Hate speech and offensive content identification in indic languages. In *Working Notes of FIRE 2022*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, and et al. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, pages 1415–1420.