girlsteam@LT-EDI-2025: Caste/Migration based hate speech Detection.

Towshin Hossain Tushi, Walisa Alam, Rehenuma Ilman, Samia Rahman
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u2004086, u2004015, u2004079, u1904022}@student.cuet.ac.bd

Abstract

The proliferation of caste- and migration-based hate speech on social media poses a significant challenge, particularly in low-resource languages like Tamil. This paper presents our approach to the LT-EDI@ACL 2025 shared task, addressing this issue through a hybrid transformer-based framework. We explore a range of Machine Learning (ML), Deep Learning (DL), and multilingual transformer models, culminating in a novel m-BERT+BiLSTM hybrid architecture. This model integrates contextual embeddings from m-BERT with lexical features from TF-IDF and FastText, feeding the enriched representations into a BiLSTM to capture bidirectional semantic dependencies. Empirical results demonstrate the superiority of this hybrid architecture, achieving a macro-F1 score of 0.76 on the test set and surpassing the performance of standalone models such as MuRIL and IndicBERT. These results affirm the effectiveness of hybrid multilingual models for hate speech detection in low-resource and culturally complex linguistic settings.

Keywords: Hate Speech Detection, Tamil, Code-Mixed Text, mBERT, BiLSTM, Caste, Tf-IDF.

1 Introduction

Hate speech encompasses expressions—verbal, written, or behavioral—that incite hostility or dehumanize individuals based on identity markers such as race, caste, gender, religion, or migration status. It reinforces prejudice and systemic discrimination, undermining individual dignity and social cohesion. Caste- and migration-based hate speech, in particular, reflects deep-rooted structural inequalities conveyed through derogatory narratives (Bhatt et al., 2022).

The rise of social media has amplified such harmful content, especially against marginalized communities in multilingual countries like India (Sharif et al., 2021). In this context, hate speech frequently appears in Tamil, English, and Tanglish—a code-mixed variant of Tamil in Latin script—posing unique challenges for automated detection. Tamil, a classical Dravidian language spoken by over 70 million people (Chakravarthi and Raja, 2020), presents significant challenges for Natural Language Processing (NLP) due to its complex morphology and limited annotated resources. Their workshop paper provided us an opportunity to engage with these challenges in processing mixed-up languages and to leverage our work (Rajiakodi et al., 2025). Recent advances in multilingual NLP have produced models tailored to Indian languages, such as IndicBERT, MuRIL, and mBERT (Khanuja et al., 2021). These transformer-based models, alongside machine learning (ML) and deep learning (DL) methods, form a robust foundation for tackling hate speech detection in such contexts. For the LT-EDI 2025 shared task, we propose a comprehensive system for identifying caste- and migration-related hate speech in Tamil social media. Our main objective was-

- To develop a robust multilingual system for detecting caste and migration related hate speech in Tamil social media by leveraging Tamil, English, and code-mixed (Tanglish) text.
- To propose a hybrid architecture that integrates transformer-based models (IndicBERT, mBERT, and MuRIL) with a BiLSTM network, combining contextual embeddings with sequential modeling to enhance classification performance in multilingual and code-mixed settings.

Our code, developed for this shared task can be accessed at [1]

[1] https://github.com/walisa810/Shared_Task_DravidianLangTech

2 Related work

Hate speech detection has gained growing attention, especially in multilingual and socio-culturally nuanced contexts. However, research focused on Tamil, particularly caste- and migration-related hate speech, remains sparse. Early work predominantly used traditional machine learning methods. For instance, Hossain et al. (2022) applied Logistic Regression to abusive Tamil text, and Bhimani et al. (2021) addressed caste and religion-based hate using similar approaches. In SemEval-2019 Task 5, Basile et al. (2019) and Almatarneh and Gamallo (2019) employed TF-IDF and lexicon-based features for multilingual hate speech detection. Sachdeva et al. (2021) used a Random Forest classifier for general hate speech classification. More recent studies have leveraged deep learning and contextual embeddings. Sharif and Hoque (2021) proposed an ensemble of CNN, BiLSTM, and GRU for Bengali hate speech, while Farooqi et al. (2021) combined Indic-BERT, XLM-R, and mBERT for code-mixed hate speech using conversational context. Romero-Vega et al. (2021) utilized SVM to detect xenophobic hate in Spanish tweets. Sajlan (2021) offered a qualitative analysis of caste-based hate speech, underscoring the need for robust computational approaches in this underrepresented area.

To address the complexities of caste and migration hate speech in Tamil, we propose a hybrid mBERT+BiLSTM model. This architecture combines the contextual understanding of mBERT with the sequential learning of BiLSTM to improve the detection accuracy in this underexplored domain.

3 Task and Dataset Description

This shared task focuses on **Caste and Migration Hate Speech Detection**. The objective of the task is to develop automatic classification models that can analyze social media texts, with a specific emphasis on content related to caste and migration. The task organizers provided a dataset containing social media posts in a mix of languages: The dataset comprises text in various language forms, including English (e.g., *"Mumbai Bangalore la 80 percentage outsiders"*), code-mixed English-Tamil, Tanglish (Tamil written in the Latin script), and pure Tamil.

Each entry in the dataset consists of the following:

- `id` – A unique identifier for the text
- `text` – The content of the text
- `label` – A binary class indicating presence of hate speech

The classification labels are defined as follows:

- **0** – Non Caste/Migration-related Hate Speech.
- **1** – Caste/Migration-related Hate Speech

The data set was segmented into training, development, and test data subsets to help with the thorough analysis and to facilitate model training. The class distribution is summarized in Table 1:

Table 1: Class distribution across datasets

Class	Train	Dev	Test
Non-Hate Speech (0)	3415	485	970
Hate Speech (1)	2097	302	606

4 Methodology

The growing prevalence of hate speech on social media has emerged as a critical issue, often targeting specific communities. In this section, we summarize the methods and strategies proposed to tackle the challenges highlighted earlier. Based on a thorough analysis, our research advocates for the adoption of a transformer-based model, using mBERT in combination with a BiLSTM architecture(Aodhora et al., 2025). Figure 2: presents a clear visualization of our methodology, illustrating the key steps in our approach.

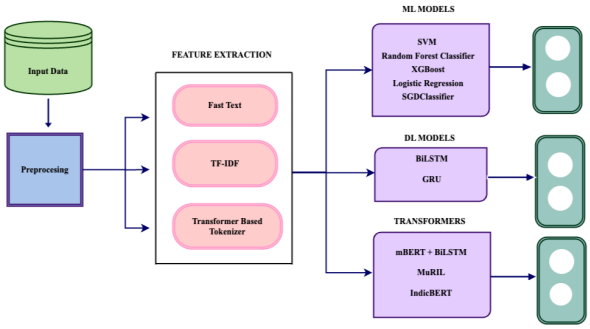

Figure 1: An abstract view of our methodology.

4.1 Data Preprocessing

The dataset provided by the problem organizing committee contains a significant amount of irrelevant and noisy content, including code-mixed

text (Ponnusamy et al., 2024). During preprocessing, we systematically cleaned the data by removing noise such as hyperlinks, emojis, punctuation, alphanumeric clutter, and special characters (e.g., slashes, brackets, and ampersands). In addition, all text was converted to lowercase to maintain consistency and improve data quality.

4.2 Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) was applied for feature extraction. In this approach, weights are assigned to words based on their frequency within a document and across the corpus, allowing the most informative terms to be identified. Furthermore, pre-trained Tamil FastText embeddings were utilized, which generate 300-dimensional word vectors by incorporating subword information through n-grams, thereby allowing the capture of semantic word relationships (Bojanowski et al., 2017).

For transformer-based models, we used model-specific tokenizers from the Hugging Face library to handle appropriate tokenization and padding, ensuring compatibility with each model's input requirements.

4.3 Model Building

In our research, we explored various machine learning (ML), deep learning (DL), and transformer-based models.

4.3.1 ML models

As a preliminary step, we evaluated the performance of several classical machine learning algorithms in our data set, including logistic regression, support vector machine (SVM), random forest, XGBoost, and stochastic gradient descent (SGD). These models served as baseline classifiers to assess the separability of the dataset and the inherent difficulty of the task. Overall, the models achieved moderate F1-scores ranging from 0.60 to 0.65, with SVM and XGBoost slightly outperforming others. These initial experiments provided a foundational benchmark for more advanced deep learning-based approaches.

4.3.2 DL models

Building on classical baselines, we developed an advanced deep learning pipeline that fused TF-IDF, FastText. These hybrid inputs were processed through sequential architectures, specifically Bidirectional LSTM and GRU networks, to capture long-range dependencies and contextual semantics in Tamil-English code-mixed hate speech. The use of BiLSTM was especially enabling it to comprehend contextual cues from both preceding and succeeding tokens—an essential capability for disambiguating morphologically rich code-mixed expressions. The BiLSTM model achieved a macro-averaged F1-score of 0.67, with the BiGRU yielding similar results, indicating consistent performance across architectures. These deep models outperformed classical approaches in recall and semantic awareness, establishing a strong benchmark for subsequent transformer-based exploration.

4.3.3 Transformer-based Models

The transformer-based models we applied include MuRIL, mBERT (Yu et al., 2024), and IndicBERT (Kakwani et al., 2020). These models were fine-tuned using their respective transformer-specific tokenizers to efficiently handle multilingual text. Table 2 presents the hyperparameters used across these models.

Among all, the mBERT + BiLSTM hybrid model demonstrated the best performance. In this architecture, we integrate the contextual strength of multilingual BERT (mBERT) with the sequential learning capability of a Bidirectional LSTM (BiLSTM). mBERT, pre-trained on over 100 languages, generates high-dimensional contextual embeddings (logits) from the input text, making it highly suitable for code-mixed and multilingual data such as Tamil-English (Tanglish) content.

These embeddings are concatenated with lexical features obtained from TF-IDF and FastText, forming a rich and diverse feature representation. This combined feature vector is reshaped and passed into a stacked BiLSTM network, which includes dropout and batch normalization layers for regularization and to prevent overfitting.We extracted logits as output features from all three models—m-BERT, MuRIL, and Indic-BERT. Token lengths of 128 and 512 were used with a batch size of 32 during feature extraction.

BiLSTM is particularly beneficial in this setup as it processes input sequences in both forward and backward directions, enabling it to capture long-range dependencies and contextual relationships that might be missed by unidirectional models. This is especially important for noisy, user-generated content, where the order and surrounding context of words can significantly impact meaning.

The fusion of mBERT's deep multilingual con-

textual embeddings with BiLSTM's sequential modeling allowed the architecture to effectively learn nuanced patterns in code-mixed text. This contributed to its superior performance, demonstrating its robustness and adaptability in multilingual hate speech detection tasks.

5 Result

In this section, we compare the performance of various machine learning (ML) and deep learning (DL) models. The effectiveness of each model is primarily evaluated using the macro F1-score. The hyperparameters for the DL models were manually fine-tuned based on their performance on the validation dataset. The sum of the precision (P), recall (R) and macro-F1 (MF1) scores for each model in the test set is presented in the Table 3

Table 2: The hyperparameters in BiLSTM model

Hyperparameters	**Values**
Input Shape	(1, 768)
Units	[256, 128, 64]
Dropout Rate	0.3
Optimizer	Adam
Loss Function	Binary Crossentropy
Batch Size	32
Epochs	100
Early Stopping	Patience = 5
Class Weights	Balanced (computed)

Table 3: Performance Comparison of All Classifiers

Classifier	**P**	**R**	**MF1**
SVM	0.65	0.65	0.65
RF	0.63	0.63	0.63
XGBoost	0.65	0.65	0.65
LR	0.64	0.64	0.64
SGD	0.64	0.64	0.64
BiLSTM	0.69	0.67	0.67
GRU + Attention	0.66	0.66	0.66
Muril	0.74	0.70	0.71
IndicBERT	0.65	0.65	0.64
mBERT + BiLSTM	**0.76**	**0.76**	**0.76**

We found that the m-BERT+BiLSTM model achieved the highest macro-F1 score of 0.76 on the test dataset using FastText embeddings, outperforming other machine learning and deep learning models. The combination of TF-IDF and FastText provided improved contextual understanding, while BiLSTM effectively captured sequential patterns in the data.

6 Error Analysis

6.1 Quantitative Discussion

To evaluate model performance, both quantitative and qualitative analyses were conducted. The confusion matrix Figure 2 showed 395 true negatives, 211 true positives, 90 false positives, and 91 false negatives, indicating a slight bias toward the negative class. Qualitatively, the mBERT+BiLSTM model effectively detected explicit caste- and migration-related hate speech in Tamil-English code-mixed text but struggled with satirical or metaphorical expressions. Errors were often due to class imbalance and subtle language use, highlighting the need for sociocultural and discourse-level understanding.

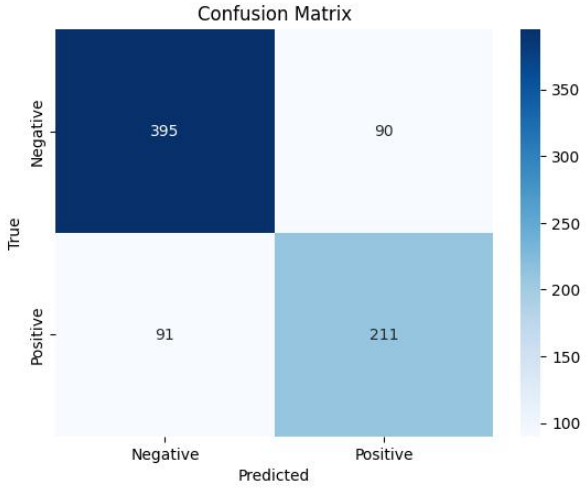

Figure 2: Confusion matrix of proposed model

7 Conclusion and Future Work

This work addressed the nuanced challenge of caste- and migration-based hate speech detection in Tamil by proposing a hybrid multilingual architecture. Integrating m-BERT's contextual embeddings with BiLSTM sequential modeling, and enriched by TF-IDF and FastText lexical features, our model achieved a macro-F1 score of 0.76, outperforming standalone baselines and affirming the utility of hybrid approaches in low-resource code-mixed settings. Looking ahead, future research can benefit from advanced ensemble techniques, such as weighted fusion of various transformer outputs or attention-based integration. Domain-adaptive pretraining on culturally grounded hate speech corpora

also holds promise for further boosting generalization. As the landscape of online discourse evolves, continuous innovation in modeling and data curation will be crucial in fostering safer, more inclusive digital environments.

Limitations

Despite the effectiveness of the mBERT+BiLSTM hybrid model in capturing contextual information, it struggled with nuanced and implicit hate speech, particularly in caste- and migration-related contexts. The training data set exhibited significant class imbalance, which led to biased learning and limited performance in minority classes. Furthermore, the multilingual and code-mixed nature of the data—especially Tanglish samples characterized by informal grammar, inconsistent spelling, and frequent code switching—posed considerable challenges. Data augmentation techniques such as synonym replacement and back translation were applied to address low-resource samples, but often introduced semantic noise due to transliteration inconsistencies, ultimately reducing the F1-score. The translation of Tamil text was also not feasible due to the lack of reliable parallel data. Furthermore, the model showed limitations in handling sarcasm, implicit hate, and vague contexts, which require deeper semantic and pragmatic understanding. Future work may explore transliteration-aware pretraining, more robust augmentation methods, and the use of pragmatic cues such as user metadata, comment threads, or conversational context.

References

Sattam Almatarneh and Pablo Gamallo. 2019. Citiuscole at semeval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 389–393, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sumaiya Rahman Aodhora, Shawly Ahsan, and Mohammed Moshiul Hoque. 2025. CUET_HateShield@NLU of Devanagari script languages 2025: Transformer-based hate speech detection in Devanagari script languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 260–266, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in nlp: The case of india. *arXiv preprint arXiv:2209.12226*.

Darsh Bhimani, Rutvi Bheda, Femin Dharamshi, Deepti Nikumbh, and Priyanka Abhyankar. 2021. Identification of hate speech using natural language processing and machine learning. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–4. IEEE.

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.

Asoka Chakravarthi and Bharathi Raja. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Z. M. Farooqi, S. Ghosh, and R. R. Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. COMBATANT@TamilNLP-acl2022: Fine-grained categorization of abusive comments using logistic regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228, Dublin, Ireland. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, and 1 others. 2021. Muril: Multilingual representations for indian languages. In *arXiv preprint arXiv:2103.10730*.

Rahul Ponnusamy, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sathiyaraj Thangasamy, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste/Immigration 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.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya MC, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Raúl R Romero-Vega, Oscar M Cumbicus-Pineda, Ruperto A López-Lapo, and Lisset A Neyra-Romero. 2021. Detecting xenophobic hate speech in spanish tweets against venezuelan immigrants in ecuador using natural language processing. In *Applied Technologies: Second International Conference, ICAT 2020, Quito, Ecuador, December 2–4, 2020, Proceedings 2*, pages 312–326. Springer.

Janak Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, and Priyanka Meel. 2021. Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 661–668.

Devanshu Sajlan. 2021. Hate speech against dalits on social media. *CASTE: A Global Journal on Social Exclusion*, 2(1):77–96.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Align and conquer: An ensemble approach to classify aggressive texts from social media. In *2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 82–86.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media. In *arXiv preprint arXiv:2101.03291*.

Boyang Yu, Fei Tang, Daji Ergu, Rui Zeng, Bo Ma, and Fangyao Liu. 2024. Efficient classification of malicious urls: M-bert—a modified bert variant for enhanced semantic understanding. *IEEE Access*, 12:13453–13468.