CUET_N317@LT-EDI 2025: Detecting Hate Speech Related to Caste and Migration with Transformer Models

Md. Nur Siddik Ruman, Md. Tahfim Juwel Chowdhury, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u2004098, u2004094}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Language that criticizes, threatens, or discriminates against people or groups because of their caste, social rank, or status is known as caste and migration hate speech, and it has grown incredibly common on social media. Such speech not only contributes to social disruption and inequity, but it also puts at risk the safety and mental health of the targeted groups. Due to the absence of labeled data, the subtlety of culturally unique insults, and the lack of strong linguistic resources for deep text recognition, it is especially difficult to detect caste and migration hate speech in low-resource Dravidian languages like Tamil. In this work, we address the Caste and Migration Hate Speech Detection task, aiming to automatically classify user-generated content as either hateful or non-hateful. We evaluate a range of approaches, including a traditional TF-IDF-based machine learning pipeline using SVM and logistic regression, alongside five transformer-based models: mBERT, XLM-R, MuRIL, Tamil-BERT, and Tamilhate-BERT. Among these, the domain-adapted Tamilhate-BERT achieved the highest macro-F1 score of 0.88 on the test data, securing 1st place in the Shared Task on Caste and Migration Hate Speech Detection at DravidianLangTech@LT-EDI 2025. Our findings highlight the strong performance of transformer models, particularly those fine-tuned on domain-specific data, in detecting nuanced hate speech in low-resource, code-mixed languages like Tamil.

1 Introduction

Caste and migration related hate speech is defined as language that insults, threatens, or discriminates against individuals or groups based on their caste, social status, or immigration background, has be- come increasingly prevalent on social media (Gagliardone et al., 2015). This type of speech affects mental health in a very bad way. So, detection of caste- and migration-related hate speech is very crucial. The automatic hate speech tool may be useful to prevent such activities. Detecting hate speech in a low-resource lan- guage like Tamil is very challenging due to lim- ited resources. In addition, our task was to identify only caste- and migration-related hate speech. To find out whether a sentence expresses hate or not it is crucial to understand the intent of the language (Schmidt and Wiegand, 2017). In Tamil, it is quite difficult to identify compared to high-resource language like English. In this study, we have fine- tuned several models to facilitate automatic hate speech detection in Tamil.

1. Proposed a Transformer based model with ensembles

2. Analyzed several ML and Transformer-based models for detecting hate speech in Tamil

Our code is available in this link GitHub repository.

2 Related Work

There have several research recently in identifying hate speech related to caste and migration in Tamil. Transformer models (Vaswani et al., 2017) have become foundational, with several recent studies exploring their efficacy in this specific domain. For the LT-EDI 2024 shared task on Caste and Migration Hate Speech Detection in Tamil (Rajiakodi et al.), Alam et al. (2024) investigated various models, finding M-BERT to achieve a macro F1-score of 0.80. In the same shared task, Singhal and Bedi (2024) demonstrated the power of ensembling transformer-based models (XLM-R, mBERT, MuRIL) through majority voting, securing the 1st rank with a macro F1-score of 0.82.

Beyond monolingual text, Hossain et al. (2025a) explored multimodal fusion for Telugu hate speech, also reporting a strong unimodal text baseline with

mBERT (F1-score 0.4968). The challenge of code-mixed Dravidian languages was addressed by Sreelakshmi et al. (2024), who found a combination of MuRIL embeddings and an SVM classifier to be highly effective, achieving accuracies up to 96%. For Indonesian, another context outside of high-resource languages, Hakim et al. (2024) combined IndoBERTweet with BiLSTM and CNN, yielding an F1-score of 85.06%.

Ensemble strategies remain popular; Roy et al. (2022) proposed a weighted ensemble of BERT models and a deep neural network for offensive and hate speech in Tamil and Malayalam code-mixed data, achieving high F1-scores. Highlighting the challenges in low-resource settings, Reddy et al. (2024) investigated data augmentation and noted the ineffectiveness of POS tagging for Dravidian languages. Multimodal approaches have also been investigated by Hossain et al. (2025b), who propose a transformer-based multimodal fusion model with cross-modal attention for hate speech detection.

3 Dataset and Task Description

The dataset (Ponnusamy et al., 2024) for the Caste and Migration Hate Speech detection task consists of mixed Tamil-English social media comments that have been annotated to indicate whether hate speech related to caste or migration is present (1) or not (0). Three separate sets of data are offered: training, development, and test.

The class-wise distribution of the dataset is summed up in Table 1. The class imbalance in both splits is similar, with approximately 62% of cases falling into the No Hate Speech class. An important factor in our modeling strategy is this imbalance.

Set	Non Hate Speech (0)	Hate Speech (1)	Total
Train	3,415 (61.96%)	2,097 (38.04%)	5,512
Development	485 (61.63%)	302 (38.37%)	787
Test	970 (61.55%)	606 (38.45%)	1,576

Table 1: Class-wise distribution of the dataset.

4 System Overview

Text classification tasks are currently very challenging for low-resource languages in social media like Tamil. Therefore, to detect Tamil hate speech related to caste and migration, we followed two distinct paths: a traditional machine learning approach as a baseline and advanced transformer-based approaches to handle the complexities of the text corpus. The figure depicts the overall process flow that we applied to do the task.

4.1 Data Preprocessing

This work is part of the Caste and Migration Hate Speech Detection shared task (Rajiakodi et al., 2025). The text corpus contained emojis, URLs, and mixed scripts, so we needed a pre-processing pipeline to manage it while preserving meaning. As a result, we performed text normalization to convert the whole corpus to lowercase to avoid case-sensitive mismatches. We applied regular expressions to strip out mentions like @username, URLs, numbers, and special characters. We also kept Tamil script (Unicode range `\u0B80-\u0BFF`) and English letters for code-mixed text. For the ML approach, we used the IndicNLP tokenizer to break Tamil text into meaningful units and removed Tamil stopwords using a curated list from GitHub. For transformers, we applied model-specific tokenizers from HuggingFace.

4.2 Feature Extraction

Feature Extraction is the process to convert raw text into a format comprehensible to our models. We used different tactics related to each approach:

- **Traditional ML:** We applied TF-IDF vectorization after trying out simpler bag-of-words models. We looked at unigrams, bigrams, and trigrams to get the short phrases and restricted the functionality to 7,000 to maintain things manageable without compromising on essential patterns.

- **Transformer-Based Models:** In this, we allowed the models' pre-trained tokenizers to do the job of converting text to token IDs with a maximum of 256 tokens. We truncated longer texts and padded shorter ones.

4.3 Traditional ML Approach

For our first step in the problem task, we used an ensemble of ML classifiers on the train dataset to figure out the complexities in more broader aspect. We trained a support vector machine (SVM) and a Logistic Regression model. For SVM, we used a linear kernel and enabled probability outputs, which we found necessary for ensemble voting. For Logistic regression, we kept it straightforward with the default configuration, mostly relying on its

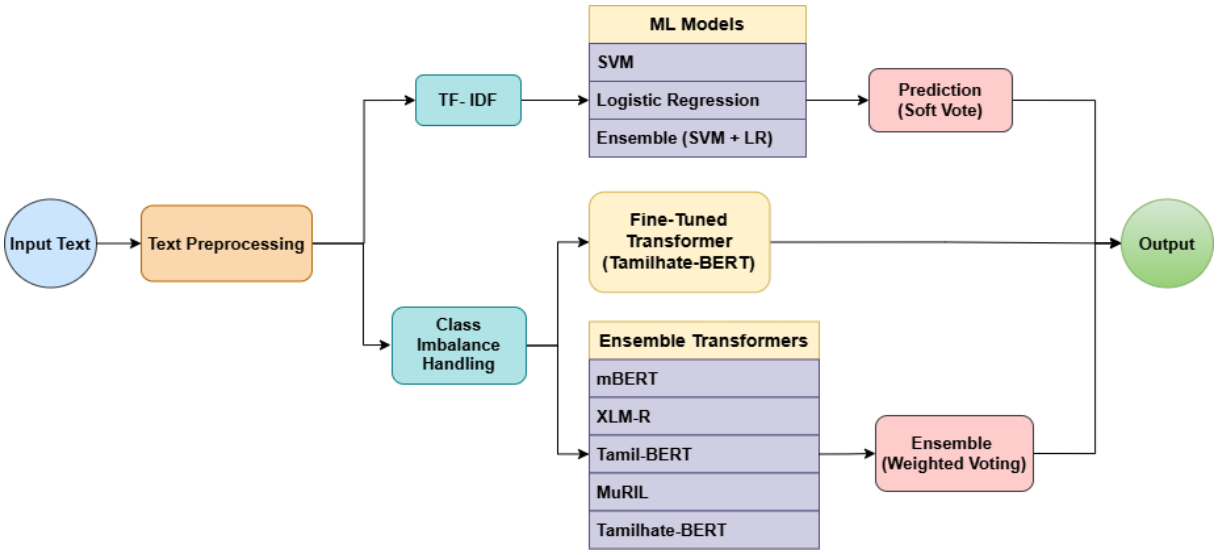

Figure 1: Schematic process of caste and migration hate speech detection in Tamil

probabilistic nature to complement the SVM. Then we combined these using a soft Voting Classifier, which averaged their predicted probabilities.

4.4 Transformer-Based Approach

Transformers are highly effective for understanding complex linguistic patterns and contextual relationships in text, especially in low-resource languages like Tamil. Therefore, we applied two strategies: a single fine-tuned model and a multi-model ensemble.

4.4.1 Fine-tuned Model

In this approach, we used a single fine-tuned transformer model to establish a strong, task-specific baseline by leveraging domain-adapted language understanding. We researched and found Tamilhate-BERT(mdo) model, a fine-tuned version of Tamil-BERT(Joshi, 2022) model on caste hate speech, which seemed like a perfect fit. To train this model, we used the AdamW optimizer with a learning rate of 1×10^{-5}, a batch size of 16, and trained for up to 10 epochs. We adopted early stopping after 2 epochs according to the validation F1 score, which saved us from overfitting. The dataset was imbalanced, so we calculated class weights based on inverse class frequencies and used them in a custom CrossEntropyLoss function. We logged metrics every 50 steps and kept the best model checkpoint based on macro F1.

4.4.2 Ensemble of Transformers

We combined multiple transformer models, hoping their diversity would make the system more robust. We fine-tuned five models: Tamil-BERT(Joshi, 2022), Tamilhate-BERT(mdo), MuRIL(Khanuja et al., 2021), mBERT(Devlin et al., 2019), and XLM-R(Conneau et al., 2020), each chosen for its strength in Tamil or multilingual aspects. Every model was fine-tuned with tailored hyperparameters, learning rates from 1×10^{-5} to 3×10^{-5} with batch sizes of 16 or 32 and trained for 12 to 14 epochs with early stopping after 3 epochs. We used class weights and gradient accumulation for some models to handle memory constraints. After that, we tested three different methods to combine predictions: majority voting, averaging probabilities, and weighted voting (using log-transformed validation F1 scores as weights). Weighted voting proved most effective after some experimentation.

Method	Classifier	Precision	Recall	Macro F1
ML	SVM	0.73	0.69	0.70
	Logistic Regression	0.72	0.65	0.65
	Ensemble	0.72	0.68	0.69
Transformers	mBERT	0.80	0.78	0.79
	XLM-R	0.80	0.77	0.78
	Tamil-BERT	0.79	0.78	0.78
	MuRIL	0.80	0.78	0.78
	Tamilhate-BERT	**0.88**	**0.88**	**0.88**
	Ensemble	0.84	0.81	0.82

Table 2: Performance of different systems on the test dataset.

5 Result and Analysis

Table 2 demonstrates the evaluation results of ML and Transformer models on the test set. Performance of the models was determined by the macro F1 score. The traditional machine learning technique with ensembling and soft voting achieved a macro F1 score of 0.69, reflecting a decent baseline but struggling to fully capture the linguistic complexity. Among the transformer-based models, XLM-R, Tamil-BERT and MuRIL each achieved macro F1 of 0.78 and the single fine-tuned Tamilhate-BERT achieved the highest macro F1 score of 0.88. This result highlights how effective domain-specific transfer learning can be. We have also explored an ensemble of multiple transformers, including mBERT, XLM-R, MuRIL, Tamil-BERT, and Tamilhate-BERT which unexpectedly reached a lower macro F1 score of 0.82. While ensembling added robustness, a well-targeted, fine-tuned model outperformed all others.

6 Error Analysis

We conducted a detailed error analysis to better understand the strengths and limitations of our best-performing model.

6.1 Quantitative Analysis

Figure 2 illustrates the performance of the top-performing model using a confusion matrix. The model correctly classified 876 out of 970 non-hate samples and 522 out of 606 hate speech samples. However, it misclassified 94 non-hate instances as hate and 84 hate instances as non-hate. The results reveal a slight bias toward the majority class (non-hate). While class weighting mitigated some imbalance, linguistic nuances in Tamil social media text, such as code-mixing, sarcasm, or context-dependent phrases, likely contributed to errors.

6.2 Qualitative Analysis

Figure 3 shows a few sample predictions of the best model on the test dataset. Some of the errors show that the model struggles with the informal or mixed language text. For example, one misclassification occurred where the text had clear caste-based insults written in mixed language. Another example written in Tamil was sarcastic and subtle, which was challenging for the model to interpret correctly. On the other hand, the model correctly classified some Tamil-English mixed texts properly.

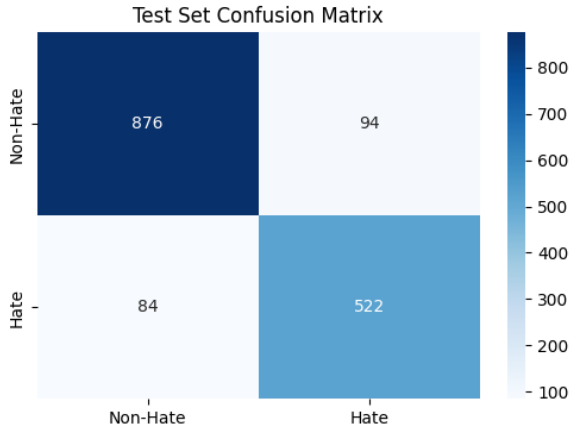

Figure 2: Confusion Matrix

Sample Text	Actual Label	Predicted Label
Holly பண்டிகை க்கு ஊருக்கு போரானுங்க திரும்பி கண்டீப்பா வருவானுங்க... (They go to their hometown for the Holi festival and will definitely come back.)	1	0
நாகப்பதனியா,நாகப்ப்பதனியா யார் பெரியவர் என்று பார்த்துவிடுவோம்… (Let's see who is greater, Nagapathan or Nagappathan.)	1	0
Nandu kadha theriyuma ? Tamizh nandu matum dhan mela yera vidama keezha thalite irundhudu.dats true (Do you know the crab story? Only Tamil crabs pull others down instead of letting them climb up. That's true.)	1	1
Avanunga holi festival ku poraanga yaa😂 (They're going for the Holi festival, yaa 😂)	1	1
தமிழர்களிடம் மட்டுமே வரவு செலவு வைத்துக் கொள்ள வேண்டும்! (You should keep accounts only with Tamils!)	0	0

Figure 3: Some predicted outcomes by the best-performing model

7 Conclusion

In this study, we analyzed three different methodologies for detecting caste and migration related hate speech in Tamil. Among these approaches, our second method Tamilhate-BERT, emerged as the top performer with a macro F1 score of 0.88, outperforming both the ML baseline and the transformer of ensemble. These findings highlight the power of transformers, especially when they are adapted to the specific linguistic and cultural characteristics of the task. For future work, we recommend enlarging the dataset with more varied examples, exploring multi-modal inputs to capture richer context, and devising strategies to further mitigate model bias.

Limitations

Our work has several limitations. First, the dataset size is relatively small, limiting the generalization of transformer-based models. A larger corpus could improve performance and robustness. Second, the code-mixed nature of the data along with slang, regional dialects, and informal spellings added extra complexity that our models may not fully capture. Third, Data augmentation techniques could be explored to improve model performance.

Acknowledgements

The authors gratefully acknowledge Centro Interuniversitario di Ricerca Scienze Umane e Sociali e Intelligenza Artificiale (ELIZA) – University of Naples 'L'Orientale' for its support in covering the registration costs, which enabled their participation.

References

mdosama39/tamil-bert-caste-hatespech_ltedi-tamil · hugging face. [Online; accessed 2025-05-08].

Md Ashraful Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. CUET_NLP_Manning@LT-EDI 2024: Transformer-based Approach on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243, Malta. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing.

Atalla Naufal Hakim, Yuliant Sibaroni, and Sri Suryani Prasetyowati. 2024. Detection of hate-speech text on indonesian twitter social media using indobertweet-bilstm-cnn. In *2024 12th International Conference on Information and Communication Technology (ICOICT)*, pages 374–381. IEEE.

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025a. SemanticCuet-Sync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection - A Transformer Based Approach with Cross-Modal Attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 567–573.

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025b. SemanticCuet-Sync@DravidianLangTech 2025: Multimodal fusion for hate speech detection - a transformer based approach with cross-modal attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 489–495, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *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, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 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*, pages 1–10.

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).

A Ankitha Reddy, Ann Maria Thomas, Pranav Moorthi, and B. Bharathi. 2024. SSN-Nova@LT-EDI 2024: POS Tagging, Boosting Techniques and Voting Classifiers for Caste And Migration Hate Speech

Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 233–237, Malta. Association for Computational Linguistics.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on NLP for Social Media*.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and Migration Hate Speech Detection in Tamil Using Ensembling on Transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, Malta. Association for Computational Linguistics.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and K.P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:12155–12168.

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