CUET_Blitz_Aces@LT-EDI-2025: Leveraging Transformer Ensembles and Majority Voting for Hate Speech Detection

Shahriar Farhan Karim*, **Anower Sha Shajalal Kashmary***, **Hasan Murad**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u2004065,u2004022}@student.cuet.ac.bd
hasanmurad@cuet.ac.bd
*

Abstract

The rapid growth of the internet and social media has given people an open space to share their opinions, but it has also led to a rise in hate speech targeting different social, cultural, and political groups. While much of the research on hate speech detection has focused on widely spoken languages, languages like Tamil, which are less commonly studied, still face significant gaps in this area. To tackle this, the Shared Task on Caste and Migration Hate Speech Detection was organized at the Fifth Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI-2025). This paper aims to create an automatic system that can detect caste and migration-related hate speech in Tamil-language social media content. We broke down our approach into two phases: in the first phase, we tested seven machine learning models and five transformer-based models. In the second phase, we combined the predictions from the fine-tuned transformers using a majority voting technique. This ensemble approach outperformed all other models, achieving the highest macro F1 score of 0.81682, which earned us 4th place in the competition.

1 Introduction

Social media is a crucial platform for accessing up-to-date information while also providing a space for individuals to exchange ideas, opinions, and thoughts, fostering meaningful conversations and building connections (Yamin et al., 2024). Although this democratization of expression enables people to express viewpoints and participate in debates, it has also contributed to the emergence of a major problem: the widespread transmission of hate speech (Watanabe et al., 2018).

Hate speech fuels division and polarization, escalating tensions between caste and migrant groups, and triggering discrimination and violence based on race, religion, gender, migration status, or other factors, threatening societal harmony and well-being (Al-Hassan and Al-Dossari, 2019). It can deeply psychologically affect its victims, hence generating emotional damage like fear, anxiety, depression, and a sense of isolation (Saha et al., 2019). Therefore, it is essential to implement automatic regulation of hateful content online to minimize the harm it can cause to society.

Significant study on the automatic identification of hate speech in text has been prompted by recent natural language processing (NLP) developments. Events to discover better techniques for automated hate speech detection have been held in major contests including SemEval-2019 (Zampieri et al., 2019) , SemEval-2020 (Zampieri et al., 2020), and GermEval-2018 (Wiegand, 2018). From many sources, researchers have built large-scale datasets that have inspired more research in this field, including studies on non-English languages and other online communities. These initiatives have opened up several processing pipelines for investigation, including different feature sets, machine learning techniques (supervised, unsupervised, and semi-supervised), and classification algorithms such as Naive Bayes, Logistic Regression (LR), Convolutional Neural Networks (CNN), LSTM, and BERT deep learning models (Jahan and Oussalah, 2023).

Our work aims to develop a system capable of distinguishing caste and migration hate speech from non-caste and migration hate speech, focusing on a low-resource language like Tamil (Chakravarthi et al., 2023), which belongs to the Dravidian language family (Krishnamurti, 2003). The primary contributions of our work are:

- Evaluated various machine learning and transformer models for hate speech detection in the Tamil language.
- Proposed a majority voting-based ensemble transformer approach for detecting hate

*These authors contributed equally to this work

speech in the Tamil language

Codes are available at GitHub Repository.

2 Related Work

Recent works have explored various computational approaches for automated detection of online hate speech targeting caste identities and migrant communities, ranging from traditional machine learning to advanced deep learning architectures.

(Alam et al., 2024) evaluated several models for Tamil hate speech detection. Their experiments showed M-BERT achieved an F1-score of 0.8049, outperforming BiLSTM (0.7490) and Tamil BERT (0.7847). In the context of Tamil-English code-mixed hate speech detection, (Pokrywka and Jassem, 2024) evaluated various transformer models, with google/muril-large-cased demonstrating strong performance with an F1-score of 0.81 on the challenge test set. They also experimented with xlm-roberta-large, bert-base-multilingual-cased, and roberta-base models. Taking a different approach, (Shanmugavadivel et al., 2024) explored traditional machine learning techniques for abusive comment detection in Tamil, reporting performance metrics for K-Nearest Neighbor (0.0772), Decision Tree (0.5862), and Naive Bayes (0.5905). (Singhal and Bedi, 2024) utilized an ensemble approach based on transformer models for Tamil hate speech detection, with MuRIL cased achieving an F1-score of 0.60, while also experimenting with XLM RoBERTa Large. (Sangeetham et al., 2024) combined traditional machine learning approaches for Tamil hate speech detection, with Support Vector Machines (SVM) performing remarkably well with an F1-score of 0.80, alongside Random Forest Classifier (RFC) and Decision Tree implementations. (Shanmugavadivel et al., 2023) proposed a machine learning approach for abusive comment detection in Tamil, achieving a macro-F1 score of 0.35. Their research revealed an important insight: traditional machine learning models can sometimes outperform sophisticated deep learning techniques when datasets are limited in size and complexity. Deep learning approaches tailored for code-mixed environments have shown promising results. (Anbukkarasi and Varadhaganapathy, 2023) employed a synonym-based Bi-LSTM model for Tamil-English code-mixed hate speech detection, achieving an F1 score of 0.8169. Their model demonstrated particular effectiveness in distinguishing between hate (F1 score: 0.8110) and non-hate texts (F1 score: 0.8050). (Subramanian et al., 2022) evaluated traditional machine learning models against transfer learning approaches for offensive language detection in Tamil YouTube comments.

3 Task and Dataset Description

The shared task on Caste and Migration Hate Speech Detection is part of LT-EDI@LDK 2025[1]. In this task, we focused mainly on Tamil-language content and aimed to build an automatic classification model that can analyze text from social media platforms and determine whether it contains caste-based or migration-related hate speech. The dataset was provided by the organizers of the competition (Rajiakodi et al., 2025). The training and development datasets consist of three columns: id (unique identifier), text (comment content), and label (1 for caste/immigration hate speech, 0 for non-hate speech). The test dataset includes only id and text. Figure 1 displays some samples of the training dataset.

id	text	label
4290	அவர்களிடம் வணிகம் வைத்துக்கொள்ள வேண்டாம் நம் மக்களுடைய கடையை தேடி போவோம் (We should not do business with them. Let's go find our own people's shop)	0
7377	முதல்ல வெட்டுடா யார் ஜெயிக்கிறார்கள் பார்கலாம் சும்மா பேசாதீங்க (First, let's see who wins. Don't talk nonsense)	1
126	வட இந்திய தென்னிந்தியா கிழக்கிந்திய மேற்கு இந்தியர்கள் அனைவரும் இந்தியர்கள் அல்ல😡😡😡😡 (North Indians, South Indians, East Indians, West Indians – not all of them are Indians)	1
6779	Nee thaniyave irrunthukko,athuthan elloorukkum nallathu (You should stay alone, that would be better for everyone)	0

Figure 1: Sample entries from the training dataset

Table 1 presents the class-wise distribution of the dataset across three subsets: Train, Dev, and Test. The training dataset consists of 5,512 instances, with 3,415 labeled as non-hate speech (label 0) and 2,097 as caste/immigration hate speech (label 1). The development dataset contains 787 instances, with 485 labeled as non-hate speech and 302 as hate speech. The test dataset includes 1,576 instances, with 970 non-hate speech and 606 hate speech instances. The datasets show an imbalance, with non-hate speech being more prevalent in each subset.

4 System Overview

Our approach included two stages. Figure 2 provides an overview of the first stage, which com-

[1] https://sites.google.com/view/lt-edi-2025

Dataset	Label		Total
	0	1	
Train	3,415	2,097	5,512
Dev	485	302	787
Test	970	606	1,576

Table 1: Class-wise distribution of the dataset

bines the application of various machine learning algorithms and the fine-tuning of transformers. In the second stage, we employed an ensemble transformer technique, combining the fine-tuned models with majority voting, which outperformed all other evaluated models.

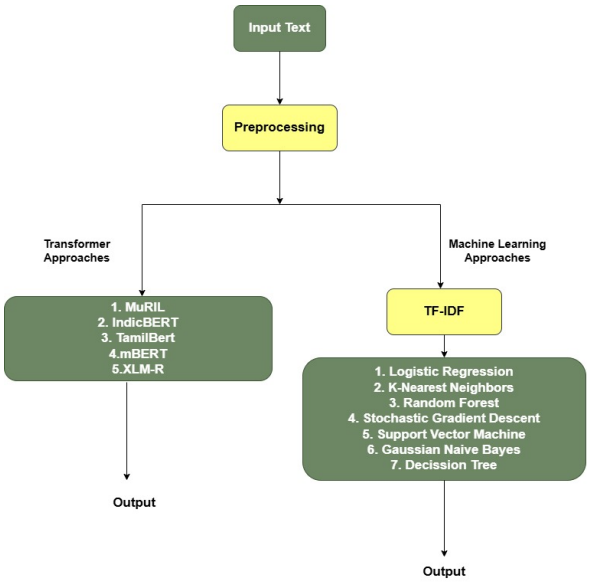

Figure 2: ML and transformer-based approaches for caste and migration hate speech detection in Tamil language

4.1 Data Preprocessing

In our preprocessing phase, we found that the corpus contains Latin characters of both upper and lowercase. Additionally, various emojis and emoticons were also used. We converted the text to lowercase and also removed emojis and punctuations. We took extra caution by replacing chat abbreviations (e.g., "LOL", "BRB", "IMO") with their full meanings.

4.2 Textual Feature Extraction:

ML algorithms cannot learn from raw texts. So Feature Extraction is necessary. We used TF-IDF (Tokunaga, 1994) technique to extract features from ML models

4.3 Classifiers

We used seven machine learning models and five transformer-based models to classify hate speech.

4.3.1 ML-based Approaches:

The experimented system used traditional ML approaches such as Logistic Regression, Support Vector Machine, KNN, Gausian Naive Bayes, Stochastic Gradient Descent (SGD), Random Forest and Decision Tree to establish the caste and migration-related hate speech detection system.

4.3.2 Transformer-based Approaches:

We conducted a comparative study using multiple transformer-based models to identify the most effective architecture for Tamil hate speech detection. We evaluated several state-of-the-art multilingual and language-specific models: MURIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), XLM-RoBERTa (Conneau et al., 2019), mBERT (Multilingual BERT) (Pires et al., 2019), IndicBERT (Dabre et al., 2021), and TamilBERT (Joshi, 2022).

To address class imbalance in the dataset, we computed class weights using the scikit-learn class_weight module. These weights were used in the cross-entropy loss function to appropriately penalize the misclassification of minority classes. A training loop was built using the following parameters: a learning rate of 1e-5, AdamW optimizer, a maximum sequence length of 115 tokens, and a batch size of 16. Each model was trained for up to 100 epochs with an early stopping criterion that halted training if no improvement in validation F1 score was observed for 5 consecutive epochs. The entire process utilized two T4 GPUs.

4.3.3 Proposed Majority Voting based Ensemble Approach:

To improve performance, we created an ensemble model (figure 3) that combines several top transformer models, including MURIL, XLM-RoBERTa, mBERT (Multilingual BERT), IndicBERT, and TamilBERT. Here's how it works: each model in the ensemble makes its own prediction, and the final decision is made through a majority voting system. In simple terms, the model that gets the most "votes" from the individual models becomes the final prediction. This method helps to balance out the weaknesses of any single model, making the overall predictions more reliable and consistent. Each model in the ensemble was fine-

tuned separately with the same training setup discussed earlier, ensuring that they work together effectively.

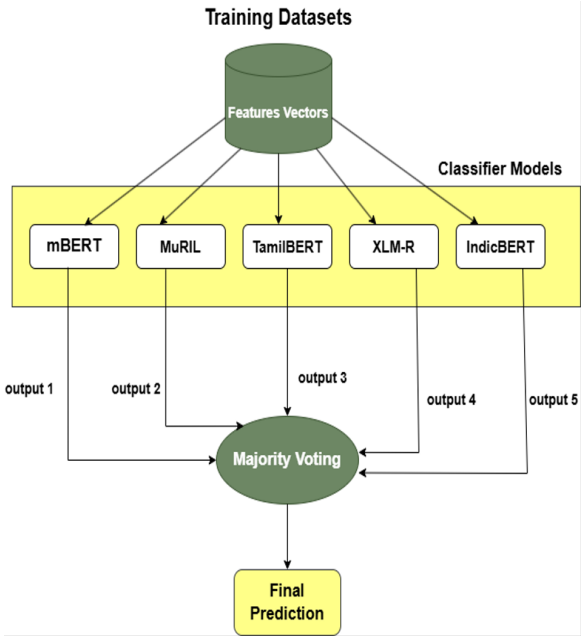

Figure 3: Proposed ensemble method for caste and migration hate speech detection in Tamil language

5 Results and Analysis

The table 2 shows the performance of several Machine Learning (ML) and Transformer models, hence highlighting the success of our suggested ensemble approach. Among the ML models, Random Forest has the highest precision (0.81401) and F1-score (0.77275), while Support Vector Machine performs best in recall (0.74413). Our suggested ensemble approach, which shows the strength of combining models for enhanced accuracy, beats everyone with the greatest precision (0.82492), recall (0.81139), and F1-score (0.81682). Though our ensemble approach performs better across all criteria, other Transformer models such as XLM-RoBERTa and MuRIL-BERT also produce good outcomes.

6 Conclusion

In this task, we created an automatic model to detect caste and migration-related hate speech in Tamil-language content on social media. By combining traditional machine learning techniques with transformer models, our ensemble approach achieved impressive results with a precision of 0.82492, recall of 0.81139, and an F1-score of 0.81682, outperforming all other models.

Machine Learning Models			
Model	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.72576	0.64925	0.64976
Support Vector Machine	0.82221	0.74413	0.75720
K-Nearest Neighbors	0.62712	0.52435	0.45024
Gaussian Naive Bayes	0.67062	0.66554	0.62345
Stochastic Gradient Descent	0.71073	0.69837	0.70249
Random Forest	0.81401	0.76053	0.77275
Decision Tree	0.72792	0.73169	0.72953
Transformer Models			
Model	**Precision**	**Recall**	**F1-score**
Tamil BERT	0.79411	0.78602	0.78947
M-BERT	0.78880	0.78933	0.78996
MuRIL-BERT	0.80655	0.79252	0.79799
Indic-BERT	0.77295	0.76746	0.76988
XLM-RoBERTa	0.81258	0.80871	0.81051
Ensemble (Proposed)	**0.82492**	**0.81139**	**0.81682**

Table 2: Performance of various models on the test set

Looking ahead, we plan to improve the model by expanding the dataset and incorporating multimodal data like images and emojis, which are common in social media posts. We also aim to explore more advanced transformer-based models for better context understanding, handling of informal language and classifying implicit hate speech. Lastly, implementing real-time detection for social media could make the model even more effective in addressing hate speech as it emerges

Limitations

Our studies have several limitations. Using a small and unbalanced dataset affects the model's ability to generalize. Our model also has trouble with mixed-language texts, slang, and emojis. It also failed to classify texts with regional dialect and implicit hate speech. More insights into Tamil hate speech could improve the model. We need to better handle informal expressions, regional dialects, and sarcasm. We could explore techniques like SMOTE, focal loss, and cost-sensitive learning to fix class imbalance and boost performance. Furthermore, our current analysis lacks statistical significance testing, ablation studies, and a detailed error analysis, especially regarding model interpretability and identifying Tamil-specific patterns. Additionally, incorporating domain expertise and cultural context could enhance model understanding. Future work should focus on expanding the dataset, including multimodal data for better understanding, and exploring normalization techniques for emojis and slang. There should also be an emphasis on real-time hate speech detection for social media and the integration of more sophisticated linguistic models for further improvement.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. pages 83–100.

Md 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, St. Julian's, Malta. Association for Computational Linguistics.

S Anbukkarasi and S Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. *arXiv preprint arXiv:2109.02903*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

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, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Bhadriraju Krishnamurti. 2003. *The dravidian languages*. Cambridge University Press.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Jakub Pokrywka and Krzysztof Jassem. 2024. kubapok@ lt-edi 2024: Evaluating transformer models for hate speech detection in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199.

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).

Koustuv Saha, Eshwar Chandrasekharan, and M. Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. *Proceedings of the 10th ACM Conference on Web Science*.

Saisandeep Sangeetham, Shreyamanisha Vinay, A Abishna, B Bharathi, et al. 2024. Algorithm alliance@ lt-edi-2024: Caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 254–258.

Kogilavani Shanmugavadivel, Malliga Subramanian, M Aiswarya, T Aruna, and S Jeevaananth. 2024. Kec ai dsnlp@ lt-edi-2024: Caste and migration hate speech detection using machine learning techniques. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 206–210.

Kogilavani Shanmugavadivel, Malliga Subramanian, Sree Harene JS, et al. 2023. Kec_ai_nlp@ dravidianlangtech: abusive comment detection in tamil language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 293–299.

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.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Takenobu Tokunaga. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6:13825–13835.

Michael Wiegand. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. Online available: https://epub.oeaw.ac.at/?arp=0x003a10d2 - Last access:13.5.2025.

Muhammad Mudassar Yamin, Ehtesham Hashmi, Mohib Ullah, and Basel Katt. 2024. Applications of llms for generating cyber security exercise scenarios. *IEEE Access*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

A Error Analysis

To obtain deeper insights by means of both quantitative and qualitative approaches, we carried out a thorough error analysis of our suggested ensemble model.

A.1 Quantitative Analysis

Figure 4 the confusion matrix for our proposed ensemble model indicates that it correctly recognized 865 cases of non-caste/migration hate speech (label 0) and 443 instances of caste/migration hate speech (label 1), for a total of 1308 accurate predictions. The model produced 268 erroneous predictions, comprising 105 false positives, where non-hate speech was misidentified as hate speech, and 163 false negatives, where hate speech was inaccurately categorized as non-hate speech. The errors likely arise from data imbalance and the variety of languages (English, Tamil, code-mixed, and code-switched) in the dataset, which hinder the model's capacity to accurately differentiate between hate and non-hate speech, particularly in intricate, context-dependent scenarios.

A.2 Qualitative Analysis

Figure 5 shows some random examples from the test data where the model's predictions are compared with the true labels. While it gets some right,

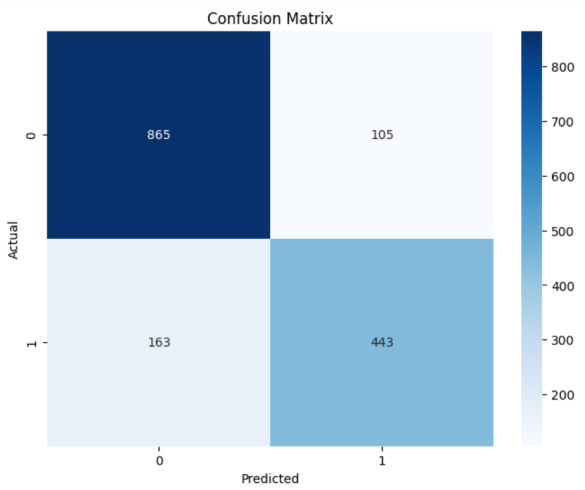

Figure 4: Confusion matrix for our proposed ensemble model

it misses the mark on others, especially when the text is a mix of Tamil and English. This code-mixed text, often full of slang, abbreviations, and informal expressions, can confuse the model, leading to incorrect predictions. Emojis and other non-standard characters add another layer of complexity, which the model might not always handle well. Another

Sample Texts	True Label	Predicted Label
அந்தவடநாட்டுவந்தேறியைவிரட்டிஅடியுங்கள் (Chase away the intruders who have come from the foreign land)	1	1
வடக்கன் என்ற ஒரு காரணத்தால் நான் உங்களுக்கு தேர்தலில் (Because of being a northerner, I will vote for you in the election)	0	0
Neenga dubai la pichai edukalaya?, Avan Inga edukiran (Are you begging in Dubai? He is begging here)	1	0
Avangala avanga ooruke anupura vazhiya parunga (Look at the way they are living in their town)	0	1

Figure 5: Some randomly selected samples from the dataset along with predictions from our model

issue that affects the model's performance is the class imbalance in the dataset. There are more examples of non-hate speech than hate speech, which can cause the model to favor predicting the majority class. This means it might struggle more when it comes across less obvious hate speech, or more subtle expressions of hate, especially when these are mixed with sarcasm or indirect language. So, the model's tendency to focus on non-hate speech

and its struggle with understanding the nuances of mixed-language content are some of the key reasons for these misclassifications. Improving how the model deals with imbalanced data and mixed, informal language could help it perform much better.