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

Nishanth S, Shruthi Rengarajan, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4aie22149, cb.en.u4aie22154}@cb.students.amrita.edu
s_sachinkumar@cb.amrita.edu

Abstract

Hate speech directed at caste and migrant communities is a widespread problem on social media, frequently taking the form of insults specific to a given region, coded language, and disparaging slurs. This type of abuse seriously jeopardizes both individual well-being and social harmony in addition to perpetuating discrimination. In order to promote safer and more inclusive digital environments, it is imperative that this challenge be addressed. However, linguistic subtleties, code-mixing, and the lack of extensive annotated datasets make it difficult to detect such hate speech in Indian languages like Tamil. We suggest a supervised machine learning system that uses FastText embeddings specifically designed for Tamil-language content and Whisper-based speech recognition to address these issues. This strategy aims to precisely identify hate speech connected to caste and migration, supporting the larger endeavor to reduce online abuse in low resource languages like Tamil.

Keywords: Hate Speech, Machine Learning models, FastText

1 Introduction

The rapid and sudden expansion of social media websites has transformed and revolutionized how people communicate with each other, exchange different types of information, and interact with a vast spectrum of disparate communities and communities of interest across the world (Sharma et al., 2025). Nevertheless, it also needs to be observed that such cyberspaces have also turned into important hotspots and breeding grounds for spreading hate speech, with the focus being specifically placed on targeting vulnerable communities, i.e., marginalized caste groups and migrant communities (Barman and Das, 2023)(Jahan and Oussalah, 2023). This malicious kind of abuse usually tends to manifest itself in the form of region-based abuses, coded language imparting veiled messages, and a range of derogatory statements that are employed to maintain prevailing social hierarchies and legitimize systemic discrimination (Kumar et al., 2017). The negative effects and negative implications stemming from such toxic content extend far and beyond the specific harm that results, and pose a very realistic threat to social cohesion, community harmony, and overall inclusivity in society(V P et al., 2023).

Although a great deal of research has been conducted and a great deal of investments have been made towards hate speech detection in high-resource languages like English, much remains to be done in addressing this very serious problem in low-resource languages properly (Papcunová et al., 2023)(MacAvaney et al., 2019) (K et al., 2021). This is a very acute problem in linguistically diverse and rich countries like India, where there are numerous languages and dialects spoken (Chakravarthi et al., 2023). Tamil, being one of the major languages of South India, has very distinctive problems concerning this, mainly because of its very unique linguistic properties, difference among various dialects, and the widespread use of code-mixing by the speakers. These cumulative problems make it particularly challenging to create useful automated hate speech detection systems. Moreover, the unavailability of properly annotated datasets only serves to further add to the issues of creating useful detection systems for this kind of abusive language (Rajiakodi et al., 2025).

In an effort to efficiently solve and address these major challenges, the Shared Task on Caste and Migration Hate Speech Detection, to be hosted at the well-regarded LT-EDI@LDK 2025 conference, has been crafted with the specific goal of encouraging and enabling large-scale research and development of robust machine learning models for addressing this major and urgent issue. The new proposed solution exploits FastText-based embeddings that

have been specifically designed for Tamil text processing and analysis. With a robust emphasis on hate speech detection and solving based on caste and migration challenges, this major task not only progresses but also fits within the wider and more general goal of creating safer, more respectful, and more inclusive digital spaces. This is especially critical for communities that use low-resource languages, which are likely to be exposed to special challenges and vulnerabilities in the digital space.

More details about the shared task can be found at[1].

2 Dataset

The dataset was distributed by the shared task organisers of Caste and Migration Hate Speech Detection - LT-EDI@LDK 2025 (Rajiakodi et al., 2024).

Data Type	Sample Size
Training	5512
Development (Dev)	787
Testing	1566

Table 1: Dataset split of the speech samples

The dataset used for this study comprises sentences in the Tamil language, categorized into two classes: Abusive and Non-Abusive (Class distribution). The data set is split into training and test sets. It was specifically developed for evaluating the suitability of language models for identifying abusive language in low-resource Dravidian languages, ensuring near-balanced Abusive and Non-Abusive example representations to provide more efficient training and evaluation.

3 Methodology

3.1 Data Preprocessing

The data pre-processing involves a series of steps such as conversion of data into lowercase, ensuring uniformity in the labels. The URLs and special characters are removed from the data to ensure consistency and to make sure the data is model friendly.

3.2 Embedding Generation

FastText offers character n-gram embeddings to enhance vector representation for morphologically rich languages. Words are represented as the average of these embeddings. It is a word2vec model extension. While FastText offers embeddings for character n-grams, the Word2Vec model offers embeddings for words. Similar to the word2vec model, fastText computes the vectors using CBOW and Skip-gram, using the subword information, allowing it to generate embeddings for out-of-vocablury words. This is especially useful for morphologically rich languages like Tamil.

In the proposed methodology, a 300-dimensional word embedding models for English and Tamil, trained on the Common Crawl Data is used.

The FastText model internally performs a series of functionalities. This includes:

- **Tokenization:** Each sentence is split into words.
- **Word Lookup:** For each word, the corresponding embedding vector is retrieved from the FastText model. Only vectors for words present in the model's vocabulary are kept.
- **Sentence Embedding:** The embedding of each sentence is computed by taking the mean of its word vectors, resulting in a fixed-size (300-dimensional) vector for each sentence.
- **Fallback for Empty Text:** If a sentence contains no valid words (e.g., only punctuation), a zero vector is assigned as its embedding.
- **Final Output:** A tensor of shape `[num_sentences, 300]` is returned, representing the sentence embeddings.

3.3 Machine Learning Models

The embeddings created are first loaded. Then, to make the final model more robust, the methodology combines the training and validation embeddings. This methodology performs feature-level fusion by concatinating the Tamil original embeddings and the Tamil-English translated embeddings so as to capture the semantics from both the languages in one input vector, enriching the feature space.

Several machine learning models have been used for this classification task (S et al., 2025):

- **XGBoost:** A powerful gradient-boosted tree model with GPU acceleration.
- **Logistic Regression:** A simple, linear classifier used as a baseline model.

[1] https://codalab.lisn.upsaclay.fr/competitions/21884

- **Random Forest:** An ensemble of decision trees that effectively handles feature interactions.

- **SVM (Support Vector Machine):** Particularly effective in high-dimensional spaces, making it suitable for text embeddings.

- **KNN (K-Nearest Neighbors):** A non-parametric method that bases predictions on the nearest neighbors.

- **MLPClassifier:** A shallow neural network (multi-layer perceptron).

4 Evaluation

The model is trained using the training embeddings and evaluated using the test embeddings. Metrics like the Accuracy, F1 score, Precision and Recall and the Training & Evaluation time has been taken into account. When all inferences of the models were submitted, the XGBoost model, which was trained in Tamil, emerged on top of others, securing us the rank **7th** with a F1 score of 0.80095

4.1 Final Model Performance

Dataset	Model	Acc	F1	Prec	Rec
Tamil	XGBoost	0.7878	0.7835	0.7859	0.7878
Tamil	Logistic Regression	0.6366	0.5697	0.6226	0.6366
Tamil	Random Forest	0.6607	0.5819	0.7026	0.6607
Tamil	SVM	0.6264	0.5088	0.6411	0.6264
Tamil	KNN	0.6163	0.6163	0.6163	0.6163
Tamil	MLP	0.7446	0.7457	0.7473	0.7446
Tamil-English	XGBoost	0.7700	0.7631	0.7685	0.7700
Tamil-English	Logistic Regression	0.6302	0.5675	0.6084	0.6302
Tamil-English	Random Forest	0.6595	0.5755	0.7120	0.6595
Tamil-English	SVM	0.6353	0.5369	0.6489	0.6353
Tamil-English	KNN	0.6455	0.6376	0.6357	0.6455
Tamil-English	MLP	0.7598	0.7604	0.7610	0.7598
Tamil (Orig+Eng)	XGBoost	0.7916	0.7860	0.7911	0.7916
Tamil (Orig+Eng)	Logistic Regression	0.6544	0.6150	0.6428	0.6544
Tamil (Orig+Eng)	Random Forest	0.6696	0.5882	0.7438	0.6696
Tamil (Orig+Eng)	SVM	0.6544	0.5941	0.6559	0.6544
Tamil (Orig+Eng)	KNN	0.6226	0.6154	0.6127	0.6226
Tamil (Orig+Eng)	MLP	0.7827	0.7829	0.7831	0.7827

Table 2: Performance of Various Models on Tamil and Tamil-English Datasets

The code files for this project can be accessed from[2]

[2] https://github.com/NishanthSaravanamurali/NS-LT-EDI-2025-Caste-Migration-based-hate-speech-Detection.git

5 Conclusion

This paper presents the results of a task performed as part of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages, focusing on Caste and Migration Hate Speech Detectio in Tamil speech dataset. The conference provided the dataset for the proposed task. The proposed methodology makes use of the FastText model and embedding generation to train the models and compare the accuracies.

6 Limitations

While working on this topic, the major limitation we faced is the mixed language, as some text was in English and some was in Tamil. Translating one language causes loss of contextual information which in this case is important even though it provided additional features that can help ML classifiers help classify better.

References

Shubhankar Barman and Mithun Das. 2023. hate-alert@dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *DRAVIDIANLANGTECH*.

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.

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

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

S. Sachin Kumar, M. Anand Kumar, and K. P. Soman. 2017. Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In *Mining Intelligence and Knowledge Exploration*, pages 320–334, Cham. Springer International Publishing.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019.

Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):e0221152.

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex Intelligent Systems*, 9:2827–2842.

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.

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).

Nishanth S, Shruthi Rengarajan, S Ananthasivan, Burugu Rahul, and Sachin Kumar S. 2025. ANSR@DravidianLangTech 2025: Detection of abusive Tamil and Malayalam text targeting women on social media using RoBERTa and XGBoost. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 711–715, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar Singh. 2025. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(3).

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.