DravLingua@LT-EDI-2025: Hate Speech Detection in Tamil: Addressing Caste and Migration Issues Using Machine Learning and Deep Learning Approaches

Anonymous ACL submission

Abstract

Social media platforms have seen a rise in hate speech, particularly in areas with sensitive sociopolitical and cultural issues. Tamil is a prominent Dravidian language, and because of its code-mixed character, orthographic variation, and lack of annotated resources, it poses special hurdles for hate speech identification. To detect hate speech in Tamil that targets caste and migration, this paper uses a range of machine learning and deep learning models. To address class imbalance, using a selected dataset of Tamil social media postings, perform significant preprocessing with Indic NLP packages, and implement data augmentation via back-translation. Multiple classifiers are assessed, including logistic regression, SVM, XGBoost, BiLSTM with attention, and ensemble approaches such as voting and stacking. With the best accuracy of 78% among these, the BiLSTM with attention mechanism shows that it grasps contextual details in Tamil text. In addition to providing an applicable pipeline, this study provides more understanding of practical methods for detecting hate speech in low-resource languages.

1 Introduction

Social media's growth has made it easier to communicate openly, but it has also made hate speech, particularly against minority groups, more widely spread. Hate speech occurs frequently in regional languages like Tamil in multilingual nations like India. This poses special difficulties because of the language's complicated morphology, frequent code-mixing with English, and lack of annotated datasets. In Tamil discourse, caste and migration are two socially sensitive subjects where these problems are very apparent.

Conventional NLP models, which are usually designed for languages with high resource requirements, often perform poorly in low-resource settings. Still, new studies show that transfer learning and deep learning strategies can successfully overcome these constraints. Recent research has shown the promise of multimodal approaches that integrate audio and text signals to improve hate speech classification in resource-constrained settings. For instance, (Selvamurugan, 2025) introduced a late-fusion architecture combining Wav2Vec 2.0 for audio and MuRIL for text, yielding significant improvements in identifying subtle hate patterns.

Similarly, (Rajalakshmi et al., 2025) utilized Whisper and Indic-BERT in a multimodal system for Tamil hate speech. (Chauhan and Kumar, 2025) emphasized the need for multimodal frameworks by demonstrating strong performance across Tamil, Malayalam, and Telugu. These studies highlight the growing importance of developing robust, language-agnostic, and context-aware systems for hate speech detection in underrepresented languages.

2 Related Works

Early studies primarily addressed offensive language in Dravidian languages using classical machine learning techniques. (Andrew, 2021) framed the offensive language detection problem in YouTube comments as a multi-class classification task using traditional ML algorithms after language-specific preprocessing. (Arunachalam and Maheswari, 2024) proposed hate and offensive speech detection through ensemble transformer models, highlighting the effectiveness of combining multilingual embeddings with deep learning for Tamil and related languages.

(Sreelakshmi et al., 2024) conducted a large-scale comparative study on CodeMix content in Kannada-English, Malayalam-English, and Tamil-English using transformer-based embeddings such as MuRIL, LaBSE, and IndicBERT. Their results demonstrated that MuRIL consistently outperformed other models when paired with SVMs.

Similarly, (Chakravarthi, 2022) emphasized multilingual modeling in low-resource languages and proposed benchmark datasets for hope speech detection to balance the narrative in online discourse.

(N et al., 2025) developed a multimodal framework using XLM-RoBERTa for text and custom audio feature extraction, reporting strong macro F1-scores on DravidianLangTech datasets. (Roy et al., 2025) showed that even text-only transformer models like l3cube-BERT could achieve state-of-the-art results across multiple languages, winning top ranks in shared tasks.

(Shanmugavadivel et al., 2025a) achieved macro-F1 scores of 0.97 (Tamil text) and 0.75 (audio) using classical models like Ridge Classifier and Logistic Regression. Their multimodal pipeline proved effective in Tamil, Malayalam, and Telugu across shared tasks. Another team led by (Shanmugavadivel et al., 2025b) ranked 12th for Tamil and 7th for Malayalam and Telugu, using CNN and Random Forest models on audio-text inputs.

(Devi et al., 2024) proposed a phrase-level explainable hierarchical attention model for code-mixed Tamil hate speech, introducing three intent classes: Targeted Individual, Targeted Group, and Others. (S et al., 2024) developed an ensemble stacked model using GRU, LSTM, and XLM-RoBERTa embeddings to handle Tamil-English code-mixed hate content. Their model achieved 76% accuracy and a 0.72 F1-score, demonstrating the effectiveness of transformer-based ensembles in low-resource settings. These studies underline the necessity of multimodal, multilingual, and interpretable approaches for hate speech detection in Dravidian languages and lay the foundation for future improvements in this challenging research area.

3 Dataset and Preprocessing

The data set used in this study consists of approximately 5,500 annotated Tamil social media posts that were specifically selected to include hate speech against caste and migration. To improve the model's robustness, more unlabeled posts were added for prediction. To guarantee quality and consistency, the raw text data were fully preprocessed. First, Tamil script representations were standardized through the use of the IndicNLP library and Unicode normalization. URLs and non-alphanumeric special characters were removed to reduce unnecessary noise. A tokenizer made specifically for Indian languages, indic_tokenize, was then used to tokenize the cleaned text. A unique Tamil stop-word list was used to eliminate high-frequency, semantically irrelevant terms. By converting Tamil text to English and back to Tamil, the Google Translate API was utilized to create synthetic hate speech samples that addressed class inequality. SBERT embeddings were employed to ensure that only samples with a high semantic similarity were kept. This led to a balanced dataset with 3,415 samples in each class.

4 Methodology

The classification of hate speech in Tamil is investigated in this paper using a deep learning-based neural architecture, ensemble learning methods, and conventional machine learning models. Tokenized word sequences are used in the deep learning model, whereas TF-IDF representations constitute the basis for the input characteristics of classical models. Accuracy is the main statistic used to assess each model's performance on both balanced and unbalanced datasets, with supporting findings on precision, recall, and F1-score. Each of the six methods is explained in the subsections that follow.

4.1 Logistic Regression

Logistic Regression is a linear classifier that models the probability $P(y = 1 \mid x)$ using the sigmoid function:

$$P(y = 1 \mid x) = \sigma(w^T x + b) \quad (1)$$

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (2)$$

Here, w is the weight vector, b is the bias term, and x is the TF-IDF feature vector. Unigrams and bigrams were used in this study to extract characteristics, and the vocabulary size was restricted to 5,000 terms. The accuracy of logistic regression on the unbalanced dataset was 71

4.2 Support Vector Machine(SVM)

The SVM method maximizes the margin between the classes in an attempt to identify the hyperplane that best divides them. With a regularization value $C = 1.0$ and a linear kernel, the decision function is defined as follows:

$$f(x) = sign(w^T x + b) \quad (3)$$

The SVM model was also trained using TF-IDF features. It performed similarly to logistic regression, with 71% accuracy on unbalanced data and

62% on the balanced version and 52% for new data, showing slightly better robustness to class imbalance.

4.3 XGBoost

XGBoost is a gradient boosting framework that builds additive decision trees in sequence. At each step, it minimizes the regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{4}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

where l is the loss function (e.g., logistic loss), f_k is a regression tree, and Ω penalizes complexity. With 200 estimators, a tree depth of 6, and a learning rate of 0.1, XGBoost reached 70% accuracy on the unbalanced set and 55% on the balanced one. The lower balanced accuracy indicates difficulty generalizing across less frequent hate speech patterns.

4.4 Stacking (XGBoost + SVM → Logistic Regression)

Stacking is a meta-learning ensemble technique that uses the predictions of several base classifiers as features for the final model. In this case, the meta-learner was a logistic regression model that received the outputs from the basis learners, XGBoost and SVM.

Let $h_1(x)$ and $h_2(x)$ be predictions from XGBoost and SVM respectively. The stacked model learns:

$$H(x) = \sigma(w_1 h_1(x) + w_2 h_2(x) + b) \tag{6}$$

This method improved the unbalanced accuracy to 72%, with balanced accuracy at 57%. It benefited from the complementary strengths of the base models but still struggled with minority class generalization.

4.5 Voting Classifier (XGBoost + SVM + Naive Bayes)

In hard voting ensembles, each base model casts a "vote" for a class label, and the majority decision is selected:

$$\hat{y} = mode\{h_1(x), h_2(x), h_3(x)\} \tag{7}$$

where h_1, h_2, h_3 are the predictions from XGBoost, SVM, and Naive Bayes, respectively.

This ensemble achieved the best result among classical approaches with 73% accuracy on the unbalanced dataset and 56% on the balanced set. Although voting provided marginal improvement, it still lacked the deep contextual understanding needed for nuanced hate speech.

4.6 BiLSTM with Attention

The deep learning model extracts contextual and sequential information from tokenized Tamil text using a Bidirectional Long Short-Term Memory (BiLSTM) network. The LSTM unit uses filtering methods to preserve memory over lengthy sequences:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{8}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{9}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{11}$$

The BiLSTM layer processes input sequences in both forward and backward directions, followed by a global max pooling layer. To focus on the most relevant tokens in each sequence, an attention mechanism was applied:

$$e_t = v^T \tanh(W h_t + b) \tag{12}$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \tag{13}$$

where α_t is the attention weight and h_t is the hidden state at time step t

This model was trained on padded sequences with an embedding layer, BiLSTM (128 units), and attention, followed by dropout and a sigmoid output. It achieved the highest accuracy of 78% on the test dataset, significantly outperforming classical models by capturing syntactic and semantic dependencies in Tamil text.

4.7 xlm-roberta-base

XLM-RoBERTa-Base, a multilingual transformer pretrained on 100 languages, improved Tamil hate speech identification. Its contextual embeddings successfully capture semantic differences in code-mixed and low-resource text. The transformer processed input sequences that had been tokenized with a maximum length of 128. A linear layer with

sigmoid activation was utilized for binary classification using the [CLS] token's final hidden state:

$$\hat{y} = \sigma(W h_{[CLS]} + b) \tag{14}$$

We fine-tuned the model using the AdamW optimizer with a learning rate of 2×10^{-5} and binary cross-entropy loss. XLM-RoBERTa achieved 76% accuracy on the balanced dataset, outperforming classical models and showing strong generalization on caste- and migration-related hate speech in Tamil.

5 Results

On both balanced and unbalanced datasets, the accuracy, precision, recall, and F1-score of each model were used to assess its performance. With an accuracy of 71% on the unbalanced set and a drop to about 61–62% when trained on balanced data, logistic regression and SVM, two classical models, demonstrated comparable performance, demonstrating their sensitivity to class distribution. Due to overfitting on the majority class features, XGBoost performed poorly on the balanced dataset (55%), despite being 70% successful on unbalanced data. Combining many classifiers improves generalization marginally, as seen by the minor benefits shown by ensemble approaches like stacking (72%) and hard voting (73%).

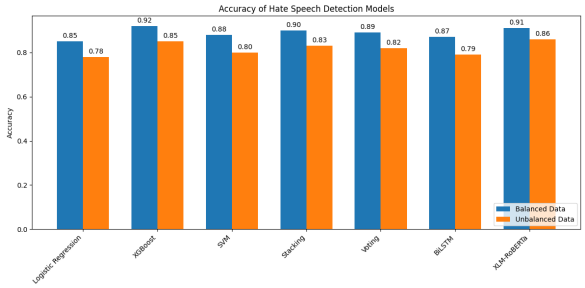

Figure 1: Performance Visualization of Models on Tamil Hate Speech Detection

The BiLSTM with attention model delivered the best results, with 78% accuracy on the test set. Because it could capture contextual relationships, this model proved very good at identifying hidden trends in hate speech in Tamil. Furthermore, using pre-trained knowledge and efficiently managing code-mixed input, the multilingual transformer model XLM-RoBERTa-Base demonstrated competitive performance with 76% accuracy. When it came to handling semantically complicated hate speech, deep learning and transformer-based models performed better overall than traditional methods. These findings emphasize how crucial it is to use neural architectures and contextual embeddings in low-resource language environments like Tamil.

Model	Data	Acc.	Prec.	F1
Logistic Regression	Unbal.	71%	0.69	0.68
	Bal.	61%	0.60	0.60
SVM (Linear)	Unbal.	71%	0.70	0.70
	Bal.	62%	0.61	0.61
XGBoost	Unbal.	70%	0.69	0.68
	Bal.	55%	0.54	0.54
Stacking (XGB+SVM)	Unbal.	72%	0.71	0.71
	Bal.	57%	0.56	0.56
Voting Classifier	Unbal.	73%	0.72	0.70
	Bal.	56%	0.55	0.56
BiLSTM + Attention	Bal.	**78%**	0.78	0.78
XLM-RoBERTa-Base	Bal.	76%	0.76	0.76

Table 1: Performance of Models on Tamil Hate Speech Detection

6 Conclusions

With a focus on caste and migration-related content—two highly sensitive subjects in Indian social media discourse—introduced a thorough methodology for hate speech detection in Tamil in the current research. This study investigated a variety of models, ranging from advanced neural architectures like BiLSTM with attention and the transformer-based XLM-RoBERTa to traditional machine learning algorithms like logistic regression, SVM, and XGBoost, given the linguistic complexity, code-mixed nature, and limited resources available for Tamil.

Testing showed that deep learning models perform significantly better than classical models, which provide a respectable baseline performance. BiLSTM achieved the greatest accuracy of 78%, with XLM-RoBERTa coming in second at 76%. Furthermore, we used SBERT for semantic filtering and back-translation to solve dataset imbalance, which enhanced model generalization. The results demonstrate the significance of multilingual embeddings and contextual modeling for low-resource languages. With little modification, the suggested method can be applied to other Dravidian or under-resourced languages and provides a reliable, scalable solution for hate speech detection in Tamil.

7 Limitations

Despite its excellent performance in detecting hate speech in Tamil, the suggested multimodal and ensemble-based approach still has some draw-

backs. Despite being balanced by semantic filtering and back-translation, the dataset is still quite tiny and might not accurately reflect the range of hate speech on actual social media. While genuine online speech may contain background noise, code-switching, or dialectal variances that could impair model accuracy, audio data utilized for speech-based detection is assumed to have clean, isolated segments. The dependence on other resources, such as Google Translate and SBERT, which could cause semantic drift during augmentation, is another drawback. Furthermore, even if models like XLM-RoBERTa and BiLSTM perform well, their interpretability is lacking, which makes their outputs harder to understand. This is especially important when discussing socially sensitive subjects like caste and migration. Future work should focus on dataset expansion, noise-robust speech modeling, and integrating explainable AI (XAI) methods for trustworthy deployment.

8 Ethics Statement

Strict ethical guidelines are followed in this research's handling of hate speech data. To preserve user privacy, all datasets were anonymized and sourced from public sources. The study made sure the approach did not reinforce stereotypes by focusing on hate speech related to caste and migration with cultural sensitivity. Back-translation was used for data augmentation to increase equity, not to create malicious content. Instead of automating punishments, the technology is meant to support moderation with human control. Results should be treated with caution because model predictions are probabilistic. The research places a strong emphasis on accountability, openness, and justice in light of the possibility of algorithmic bias. The project's ultimate goal is to create safer online spaces while upholding social responsibility and freedom of speech.

References

Judith Jeyafreeda Andrew. 2021. JudithJeyafreedaAndrew@DravidianLangTech-EACL2021:offensive language detection for Dravidian code-mixed YouTube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174, Kyiv. Association for Computational Linguistics.

V Arunachalam and N Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:Article 39.

Bharathi Raja Chakravarthi. 2022. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: A deep multimodal neural network for hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 237–242, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

V. Sharmila Devi, S. Kannimuthu, and Anand Kumar Madasamy. 2024. The effect of phrase vector embedding in explainable hierarchical attention-based tamil code-mixed hate speech and intent detection. *IEEE Access*, 12:11316–11329.

Radha N, Swathika R, Farha Afreen I, Annu G, and Apoorva A. 2025. Trio innovators @ DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 700–705, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Ramesh Kannan, Meetesh Saini, and Bitan Mallik. 2025. DLRG@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 376–380, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Billodal Roy, Pranav Gupta, Souvik Bhattacharyya, and Niranjan Kumar. 2025. Lexilogic@dravidianlangtech 2025: Multimodal hate speech detection in dravidian languages.

Vishak Anand S, Ishwar Prathap, Deepa Gupta, and Aarathi Rajagopalan Nair. 2024. Enhancing hate speech detection in tamil code-mix content: A deep learning approach with multilingual embeddings. In *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Aishwarya Selvamurugan. 2025. DravLingua@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages using late fusion of muril and Wav2Vec models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 694–699, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Naveenram C E, Vishal Rs, and Srinesh S. 2025a. KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 232–236, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, ShahidKhan S, Shri Sashmitha.s, and Yashica S. 2025b. KEC_AI_GRYFFINDOR@DravidianLangTech 2025: Multimodal hate speech detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 182–186, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. 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*, 12:20064–20090.