CUET's_White_Walkers@LT-EDI 2025: Transformer-Based Model for the Detection of Caste and Migration Hate Speech

Jidan Al Abrar, Md Mizanur Rahman, Ariful Islam,
Md Mehedi Hasan, Md Mubasshir Naib, Mohammad Shamsul Arefin
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{080, 116, 129, 067, 089}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Hate speech on social media is an evolving problem, particularly in low-resource languages like Tamil, where traditional hate speech detection approaches remain under developed. In this work, we provide a focused solution for cast and migration-based hate speech detection using Tamil-BERT, a Tamil-specialized pre-trained transformer model. One of the key challenges in hate speech detection is the severe class imbalance in the dataset, with hate speech being the minority class. We solve this using focal loss, a loss function that gives more importance to harder-to-classify examples, improving the performance of the model in detecting minority classes. We train our model on a publicly available labeled dataset of Tamil text as hate and non-hate speech. Under strict evaluation, our approach achieves impressive results, outperforming baseline models by a considerable margin. The model achieves an F1 score of 0.8634 and good precision, recall, and accuracy, making it a robust solution for hate speech detection in Tamil. The results show that fine-tuning transformer-based models like Tamil-BERT, coupled with techniques like focal loss, can substantially improve performance in hate speech detection for low-resource languages. This work is a contribution to this growing amount of research and provides insights on how to tackle class imbalance for NLP tasks.

1 Introduction

The sudden rise in social networking websites has come with the ever-mounting responsibility of monitoring and regulating harmful content, primarily hate speech. Hate speech is defined as any form of speech that incites violence or is directed against individuals on the basis of race, religion, gender, or any other quality and hence is an emerging menace to the cyber world. Though most of the hate speech detection research has focused on high-resource languages like English, hate speech detection in low-resource languages is not yet explored. Tamil, a Dravidian language with millions of speakers, is a typical example of a low-resource language where hate speech detection tools are nonexistent or inefficient. Recognizing hate speech in Tamil is particularly challenging due to its complex syntax, local dialects, and lack of massive annotated datasets.

Latest advances in natural language processing (NLP) have revealed that transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) represent an effective remedy for text classification issues, such as hate speech recognition. But one of the persistent problems in training such models for hate speech detection is class imbalance. In the majority of datasets, hate speech instances are significantly less than non-hate speech instances, which may lead to model bias and bad generalization.

To address these issues, we propose a novel Tamil hate speech detection method based on Tamil-BERT, the language-specific variant of BERT. We enhance the model's performance with the help of focal loss, which is a technique that allows the model to focus more on the minority class and therefore mitigate the effects of class imbalance. Our contributions are as follows:

1. Developing a deep learning framework for Tamil hate speech detection using a Pretrained transformer model.

2. Addressing class imbalance with focal loss, improving the detection of minority hate speech instances.

3. Evaluating the model on the available Tamil hate speech dataset with a competitive performance of F1 score 0.8634.

The implementation details have been provided in the following GitHub repository:-

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 75–79

https://github.com/Mizan116/LT-EDI-LDK-2025/Hate Speech.

This study builds on the earlier studies of (Chhabra, 2022), who employed transformer models for hate speech detection across multiple languages, and (Zhao, 2020), who demonstrated the efficacy of BERT across low-resource languages. Our study goes one step further by applying these approaches to the special case of Tamil, demonstrating how fine-tuning transformer-based models and focal loss can be used to overcome the specific challenges posed by language-specific characteristics and dataset skew.

2 Related Work

Hate speech detection has attracted significant interest as a consequence of the rapidly rising amount of toxic material on social media platforms. Early approaches to coping with this problem used typical machine learning methods like Support Vector Machines (SVMs) and Naive Bayes (Waseem and Hovy, 2016), where manually designed features like n-grams were common. However, these approaches often struggled to accommodate the complex linguistic and contextual nature of hate speech.

With the arrival of deep learning, models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) emerged, with better performance in classification tasks, especially in recognizing sequential dependencies (Zhang et al., 2018). Yet these models were not yet able to deal with long-range dependencies and deeper context in text, especially in very morphologically complex languages such as Tamil.

The recent progress in Natural Language Processing (NLP) has been transformer-based models, such as BERT (Devlin et al., 2019), which has been the key driver for text classification. Models based on BERT have particularly excelled at hate speech detection since they can learn contextual information through attention mechanisms (Zhang et al., 2020). Other than this, experimentation has also demonstrated that multilingual BERT models can be used for hate speech detection in low-resource languages such as Hindi and Tamil (Chhabra, 2022), thereby demonstrating that transformer models can also be fine-tuned for low-resource languages.

Class imbalance is the greatest challenge in hate speech classification. It has been addressed by focal loss (Lin et al., 2017) that focuses on hard-to-class samples more, thus improving the identification of minority classes like hate speech. Focal loss has shown great promise for its application in NLP such as sentiment analysis and hate speech classification.

In addition, research involving adversarial training (Ta, 2022) and data augmentation strategies like paraphrasing (Bora, 2022) has indicated improvements in model robustness and performance in detecting aggressive language on social media.

This research is grounded on these breakthroughs and uses a Tamil-specific BERT model (Tamil-BERT) and focal loss to address class imbalance in hate speech detection in Tamil. In a related shared task on Dravidian languages, the authors of Rahman et al. (2025) employed transformer models like XLM-R and MuRIL, demonstrating high performance in abusive language detection.

3 Dataset

We have applied the given dataset for Shared Task on Caste and Migration Hate Speech Detection in LT-EDI@LDK 2025 (Rajiakodi et al., 2025), which is focused entirely on caste and migration hate speech detection in Tamil language (Ponnusamy et al., 2024). The dataset is of type two classes: Caste/Migration-related Hate Speech and Non-Caste/Migration-related Hate Speech. The dataset is divided into training, development, and test datasets. The training dataset contains 2,790 samples, and the validation and test datasets contain 598 samples each.

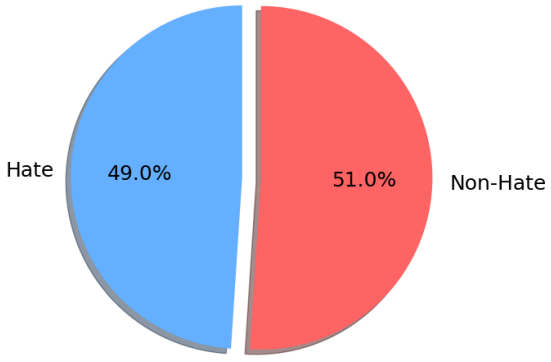

Figure 1: Class Distribution of Train Data

We used the dataset of Rajiakodi et al. (2024) for training, testing and evaluation our model. The distribution of the dataset is as follows:

Split	Hate	Non-Hate	Total Samples
Train	1,366	1,424	2,790
Dev	278	320	598

Table 1: Training and Development Split

Split	Hate	Non-Hate	Total Samples
Test	278	320	598

Table 2: Test Split

The information was collected from Tamil social media, with real-world forms of caste and migration-based hate speech. As there exists class imbalance, whereby non-hate speech instances are greater than hate speech samples, we used focal loss when training the model to give more importance to the minority class.

Text was tokenized using a Tamil-specific tokenizer and padded to 128 tokens. The evaluation metric for the task is macro F1-score to ensure balanced evaluation of both classes.

4 Methodology

In this section, we provide an overview of the methodology and approaches utilized to build the system using the previous Tamil-BERT transformer model. Methodology of our work is shown in Figure 2.

4.1 Preprocessing

The dataset used in this study consists of Tamil language social media texts annotated with binary labels as Hate: 1 and Non-Hate: 0. Preprocessing is crucial to ensure that the model receives clean and consistent input.We have done text cleaning, label encoding, data splitting and tokenization. In data splitting the data split into 80% training and 20% validation. Padding and truncating sequences to a maximum length of 128 tokens.

4.2 Model Selection

We selected Tamil-BERT for hate speech classification due to its strong language specific capabilities and superior performance in Tamil language.The Tamil-BERT model have the highest accuracy, precision, recall and F1-score compared to the other baseline models. The model was fine-tuned using cross-entropy loss and Adam optimizer to avoid overfitting.

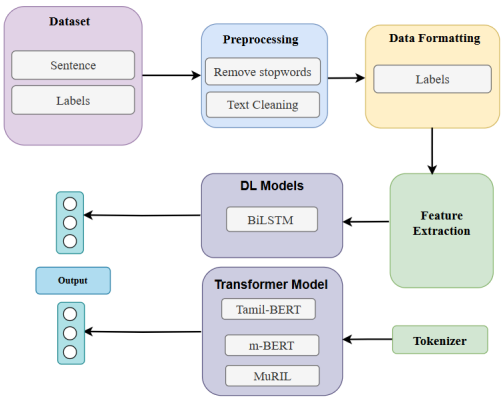

Figure 2: Methodology of our work

4.3 Evaluation and Testing

The Tamil-BERT model was evaluated on a 20% test set using Accuracy, Precision, Recall and F1-Score.A confusion matrix was used for error analysis.Tamil-BERT outperformed baseline models, showing strong generalization and reliable performance.Metrics were computed using macro-averaging to handle class imbalance, ensuring fair evaluation across categories.

5 Results and Analysis

In this section, we evaluate the Tamil-BERT model for hate speech binary classification through four components : parameter setting, comparative analysis, performance metrics and error analysis.

5.1 Parameter Setting

Table 3 shows parameter setting for Tamil-BERT model. In Table 3, lr, optim, bs and wd represent `learning_rate`, optimizer, `batch_size` and `weight_decay` respectively.

Model	lr	optim	bs	ep	wd
Tamil-BERT	2e-5	AdamW	16	8	0.05
MuRIL	3e-5	AdamW	32	7	0.1
m-BERT	2e-5	AdamW	32	5	-
BiLSTM	2e-5	Adam	16	8	-

Table 3: Parameter Setting in different model

5.2 Comparative Analysis

To validate the effectiveness of the proposed Tamil-BERT model, we compared its performance with several baseline models such as MuRIL, m-BERT and BiLSTM. To ensure robustness, we trained all

models across five different random seeds. The mean F1-score for Tamil-BERT was 0.8634, which consistently outperformed all baselines. Each model was trained on the same dataset and evaluated under identical conditions using macro-averaged metrics: Accuracy, Precision, Recall and F1-score.

The result, summarized in Table 4, shows that Tamil-BERT significantly outperforms the other models across all metrics. Here Loss, A, F1 denotes Loss, Accuracy and F1-Score. This is expected as Tamil-BERT is pre-trained specifically for the Tamil language and caste-related content.

Model	Loss	A	F1
Tamil-BERT	0.3571	87.22%	0.8634
MuRIL	0.4725	81.4%	0.77
m-BERT	0.5093	79.8%	0.74
BiLSTM	0.5761	74.2%	0.69

Table 4: Comparison of different models

5.3 Performance Metrics

The performance of various models has been evaluated using various metrics such as Accuracy, F1 Score, Precision, Recall and Confusion Matrix. Figure 3 show the confusion matrics of Tamil-BERT model for Tamil language.

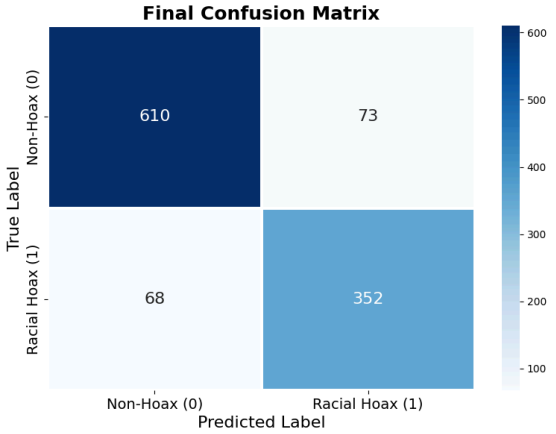

Figure 3: Confusion matrix of Tamil-BERT model

5.4 Error Analysis

The confusion matrix of Tamil-BERT model shows that the model correctly classified 610 negative and 352 positive instances with 73 false positives and 68 false negatives. The overall accuracy of Tamil-BERT is approximately 87.31% and the model achieves precision of 86.99%, recall of 85.85% and an F1-score of 86.34% for the positive class. The other baseline model have lower accuracy, lower precision and recall compare to the Tamil-BERT model. This error occur due to ambiguous or sarcastic language, code-mixed and formal text, implicit hate speech that is expressed indirectly.We also performed an ablation analysis by removing focal loss from the traingin pipleline and replace it with cross-entropy loss. This led to drop of 4.3% in macro F1-score, confirming that focal loss contributes significantly to model performance.

6 Conclusion

In this paper, we proposed a deep learning-based approach for caste and migration hate speech detection in Tamil based on the pre-trained transformer model Tamil-BERT. Using focal loss to handle class imbalance, our method attained a high F1-score of 0.8634, which testifies to its efficiency in classifying hate speech and non-hate speech. Although the model achieved promising performance, there is still room for further improvement, specifically in reducing overfitting and enhancing generalization. Future work will explore added features, better loss functions, and hyperparameter tuning to continue to advance performance. Our study contributes another entry to the growing corpus of work in low-resource hate speech detection and calls for responsible AI innovation to create secure digital spaces.

References

A. Bora. 2022. Data augmentation for hate speech detection using paraphrasing and back-translation. *Proceedings of the 2022 Annual Conference on Natural Language Processing*.

A. Chhabra. 2022. Hate speech detection using transformers: A comparative study. *Conference Name*.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*.

T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. Focal loss for dense object detection. *Proceedings of ICCV 2017*.

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.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of abusive Tamil and Malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 243–247, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

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).

H. T. Ta. 2022. Gan-bert: Adversarial learning for aggressive text detection on social media. *Proceedings of the 2022 IEEE/ACM International Conference on Computer-Aided Design*.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of NAACL-HLT 2016*.

Q. Zhang, K. Zhao, and X. Li. 2020. Bert for hate speech detection: A comparative study. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Z. Zhang, Y. Zhao, and Y. LeCun. 2018. Deep learning for hate speech detection. *Proceedings of the International Conference on Learning Representations (ICLR)*.

D. Zhao. 2020. Multilingual bert for hate speech detection. *Journal Name*.