CUET_NLP_Manning@LT-EDI 2024: Transformer-based Approach on Caste and Migration Hate Speech Detection

Md Ashraful Alam, Hasan Mesbaul Ali Taher, Jawad Hossain,
Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1804061, u1804038, u1704039, u1704057}@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

The widespread use of online communication has caused a significant increase in the spread of hate speech on social media. However, there are also hate crimes based on caste and migration status. Despite several nations efforts to bring equality among their citizens, numerous crimes occur just based on caste. Migration-based hostility happens both in India and in developed countries. A shared task was arranged to address this issue in a low-resourced language such as Tamil. This paper aims to improve the detection of hate speech and hostility based on caste and migration status on social media. To achieve this, this work investigated several Machine Learning (ML), Deep Learning (DL), and transformer-based models, including M-BERT, XLM-R, and Tamil BERT. Experimental results revealed the highest macro f_1-score of 0.80 using the M-BERT model, which enabled us to rank 3^{rd} on the shared task.

1 Introduction

The advent of social media has reshaped the contours of communication, enabling individuals to share their thoughts and interact with a global audience instantaneously. While this has led to the democratization of information exchange, it has also given rise to an insidious byproduct of hate speech and hostility (Sharif et al., 2021). Hate speech, mainly rooted in caste discrimination and migration bias, is a pervasive element in online discourse, highlighting societal prejudices and perpetuating exclusion and animosity. In several nations, caste discrimination remains a persistent issue despite the country's legal strides toward equality (Bhatt et al., 2022). The caste system, an ancient social hierarchy, continues to influence individual and collective identities and relationships, often manifesting in the form of hate speech that targets marginalized communities (Sajlan, 2021). The repercussions of such expressions are not confined to the digital realm; they spill over into the real world, reinforcing social divisions and impeding efforts to establish a more equitable society.

The issue of migration discrimination is similarly problematic, affecting nations worldwide (Chulvi et al., 2023). As people migrate across borders in search of better opportunities or refuge, they often face hostile attitudes and vilification on social media, contributing to xenophobia and nationalism, fostering fear and suspicion, and leading to divisive policies. Thus, addressing these forms of hate speech is crucial, and computational linguistics can help us identify them effectively (Paasch-Colberg et al., 2021).

The goal of this study is to develop a system capable of discerning caste and migration hate speech from non-caste and migration hate speech. The primary accomplishments include:

- Examined various ML, DL, and transformer-based models to detect caste and migration hate speech in Tamil social media, analyzing errors for deeper insights.
- Presented a suitable transformer-based model (M-BERT) tuned with task dataset to classify Tamil text into caste and migration hate speech (CMHS) and not caste and migration hate speech (NCMHS).

2 Related Work

Social media and blogging platforms offer a platform for individual expression, but they can also promote antisocial conduct, such as hate speech and cyberbullying (Hossain et al., 2023). A shared task was conducted (Basile et al., 2019) to detect multilingual hate speech against immigrants and women on Twitter. Almatarneh et al. (2019) used TF-IDF and Lexicon to identify hate speech against migrants and women in English and Spanish tweets, achieving f_1 scores of 0.36 and 0.54, respectively. Romero-Vega et al. (2021) addressed xenophobic

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 238–243
March 21, 2024 ©2024 Association for Computational Linguistics

hate speech in Spanish tweets about Venezuelan migrants in Ecuador, with the SVM model showing the highest performance f_1-score of 0.98. Farooqi et al. (2021) addressed hate speech in social media, emphasizing the need to consider conversation context; their system achieved the highest macro f_1-score of 0.7253 leveraging neural networks and the ensemble of Indic-BERT, XLM-RoBERTa, and Multilingual BERT. A recent study Bhimani et al. (2021) utilized NLP and ML techniques to analyze hate speech on social media, considering aspects such as caste and religion, and gained 96.29% accuracy using Logistic Regression (LR). Sachdeva et al. (2021) addressed the issue of hate speech on social media, underscoring the pressing demand for automated approaches in light of the increasing spread of biased content. They achieved an f_1-score of 0.84 by using the Random Forest (RF) classifier. Dhanya and Balakrishnan (2021) surveyed hate speech detection in Asian languages, focusing on developing an automated system for Malayalam, addressing negativity related to societal factors with varying dataset sizes. Hossain et al. (2022) identified abusive comments from Tamil texts using LR and achieved a f_1-score score of 0.39. Sharif and Hoque (2021) addressed aggressive content on social media, especially in regional languages like Bengali, proposing an ensemble classifier trained on 10,095 annotated texts. Using CNN, BiLSTM, and GRU with diverse embeddings and ensemble strategies, their framework achieved the highest coarse-grained f_1-score of 0.89 and fine-grained weighted f_1-score of 0.84 on the dataset. Despite extensive research in natural language processing, there is a lack of studies on detecting hate speech related to caste and migration.

3 Task and Dataset Description

Due to the complexity of code-mixed data in social media texts, it is challenging for systems trained on monolingual data to classify. This task aims to implement a system to identify hate speech related to caste and migration. In order to detect caste and migration hate speech from text data, task organizers[1] developed a code-mixed (Tamil-Engilsh) corpus. To develop such a system, we analyzed the corpus given by the task organizers (Rajiakodi et al., 2024). Table 1 shows the number of instances for each class in training, validation, and test sets. Datasets are imbalanced, where the number of instances in the NCMHS class is higher compared to the CMHS class.

Classes	Train	Valid	Test	Total Words
NCMHS	3,303	594	973	58,029
CMHS	2,052	351	602	36,654
Total	5,355	945	1,575	94,683

Table 1: Class-wise distribution of train, validation, and test set for the Tamil language

The corpus is split into training (5,355 texts), validation (945 texts), and test (1,575 texts) sets. The task involves a binary classification problem to identify caste and migration hate speech from the corpus. The classes are caste and migration hate speech (CMHS), containing 4,870 texts, and not caste and migration hate speech (NCMHS), containing 3,005 texts.

We analyzed the dataset in further detail concerning sentence length. Figure 1 displays the dataset's length-frequency distribution. According to the length-frequency distribution study, a few text samples had text lengths of more than 100 words. As a result, the maximum sentence length for this work was 100 words. The average sentence length is 18, with one word as the minimum.

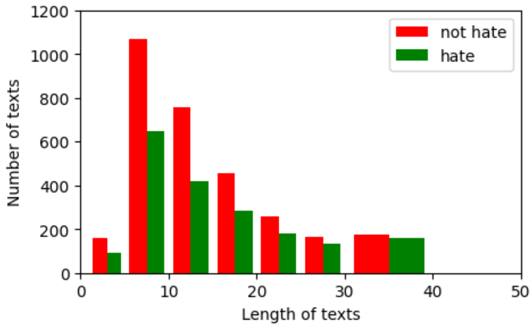


Figure 1: Distribution of sentences frequency in terms length

4 Methodology

Various ML and DL techniques are used for the baseline evaluation with appropriate feature extraction techniques. Moreover, a few transformer models, such as m-BERT, XLM-R, and Tamil-BERT, are examined. Figure 2 depicts a schematic representation of the overall system and employed techniques to tackle the task.

Data Preparation: The corpus text contains unnecessary symbols, punctuation, and letters. Thus, the data in the corpus undergoes a cleaning proce-

[1] https://sites.google.com/view/lt-edi-2024/

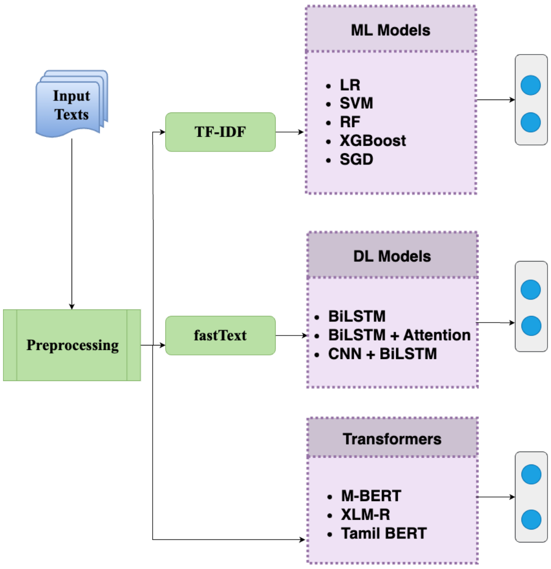


Figure 2: Abstract process of caste and migration hate speech detection in Tamil

dure before system development. This stage prepared a cleaned dataset for the language by removing unnecessary letters, symbols, punctuation, and numbers from the texts. We used this pre-processed data as input for the ML and DL-based models. For transformer-based models, this work used the raw data as input. Additionally, class weighting addresses class imbalance during the model's training.

Textual Feature Extraction: Feature extraction methods are necessary for training classifier models, as ML and DL algorithms cannot learn from raw texts. The TF-IDF technique (Takenobu, 1994) is applied to extract the features for ML models. On the other hand, fastText embeddings (Grave et al., 2018) are used as feature extraction techniques for DL models.

4.1 Classifiers

Six ML, three DL, and three transformer-based models are exploited to classify hate speech in Tamil.

ML-based Classifiers: The suggested system starts with traditional ML approaches such as LR, RF, SGD, and SVM to establish the caste and migration-related hate speech detection system. We chose 'linear' SVM with $C = 10$ and RF. The ensemble approach is built using LR in addition to SVM, Gradient Boosting, and RF. The ensemble method employs the majority voting and stacking techniques. For SGD models, we used the 'log' loss function.

DL-based Classifiers: DL techniques consistently outperformed traditional ML methods. This work uses BiLSTM, Attention, and BiLSTM-CNN to classify hate speech. A 200-cell bidirectional LSTM with 0.2 dropout captures states. The sigmoid function predicts output, and the attention mechanism highlights keywords. The BiLSTM+Attention includes a 20-neuron layer, and CNN+BiLSTM uses 1D convolutional layer (128 filters, kernel 3), bidirectional LSTM (256 units, 0.3 dropouts), and embedding (128). Flattening and dense layers conclude with sigmoid activation for classification. In this work, we used *optuna* (Akiba et al., 2019) for finding the optimal hyper-parameters.

Transformer-based Classifiers: Transformers have grown in popularity in recent years due to their exceptional performance in nearly every NLP domain. As the given dataset consists of code-mixed texts, we choose three transformers such as M-BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019), and Tamil-BERT (Joshi, 2022) to develop our models. A self-supervised cross-lingual understanding training method called XLM-R is beneficial for low-resourced languages. The transformer model m-BERT, on the other hand, has been pre-trained in more than 104 languages. Tamil-BERT is a type of BERT designed explicitly for the Tamil language. It is trained on a large corpus of Tamil text to improve monolingual understanding and natural language processing tasks for Tamil speakers. These models were extracted from the Huggingface[2] transformer library and fine-tuned on our dataset with the Ktrain (Maiya, 2022) package. To fine-tune those models, we used the 'fit_onecycle' method with a learning rate of $2e^{-5}$. All the models have trained up to 15 epochs, with batch size 12.

5 Results and Analysis

Table 2 demonstrates the performance of the various methods employed on the test set. The models dominance is determined by the macro f_1-score. On the other hand, we closely monitor the other metrics, including macro recall (R) and macro precision (P) scores. These additional measures comprehensively evaluate the models performance across different aspects.

The results showed that the LR and SVM models obtained a macro f_1-score of 0.75. When trained

[2] https://huggingface.co/

Methods	Classifiers	P	R	MF1
ML Models	LR	0.7489	0.7308	0.7531
	SVM	0.7512	0.7248	0.7509
	RF	0.7439	0.7908	0.7589
	XGB	0.6337	0.6892	0.6309
	SGD	0.7143	0.7798	0.7275
	Ensemble	0.7931	0.7452	0.7629
DL Models	BiLSTM	0.7473	0.7429	0.7490
	BiLSTM + Attention	0.6952	0.6438	0.6418
	BiLSTM + CNN	0.7671	0.7342	0.7409
Transformer	**M-BERT**	**0.7823**	**0.8246**	**0.8049**
	XLM-R	0.7598	0.7647	0.7638
	Tamil BERT	0.7794	0.7849	0.7847

Table 2: Performance of various models on the test set. The acronyms P, R, and MF1 denote Precision, Recall, and macro f_1-score.

on fastText feature vectors, the BiLSTM approach yielded a macro f_1-score of 0.74. Deep learning-based models obtained comparatively worse results than the ML-based models. The small size of the training data maybe the reason behind this. Transformer-based models outperformed all other models. M-BERT obtained the best performance, macro f_1-score of 0.80.

5.1 Error Analysis

We performed an in-depth error analysis to get insights into the best-performed model (M-BERT) performance using quantitative and qualitative analysis.

5.1.1 Quantitative Analysis

Table 2 shows that M-BERT is the best-performing model for detecting hate speech related to caste and migration in the given dataset. The confusion matrix (Figure 3) of the best-performing model shows that a total 1,211 number of labels were classified correctly.

Misclassified hate/Non-hate labels totaled 301, with 169 NCMHS and 132 CMHS texts. This is likely due to data imbalance and the dataset's diverse languages (English, Tamil, code-mixed, and code-switched), hindering the models pattern recognition. The misclassification hints at nuanced contextual factors, posing challenges in differentiating between hate and non-hate labels.

5.1.2 Qualitative Analysis

Figure 4 illustrates a few predicted outcomes by the best model on the test dataset. Samples 2 and 3 are among those that have been classified correctly.

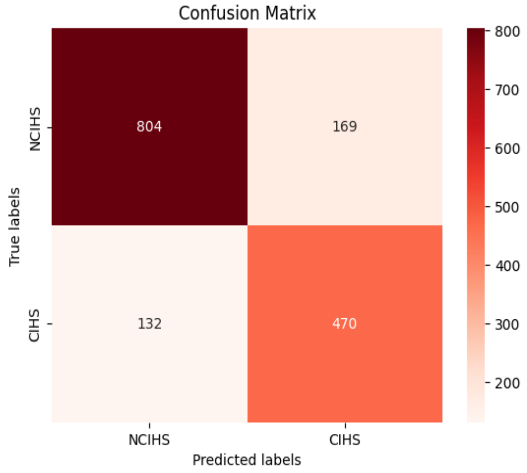
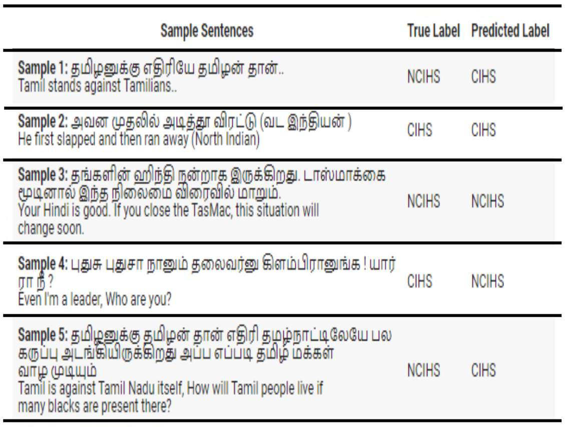


Figure 3: Confusion matrix of the best-performed model (M-BERT) for the task

Sample 1 is incorrectly classified as caste and migration hate speech, whereas sample 4 is classified wrongly as not caste and migration hate speech. These are just two examples of situations where the model misclassified data. This misclassification may have happened due to the imbalanced nature of the dataset. Additionally, the model needed help to classify the text because the corpus contained code-mixed data. These subtleties emphasize the value of qualitative analysis in figuring out how the model functions in certain situations.

Limitations

This study evaluated various transformers, ML, and DL models where M-BERT showed promising per-



Sample Sentences	True Label	Predicted Label
Sample 1: தமிழனுக்கு எதிரியே தமிழன் தான்.. Tamil stands against Tamilians..	NCIHS	CIHS
Sample 2: அவன முதலில் அடித்தூ விரட்டு (வட இந்தியன்) He first slapped and then ran away (North Indian)	CIHS	CIHS
Sample 3: தங்களின் ஹிந்தி நன்றாக இருக்கிறது. டாஸ்மாக்கை மூடினால் இந்த நிலைமை விரைவில் மாறும். Your Hindi is good. If you close the TasMac, this situation will change soon.	NCIHS	NCIHS
Sample 4: புதுசு புதுசா நானும் தலைவர்னு கிளம்பிரானுங்க ! யார்ரா நீ ? Even I'm a leader, Who are you?	CIHS	NCIHS
Sample 5: தமிழனுக்கு தமிழன் தான் எதிரி தமழ்நாட்டிலேயே பல கருப்பு அடங்கியிருக்கிறது அப்ப எப்படி தமிழ் மக்கள் வாழ முடியும் Tamil is against Tamil Nadu itself, How will Tamil people live if many blacks are present there?	NCIHS	CIHS

Figure 4: Some predicted outcomes by the best-performed model

formance detecting hate speech in Tamil. However, it struggled to detect caste and migration hatred due to limited training data. The dataset included social media content featuring regional dialects and poor grammar, posing challenges for identifying hate classes. Additionally, ambiguous statements and context gaps may affect the models performance. Enhanced methods for collecting nuanced grammar details could improve the performance of the current implementation.

6 Conclusion

This work explored several ML, DL, and transformer-based techniques and analyzed their performance in detecting caste and migration hate speech in Tamil. Experimental assessment of the test dataset revealed that the M-BERT model is the best performing model for detecting hate speech in Tamil and outperformed all models by obtaining the highest macro f_1-score (0.80). Surprisingly, the BiLSTM + Attention model performed poorly compared other ML and transformer models. These inferior results might occur because of the prevalence of local words, which still need to be discovered in the model. The future work includes adding more data in the respective classes to make a balanced dataset and investigating more sophisticated techniques such as MuRIL and GPT for improved performance.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Sattam Almatarneh, Pablo Gamallo, and Francisco J Ribadas Pena. 2019. CiTIUS-COLE at semeval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 387–390.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing Fairness in NLP: The Case of India.

Darsh Bhimani, Rutvi Bheda, Femin Dharamshi, Deepti Nikumbh, and Priyanka Abhyankar. 2021. Identification of Hate Speech using Natural Language Processing and Machine Learning. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–4. IEEE.

Berta Chulvi, Mariangeles Molpeceres, María F Rodrigo, Alejandro H Toselli, and Paolo Rosso. 2023. Politicization of Immigration and Language Use in Political Elites: A Study of Spanish Parliamentary Speeches. *Journal of Language and Social Psychology*, page 0261927X231175856.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

LK Dhanya and Kannan Balakrishnan. 2021. Hate speech detection in Asian languages: a survey. In *2021 international conference on communication, control and information sciences (ICCISc)*, volume 1, pages 1–5. IEEE.

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. COMBATANT@ TamilNLP-ACL2022: Fine-grained Categorization of Abusive Comments using Logistic Regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228.

Jawad Hossain, Hasan Mesbaul Ali Taher, Avishek Das, and Mohammed Moshiul Hoque. 2023. NLP_CUET at BLP-2023 Task 1: Fine-grained Categorization of Violence Inciting Text using Transformer-based Approach. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 241–246.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Arun S Maiya. 2022. ktrain: A low-code library for augmented machine learning. *The Journal of Machine Learning Research*, 23(1):7070–7075.

Sünje Paasch-Colberg, Christian Strippel, Joachim Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, 9(1):171–180.

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.

Raúl R Romero-Vega, Oscar M Cumbicus-Pineda, Ruperto A López-Lapo, and Lisset A Neyra-Romero. 2021. Detecting xenophobic hate speech in spanish tweets against venezuelan immigrants in ecuador using natural language processing. In *Applied Technologies: Second International Conference, ICAT 2020, Quito, Ecuador, December 2–4, 2020, Proceedings 2*, pages 312–326. Springer.

Janak Sachdeva, Kushank Kumar Chaudhary, Harshit Madaan, and Priyanka Meel. 2021. Text based hate-speech analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 661–668. IEEE.

Devanshu Sajlan. 2021. Hate Speech against Dalits on Social Media. *CASTE: A Global Journal on Social Exclusion*, 2(1):77–96.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Align and Conquer: An Ensemble Approach to Classify Aggressive Texts from Social Media. In *2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pages 82–86. IEEE.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Combating hostility: Covid-19 fake news and hostile post detection in social media. *arXiv preprint arXiv:2101.03291*.

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