IIITDWD_SVC@DravidianLangTech-2024: Breaking Language Barriers; Hate Speech Detection in Telugu-English Code-Mixed Text

Chava Srinivasa Sai[1] **Rangoori Vinay Kumar**[1] **Sunil Saumya**[1] and **Shankar Biradar**[1]
[1]Department of Data Science and Intelligent Systems,
Indian Institute of Information Technology, Dharwad, Karnatka, India
(srinivassaichava,vinaykumarrangoori33)@gmail.com
(shankar,sunil.saumya)@iiitdwd.ac.in

Abstract

Social media platforms have become increasingly popular and are utilized for a wide range of purposes, including product promotion, news sharing, accomplishment sharing, and much more. However, it is also employed for defamatory speech, intimidation, and the propagation of untruths about particular groups of people. Further, hateful and offensive posts spread quickly and often have a negative impact on people; it is important to identify and remove them from social media platforms as soon as possible. Over the past few years, research on hate speech detection and offensive content has grown in popularity. One of the many difficulties in identifying hate speech on social media platforms is the use of code-mixed language. The majority of people who use social media typically share their messages in languages with mixed codes, like Telugu–English. To encourage research in this direction, the organizers of *DravidianLangTech@EACL-2024* conducted a shared task to identify hateful content in Telugu-English code-mixed text. Our team participated in this shared task, employing three different models: Xlm-Roberta, BERT, and Hate-BERT. In particular, our BERT-based model secured the 14^{th} rank in the competition with a macro F1 score of 0.65.

1 Introduction

In contemporary society, social media plays a pivotal role in the daily lives of many individuals. Text messages across various platforms hold considerable influence, both positively and negatively. On a positive note, social media serves as a global connector, fostering creativity, enhancing skills, and providing entertainment. Additionally, it facilitates the swift dissemination of breaking news. Conversely, the prevalence of hate speech and the dissemination of inaccurate information about individuals, groups, or societies represent undesirable phenomena. Social media platforms are regrettably exploited for expressing destructive views and eliciting negative emotions through hate and fraudulent communications.

In the present era, there is a high degree of trust in social media, so misinformation propagated by media outlets or influential figures is often accepted as true. Consequently, individuals disseminate false information using inappropriate language, with hashtags like *#HateSpeech* gaining prominence on platforms such as Twitter and YouTube, particularly during the emergencies like COVID-19 pandemic. Furthermore, some individuals erroneously believe that engaging in abusive language or hate speech can confer fame and notoriety. Social media platforms are actively striving to eradicate such negative textual content, recognizing the severe consequences it can have on individuals' lives.

While social media users are increasingly cognizant of the issue, exposure to hate news persists, even when the true story is known. Efforts to address this problem involve the development of machine learning and deep learning models capable of identifying hate speech in text data (Nozza, 2021). Given the language's global prevalence, numerous models have been trained on English data (Santosh and Aravind, 2019). However, it is imperative to acknowledge that hate speech extends beyond English, with regional languages being utilized for its propagation. Telugu, a Dravidian language spoken in Andhra Pradesh and Telangana, India, is one such language.

Motivated by this realization, *DravidianLangTech@EACL-2024* initiated a shared task for the classification of hate and non-hate speech detection in Tenglish (Telugu-English) code-mixed dataset (B et al., 2024). Our team participated in the shared task; we employed various techniques, including transliteration, and translation during preprocessing. In addition, we subsequently utilized three distinct models for embedding extraction,

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 119–123

HateBERT, XLM-RoBERTa, and BERT, and secured 14th position with a F1-Score of 0.6565 for BERT-Based (cased) among all competing teams.

The article is structured as follows: Section 2 presents the background study. The details of the dataset and methodology are presented in Section 3, and finally, the results are discussed in Section 4. At last, in Section. 5 concluded and talked about Future research direction. We have written some Ethics in Section 6 for the work which had done.

2 Related work

The exploration of hate speech detection in Dravidian code-mixed text remains a relatively under explored topic, as most previous research has predominantly focused on high-resource languages such as English. However, recent attention from the research community has been directed towards hate speech detection in Dravidian code-mixed text data (Chakravarthi et al., 2020).

The gold standard corpus for detecting hate speech in three Dravidian languages: Tamil (Chakravarthi et al., 2020), Malayalam (Chakravarthi et al.), and Kannada (Hande et al., 2020) was developed by (Chakravarthi et al., 2020). This corpus was established as part of a shared task, stimulating active engagement from multiple teams. The majority of these teams concentrated on leveraging knowledge derived from pre-trained transformer models to address the challenges associated with low-resource languages. (Biradar et al., 2021; Fharook et al., 2022; Kavatagi et al., 2023) for instance, utilized a cross-lingual pre-trained model like Mbert in conjunction with Support Vector Machines (SVM) to identify hate speech in Tanglish and Manglish text. Furthermore, (Saumya et al., 2022) adopted an ensemble setup, combining machine learning (ML) and deep learning (DL) based models to effectively detect transphobic content in Tamil and Malayalam text. However, Telugu, being one of the major Dravidian languages widely spoken in Telangana and Andhra Pradesh, has been relatively less explored in this context. This marks the first attempt to identify hate content in Telugu-English code-mixed text.

3 Methodology

3.1 Task and Data

The *DravidianLangTech@EACL-2024* shared task has been a significant focus of our work. The organizers of this shared task have made available a comprehensive dataset consisting of 4000 and 500 comments in the train and test stages respectively. These comments were collected from Youtube, as stated by the organizers (B et al., 2024). The main objective of this task is to classify each Telugu-English code-mixed social media comment at the sentence level, determining whether it falls into the hate or non-hate categories. Our team actively participated in this task and achieved an impressive rank of 14^{th} position. For more detailed information about this dataset, plese refer to Table 1

	Hate	**Non-hate**	**Total**
Train	1939	2061	4000
Test	250	250	500

Table 1: Data distribution

3.2 Data pre processing

Preprocessing raw data is essential in optimizing it for compatibility with machine learning models. Even a small adequate data preprocessing improves a model's efficiency significantly. The different kinds of data preprocessing methods we used as illustarted in Figure 1.

3.2.1 Transliteration

Transliteration is a process that does not alter the meaning of a sentence; instead, it modifies the words to facilitate pronunciation in the reader's native language (Deselaers et al., 2009). Our dataset comprises Telugu content presented in the English script. In this context, we employed a transliteration model designed to accurately convert the Tenglish (Telugu-English) script into the Telugu script. The IndicXlit [1] model was selected for this task; this model proficiently transliterates Tenglish into standard Telugu, enhancing comprehension and facilitating progress in subsequent stages of the project.

3.2.2 Translation

Following the transliteration process, we now possess high-quality Telugu text, which needs to be translated into English. This step is essential because the majority of the pre-trained models were trained on English datasets. Therefore, we have implemented a translation approach. We employed the "IndicTrans2"[2] model for translating the Tel-

[1]https://ai4bharat.iitm.ac.in/indicxlit-model/
[2]https://ai4bharat.iitm.ac.in/indic-trans2/

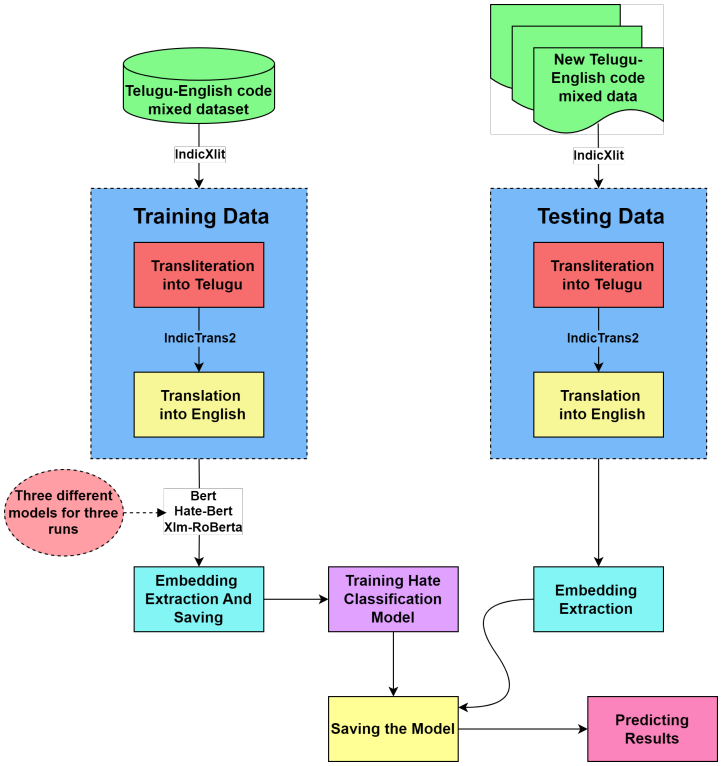

Figure 1: Flow diagram for the proposed work

ugu text into English. Table 2 illustrates the sample comments from translated text.

3.2.3 Tokenization

After translation , our training dataset is ready for tokenization , we loaded three different tokenizers for these models using the Hugging Face Transformers library[3]. We specifically used the AutoTokenizer class from that package to load the appropriate tokenizer for the chosen model architecture.

3.3 Feature Extraction

Tokenized text is subsequently employed for feature extraction. For this purpose, we deployed three encoder-based models with frozen weights: Bert, Hate-bert, and XLM-Roberta. The next step involves extracting embeddings using the mean-pooling approach, a widely adopted method in NLP applications that involve neural networks and word embedding. This method is employed to obtain a comprehensive representation by averaging embedding vectors along specific directions.

3.3.1 BERT

The BERT model, based on transformer architecture, has been widely adopted for its pretraining capabilities (Kenton and Toutanova, 2019). BERT is chosen in the proposed work due to its comprehensive language understanding capabilities. The embeddings from the CLS token are utilized in the proposed work to generate sentence-level representations. Specifically, 'bert-base-uncased' from the Hugging Face library[4] is employed to generate these sentence representations in the proposed work.

3.3.2 Xlm-RoBERTa

Xlm-RoBERTa is a widely-used RoBERTa model that supports multiple languages(English, Hindi etc.,). The model has been trained on a larger corpus comprising 100 different languages (Conneau et al., 2020). The present study utilizes the "xlm-roberta-base"[5] for comprehending cross-lingual representations.

3.3.3 Hate-BERT

It is a variant of the BERT model specifically designed for detecting abusive language in English text. This variant was derived through extensive training on the BERT uncased model, utilizing over one million posts that were banned in *Reddit* communities. Notably, this model has demonstrated superior performance compared to the original BERT

[3]https://huggingface.co/models

[4]https://huggingface.co/bert-base-uncased

[5]https://huggingface.co/xlm-roberta-base

Original comment	Translated text
Thappu chesina vaallaku vanike kaadu inka anni modalithavi . Enta kaalam students life tho aadukuntu crores earn chedtharu illegal ga.	It is not only a pity for the wrong doers, but also a first step towards them. How long will you play with the life of a student and play with him / her?
Kanipinche devudu CBN	The Seeing God CBN
Pavan Kalyan gari nayakatvam vardillali jai power star	Power Star Pawan Kalyan

Table 2: Sample comments for translated text

in understanding offensive language. The proposed method employs the "GroNLP/hateBERT"[6] model from Hugging Face to generate domain-specific representations.

3.4 Classifier

The generated features are subsequently passed through the final stage of our pipeline, which is the classifier for hate or non-hate class detection. This classifier remains consistent across all three models.

The proposed classifier is constructed using a simple feed-forward neural network with three layers: the input, hidden, and output layers. The size of the input layer is contingent upon the dimensions of the model embeddings. Proceeding to the hidden layer, it consists of a non-linear layer comprising 128 neurons and utilizes the Rectified Linear Unit (ReLU) activation function. This non-linear aspect allows the model to discern intricate patterns in the data, enhancing its capacity for producing more accurate predictions. A final sigmoid layer is incorporated to predict the output class. Subsequently, the model undergoes training for nine epochs employing the *Binary Cross-Entropy Loss* with the *Adam optimizer*.

4 Result and discussion

The proposed model was tested on three distinct embedding representations generated using pre-trained encoder-based models. The comparative results between these models are presented in Table 3. According to Table 3, the BERT and HateBERT-based models demonstrate a superior ability to comprehend the hate and non-hate nature of the text, achieving comparable results of F1-Score 0.6565 for BERT-Based (cased) on translated text data .

[6]https://huggingface.co/GroNLP/hateBERT

	Hate (F1)	Non Hate (F1)	Accuracy
HateBERT	**0.68**	**0.70**	**69**
BERT	**0.68**	**0.71**	**69**
XlmRoBERTa	0.64	0.64	64

Table 3: Comparative results

Team	F1 score	Rank
Sandalphon	0.7711	1
Selam	0.7711	2
Kubapok	0.7431	3
DLRG1	0.7101	4
IIITDWD_SVC	**0.6565**	**14**

Table 4: Leader board

In contrast, Xlm-RoBERTa slightly lags behind, likely due to the high-resource English text.

Our team presented the results of the best-performing model, BERT, in the competition. The organizers of the shared task evaluated model performance using the macro F1 score, and our team secured the 14^{th} rank among the participating teams. Table 4 present the leaderboard, depicting the position of our team in the competition.

5 Conclusion and Future research direction

The study provides the working notes of the model presented during the DravidianLangTech-2024 shared task. The experimental findings suggest that the performance of the model can be enhanced by translating the original code-mixed text and leveraging the knowledge derived from monolingual pre-trained models. Additionally, this work can be extended to incorporate the fine-tuning of language models using domain-specific data.

6 Ethics

In our study on detecting hate speech in Telugu-English code-mixed text, ethical considerations have played a crucial role. We have been careful in using language models as we have openly shared our techniques, models, and findings. Our participation in the *DravidianLangTech@EACL-2024* shared task has also been conducted with ethical standards in mind, promoting collaboration and knowledge exchange within the research community. We are committed to responsible AI practices and continuously strive to reduce biases and ensure fair representation in hate speech detection.

References

Premjth B, Bharathi Raja, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Janakiram Chandu. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. mbert based model for identification of offensive content in south indian languages.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. A sentiment analysis dataset for code-mixed malayalam-english. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 177.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Thomas Deselaers, Saša Hasan, Oliver Bender, and Hermann Ney. 2009. A deep learning approach to machine transliteration. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 233–241.

Shaik Fharook, Syed Sufyan Ahmed, Gurram Rithika, Sumith Sai Budde, Sunil Saumya, and Shankar Biradar. 2022. Are you a hero or a villain? a semantic role labelling approach for detecting harmful memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 19–23.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Sanjana M Kavatagi, Rashmi R Rachh, and Shankar S Biradar. 2023. Vtubgm@ lt-edi-2023: Hope speech identification using layered differential training of ulmfit. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 209–213.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

TYSS Santosh and KVS Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 310–313.

Sunil Saumya, Vanshita Jha, and Shankar Biradar. 2022. Sentiment and homophobia detection on youtube using ensemble machine learning techniques. In *Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR*.