Lidoma@DravidianLangTech 2024: Identifying Hate Speech in Telugu Code-Mixed: A BERT Multilingual

Muhammad Tayyab Zamir, Moein Shahiki Tash, Zahra Ahani, Alexander Gelbukh, Girigori Sidorov
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC) Mexico
Corresponding: mzamir2023@cic.ipn.mx

Abstract

Over the past few years, research on hate speech and offensive content identification on social media has been ongoing. Since most people in the world are not native English speakers, unapproved messages are typically sent in code-mixed language. We accomplished collaborative work to identify the language of code-mixed text on social media in order to address the difficulties associated with it in the Telugu language scenario. Specifically, we participated in the shared task on the provided dataset by the Dravidian-LangTech Organizer for the purpose of identifying hate and non-hate content. The assignment is to classify each sentence in the provided text into two predetermined groups: hate or non-hate. We developed a model in Python and selected a BERT multilingual to do the given task. Using a train-development data set, we developed a model, which we then tested on test data sets. An average macro F1 score metric was used to measure the model's performance. For the task, the model reported an average macro F1 of 0.6151.

1 Introduction

India is a multilingual nation with a diverse linguistic past in South Asia. As the official and administrative language of the state of Andhra Pradesh in southern India, Telugu is one of the most widely spoken languages in the country. having 96 million or more Telugu speakers as native speakers(tel).

In addition to their native, local, or regional tongue, many in the region feel at ease utilizing English for daily communication. These multilingual people prefer to share their thoughts, opinions, and comments on social media sites in several scripts and/or languages, which makes code-mixing the norm on social media(Priyadharshini et al., 2023b; Chakravarthi et al., 2021; Priyadharshini et al., 2023a). The spread of hate speech(Yigezu et al., 2023b; Shahiki-Tash et al., 2023) and objectionable content has far-reaching effects, increasing tensions, encouraging discrimination, and widening societal divisions as social media and online platforms become an essential part of daily life in India. In light of the pressing need to address such content, this study attempts to manage the complexities of Telugu codemixed (B et al., 2024; Yigezu et al., 2022) language by utilizing sophisticated natural language processing (NLP) models, most notably BERT (Bidirectional Encoder Representations from Transformers) (Bade, 2021; Tonja et al., 2022). Sentiment analysis (Kanta and Sidorov, 2023; Tash et al., 2023; Bade and Afaro, 2018), a powerful tool in natural language processing, often focuses on discerning emotions conveyed in the text. When applied to hate speech, it plays a crucial role in understanding the underlying sentiment behind abusive language, shedding light on the detrimental impact of hateful expressions within the digital sphere.

The results of this work aim to establish a basic framework for continued attempts to prevent the spread of damaging content, provide platforms with efficient tools for moderation, and foster a more favorable and supportive online environment among the diverse range of languages found in India's digital space.

2 Related work

The identification of hate speech has grown in importance in the social media and internet communication era. Due to the increase in hate speech occurrences, academics are investigating many approaches to effectively address this problem, such as deep learning (Yigezu et al., 2023a; Ahani et al., 2024),

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 101–106
March 22, 2024 ©2024 Association for Computational Linguistics

transformer-based models, Convolutional Neural Network (Bade and Seid, 2018; sha), and machine learning (Tash et al., 2022).

Traditional machine learning techniques played a major role in the early stages of hate speech identification, laying the groundwork for later studies in the area. Davidson et al(Devlin et al., 2018) made a significant addition in 2017 by offering a data set and a number of features created especially for the detection of hate speech. This groundbreaking discovery launched a trajectory of developments in the field and signaled the beginning of systematic hate speech detection research.

A crucial element of conventional machine learning methodologies was the feature engineering process. Scholars employed attributes like n-grams, sentiment analysis, and lexical aspects to efficiently depict textual content. Sentiment analysis assessed the text's emotional tone, whereas N-grams in particular demonstrated how language is sequential. Lexical features, which include language patterns and vocabulary. Hate speech detection research was first driven by traditional machine learning techniques, which were crucial in laying the groundwork for further advancements and offering insightful information. The advent of more sophisticated strategies was spurred by these systems' limits in addressing language complexity and context, despite their promising outcomes.

Zhang et al (Zhang and LeCun, 2015) introduced a Convolutional Neural Network (CNN) model presenting a novel method for detecting hate speech. This model outperformed other approaches in terms of performance. Because hate speech frequently uses certain phrases, keywords, and linguistic clues, CNNs are particularly good at identifying local patterns within the text. In 2018 (Ribeiro et al., 2018) presented a hierarchical attention-based model This strategy focused on attention mechanisms and hierarchical representations in order to address the need to record nuanced hate speech (Mathew et al., 2021) . A more thorough examination of the substance of hate speech was made possible by the use of hierarchical attention models, which made it possible to examine data at the word and sentence levels.

A breakthrough in natural language processing, BERT (Bidirectional Encoder Representations from Transformers) was introduced by Devlin et al (Devlin et al., 2018) in 2019 . The novel aspect of BERT is its capacity to comprehend a word's context by taking the complete phrase into account. This contextual awareness is especially important in the complex and frequently subtle realm of hate speech. BERT can more precisely and thoroughly identify hate speech by collecting the entire context (Dowlagar and Mamidi, 2021) . The promise of BERT in the area of hate speech identification was immediately recognised by researchers. They efficiently used BERT's potent language understanding capabilities to discern between hateful and non-hateful information by honing it on hate speech data sets (Khanduja et al.). Modern outcomes in hate speech identification can be attributed to this adaptation.The Transformer family has grown ever since BERT was introduced.

3 Data set and Task description

The task at hand focuses on hate speech classification within a data set encompassing 4000 sentences expressed in Telugu, represented both in native script and Romanized forms. Within this data set, 2061 sentences are categorized as non-hate, while 1939 sentences are designated as hate in the training set[1].Moreover, an additional test data set containing 500 sentences is provided, lacking categorized labels. The primary goal of this task is to employ BERT multilingual model to discern patterns from the labeled training data in order to predict whether the 500 test sentences fall into the categories of hate speech or non-hate (Priyadharshini et al., 2023a; B et al., 2024).

This classification task presents a significant challenge in analyzing and identifying hate speech within Telugu text, considering the multilingual aspect involving both native script and Romanized forms. With a substantial dataset comprising labeled examples of hate and non-hate speech, machine learning models can be trained to recognize intricate patterns, linguistic nuances, and context-specific features associated with hate speech (Marreddy et al., 2022).

[1] https://codalab.lisn.upsaclay.fr/competitions/16095

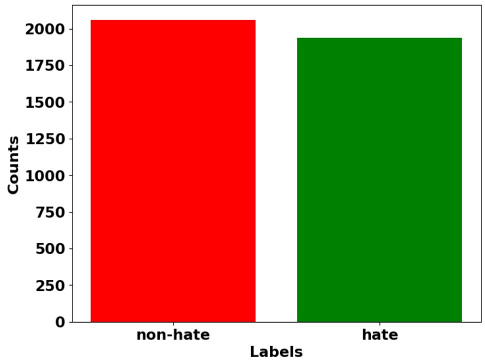

Figure 1: counts for hate and non in training data set

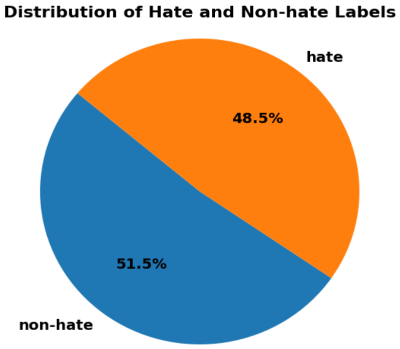

Figure 2: Percentage Distribution of Labels

4 Data preprocessing

During the preprocessing phase of our data, a series of crucial steps were undertaken to enhance the quality and applicability of the text data for hate speech identification and other Natural Language Processing (NLP) tasks. A key focus was placed on text preprocessing, involving the removal of special characters, numeric digits, and emoticons. Emojis, despite often expressing emotions or nonverbal cues, were eliminated from the text due to their limited contribution to semantic meaning in text analysis. Moreover, special characters and numerical digits were also omitted from the text corpus. These elements were considered as they could potentially complicate later processing stages and generally convey minimal linguistic information within NLP tasks. The primary objective behind this meticulous cleaning procedure was to streamline the text data, reducing noise and irrelevant features, while retaining a more refined and cohesive representation suitable for subsequent analysis or model development.

5 Training the Model

5.1 Model Description

In our hate and offensive language task, we utilized the BERT (Bidirectional Encoder Representations from Transformers) multilingual Transformer (Devlin et al., 2018), a sophisticated language model renowned for its contextual understanding of text across various languages. Through the adoption of this different model architecture, our approach involved comprehensive experimentation aimed at fully harnessing the capabilities of this advanced technology within our task domain. The overarching goal was to create a robust and precise system for detecting and categorizing hate speech accurately.

By leveraging the BERT multilingual Transformer model, our objective revolved around developing a highly capable system capable of recognizing and effectively classifying hate speech content. Through thorough exploration and experimentation with this model, our focus was on identifying the most optimal architecture and configurations that would yield superior performance in the identification and mitigation of hate and offensive content within textual data. This process involved fine-tuning the model parameters, experimenting with various training methodologies, and optimizing the model's ability to comprehend and categorize hate speech expressions, ultimately aiming for heightened accuracy and efficiency in the detection and classification process. The utilization of the BERT multilingual Transformer represented our concerted effort to leverage cutting-edge technology, exploring its potential to enhance the efficacy of hate speech identification systems through state-of-the-art natural language understanding and classification capabilities (Sohn and Lee, 2019).

5.2 Training the Model

5.2.1 Data Splitting

The data set is initially divided into a training set and a validation set and testing set, where in 70 percent of the labeled data is allocated for training the BERT multilingual model and 10 percent for validation. This substantial portion serves as the foundation for the model to learn and extract patterns, linguistic nuances, and hate speech indicators from the provided Telugu codemixed text. The model undergoes the training process using this data to adjust its parameters and optimize its understanding of hate speech expressions.

Simultaneously, a smaller subset, constituting 20 percent of the labeled data set, is set aside as the validation set. This portion is crucial for fine-tuning the model's performance and validating its effectiveness. The validation set assists in adjusting hyper parameters, evaluating the model's performance on unseen data, and preventing over fitting, ultimately enhancing the model's generalized. It provides a means to measure how well the model learns from the training data and how effectively it can predict hate speech occurrences in new, unseen instances of codemixed Telugu text.

Finally, the unlabeled test data, separate from the training and validation sets, serves as a means to assess the model's real-world performance. This data set, containing instances of Telugu codemixed text without labeled categories, enables the evaluation of how well the trained BERT multilingual model can generalize its learning and accurately.

5.2.2 Training Parameters for the Model

The training process of the BERT multilingual model involves several critical parameters aimed at optimizing its learning from the data set while ensuring computational efficiency and convergence stability. Primarily, the choice of 3 epochs for training iterations indicates that the entire labeled data set is iterated through the model 3 times.

The batch size, set at 32, determines the number of data samples processed simultaneously in each iteration during training. The learning rate, specified as 1e-5, governs the size of parameter updates during training. A lower learning rate typically facilitates more precise updates but might prolong the training process.

Additionally, determining the update size involves settings that regulate how the model's parameters are adjusted based on the calculated gradients during training. These settings aim to strike a balance between stability and efficiency during the model's learning process.

While these parameters have been set to strike a balance between computational efficiency and convergence stability, optimizing these settings might further enhance the model's performance in recognizing hate and offensive content. Fine-tuning parameters such as the learning rate, batch size, training epochs, or update size could potentially refine the model's accuracy.

6 Evaluation Metrics and Results

The F1-score, computed as the harmonic mean of precision and recall, provides a balanced assessment by considering both metrics. Achieving a macro F1-score of 0.6151 in our task indicates a moderate level of overall performance, suggesting a reasonable balance between precision and recall across multiple classes. This metric signifies the model's effectiveness in correctly identifying hate and offensive content in a multi-class classification scenario, highlighting its general capability in accurately categorizing various classes within the data set.

7 Error Analysis

The BERT multilingual model applied to Telugu hate speech exhibits notable accuracy, particularly in correctly identifying true positives. However, a significant challenge arises with false positives, misclassified instances even in a balanced dataset. This pattern necessitates thorough analysis and adjustments in the model's discriminatory capabilities. Comprehensive evaluations on validation and test sets are essential for assessing the model's adaptability. Proposed strategic modifications involve fine-tuning parameters and scrutinizing false positive occurrences to enhance overall accuracy and efficacy.

8 Limitations

Challenges in a Telugu language dataset include limited resources hindering preprocessing techniques. The dynamic nature of evolving fake news poses adaptability issues for models trained on historical data. Identifying relevant features for effective training is challenging, particularly in a language with unique linguistic characteristics not well-captured by standard NLP techniques.

9 Conclusion

Utilizing the BERT model, this study focuses on detecting Hate and Offensive Language within Telugu Codemixed Text, achieving a macro F1-score of 0.6151. It showcases the model's proficiency in identifying hate speech amidst the intricate linguistic composition of Telugu codemixed text. Despite this success, the research underscores the imperative for continual improvement in both model architecture and data set expansion to heighten the accuracy of hate speech detection. The study serves as a foundational milestone, laying the groundwork for future advancements. It sets a benchmark for the development of more sophisticated and sensitive systems crucial for accurately identifying and mitigating harmful information present in multilingual digital realms. This work not only validates the potential of the BERT model but also emphasizes the ongoing need for refinement and innovation in combating hate speech in diverse linguistic contexts.

10 Future work

The future work for this task involves enhancing the existing framework through various approaches. It includes refining model architectures tailored for codemixed languages, diversifying and augmenting data sets, fine-tuning model parameters, exploring multimodal approaches, ensuring cultural sensitivity, implementing real-time detection systems, and establishing standardized evaluation metrics. These efforts aim to develop more effective and culturally sensitive mechanisms for detecting hate speech in Telugu codemixed text, fostering safer and more inclusive digital spaces.

Ethics Statement

We affirm our commitment to ethical research practices and compliance with ACL guidelines in conducting and presenting our study. No ethical concerns or conflicts of interest arose during the course of this research.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20242138, 20241567, and 20242080 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Telugu Language - Wikipedia. Accessed on: February 6, 2024.

Zahra Ahani, Moein Shahiki Tash, Yoel Ledo Mezquita, and Jason Angel. 2024. Utilizing deep learning models for the identification of enhancers and super-enhancers based on genomic and epigenomic features. *arXiv preprint arXiv:2401.07470*.

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.

Girma Yohannis Bade. 2021. Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia. volume 3, pages 26–30.

Girma Yohannis Bade and Akalu Assefa Afaro. 2018. Object Oriented Software Development for Artificial Intelligence. volume 7, pages 22–24.

Girma Yohannis Bade and Hussien Seid. 2018. Development of Longest-Match Based Stemmer for Texts of Wolaita Language. *vol*, 4:79–83.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, RL Hariharan, John Philip McCrae, Elizabeth Sherly, et al. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 133–145.

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*.

Suman Dowlagar and Radhika Mamidi. 2021. Hasocone@ fire-hasoc2020: Using BERT and multilingual BERT models for hate speech detection. *arXiv preprint arXiv:2101.09007*.

Selam Kanta and Grigori Sidorov. 2023. Selam@ DravidianLangTech: Sentiment Analysis of Code-Mixed

Dravidian Texts using SVM Classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

Namit Khanduja, Nishant Kumar, and Arun Chauhan. Unmasking Hate: Telugu Language Hate Speech Detection Using Transformers. *Available at SSRN 4642780*.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am I a Resource-Poor Language? Data Sets, Embeddings, Models and Analysis for four different NLP Tasks in Telugu Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Bharathi Raja andS Malliga andCN SUBALALITHA Priyadharshini, Ruba andChakravarthi, Premjith and-Murugappan Abirami S V, Kogilavani andB, and Prasanna Kumar Kumaresan. 2023a. "Overview of Shared-task on Abusive Comment Detection in Tamil and Telugu". In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, S Malliga, Subalalitha Cn, SV Kogilavani, B Premjith, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87.

Ícaro JS Ribeiro, Rafael Pereira, Ivna V Freire, Bruno G de Oliveira, Cezar A Casotti, and Eduardo N Boery. 2018. Stress and quality of life among university students: A systematic literature review. *Health Professions Education*, 4(2):70–77.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word Level Language Identification in Code-mixed Kannada-English Texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. LIDOMA@ DravidianLangTech: Convolutional Neural Networks for Studying Correlation Between Lexical Features and Sentiment Polarity in Tamil and Tulu Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multilingual Hope Speech Detection using Machine Learning.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word Level Language Identification in Code-mixed Kannada-English Texts using Deep Learning Approach. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 29–33.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*.