Detecting Hate Speech in Hinglish: A BiLSTM Neural Network Approach

Kirti Kumari[1,*,†], Avishikta Bhattacharjee[2,†] and Vinayak Vijay[1,†]

[1]Indian Institute of Information Technology Ranchi
[2]Kalinga Institute of Industrial Technology

Abstract

In today's digital age, the rise of social media platforms has brought forth a concerning increase in hate speech among users. This research paper delves into a thorough examination of hate speech detection for Hinglish, shedding light on the unique challenges posed by code-mixing, transliteration, and rich morphological variations in the language. The study delves into the utilization of a deep neural network comprising LSTM layers. Furthermore, the approach encompasses tailored preprocessing pipelines designed to effectively manage code-mixed data, transliteration obstacles, and emoticon interpretation. The models underwent training and evaluation on meticulously curated datasets, demonstrating their efficacy in accurately pinpointing instances of hate speech. Our method uses a BiLSTM model to effectively categorize code-mixed Hinglish text, resulting in a macro F1 score of 84.

Keywords

Hate Speech Detection, Hinglish, Deep neural network, LSTM, BiLSTM

1. Introduction

As social media platforms ,like Instagram,Twitter keep evolving,there is a witness of rapid growth in digital communication that come along with a range of intricacies of languages .One of the prominent outcome of this transformation is the widespread adoption of code-mixed languages, particularly Hinglish-a blend of Hindi and English-across North India and South Asia.In this way users helps blending local dialects and expressions with English,that give rise to a diverse and vibrant linguistic landscape that mirrors cultural identities and societal dynamics.However,this linguistic phenomenon not only portrays a region's vibrant cultural identity but at the same time introduces significant complexities for automated systems tasked with content moderation, sentiment analysis, and hate speech detection. These complex interactions between code-mixing, transliteration, and the numerous morphological variants present in the language, poses special difficulties in detecting hate speech in Hinglish. Traditional natural language processing (NLP) algorithms are less effective since users frequently use emoticons or non-standard spellings, employ region-specific slang, and move between Hindi and English effortlessly within a single statement. Furthermore, creating reliable hate speech detection models is made more difficult by the dearth of sizable, excellent annotated datasets for code-mixed languages.

This research paper seeks to investigate the intersection of linguistic diversity within the Hinglish language and its implications for identifying harmful content on online platforms.After following and examining case studies involving the processing of the codemixed digital information, we aim to showcase the efficacy of existing algorithms in navigating this intricate linguistic terrain. Our ultimate objective is to underscore the significance of integrating cultural nuances into algorithmic frameworks to cultivate a more inclusive and culturally aware digital environment. In this research, we utilized the

Forum for Information Retrieval Evaluation, December 12-15, 2024, India
*Corresponding author.
†All the authors contributed equally.
kirti@iiitranchi.ac.in (K. Kumari); avishikta.bhattacharjee@gmail.com (A. Bhattacharjee); vinayakvijay2003@gmail.com (V. Vijay)
https://github.com/gatetub (A. Bhattacharjee); https://github.com/Vinayak164000 (V. Vijay)

BiLSTM neural network model to effectively classify code-mixed Hinglish text, achieving a macro F1 score of 84.

The remaining sections of this work are organized as follows. Section 2 offers a quick introduction to relevant literature. Section 3 outlines the proposed approach and structure for dealing with Hinglish text. Section 4 summarizes the experiments and findings. The paper concludes in Section 5.

2. Related Work

Hinglish, a code-mixed language blending Hindi and English, is widely spoken in North India, presenting significant challenges in detecting hate speech due to its inherent complexities and nuanced expressions. Varade et al. [1]highlight the difficulties of working with such data, emphasizing the need for robust machine learning models to effectively analyze Hinglish texts. They underscore the importance of preprocessing techniques to handle the intricacies of this mixed language. Similarly, Ananya et al. [2] explore the role of artificial intelligence, particularly deep learning approaches, in identifying hate speech within Hinglish content. Their research stresses the significance of meticulously curating datasets and applying thorough preprocessing strategies to improve classification accuracy.

In line with this, Rahul et al. [3] present an ensemble-based methodology for detecting hate speech in Hinglish by integrating multiple machine learning models to enhance overall classification performance. Their approach showcases the advantages of combining different models to gain a more comprehensive understanding of the language's structure. Bhaskara et al. [4] take a unique angle by incorporating emojis as an essential feature in their hate speech detection models, offering a comparative analysis between English and Hinglish datasets. Their study highlights the value of emojis in understanding the emotional tone and abusive content in Hinglish texts. Kumari et al. [5]proposed a deep learning approach based on pre-trained BERT models to identify hate speech and offensive language in code-mixed Hindi-English social media text. Their work demonstrated the effectiveness of fine-tuning BERT models for this task.

Moreover, Kumar et al. [6] introduce HSDH, a deep neural network architecture specifically designed to detect hate speech in Hinglish. Their research explores various deep learning architectures to better capture the complexity of code-mixed text, demonstrating how these approaches can be tailored for Hinglish. Shankar et al. [7] also address this challenge by proposing a transformer- and translation-based approach to combat hate speech from the perspective of bilingual Hinglish speakers, which is particularly valuable given the mixed linguistic nature of the data. Their 2022 study emphasizes the role of translation in improving the identification process. Similarly, Birdar et al. [8] (2021) explore a translation-based method for hate speech detection in code-mixed Hinglish datasets, contributing further to the understanding of how translation techniques can support accurate hate speech classification. Meanwhile, Barman et al. [9] focus on the broader challenge of identifying languages within code-mixed texts like Hinglish, which is prevalent across social media platforms. Their work emphasizes the need for specialized models to effectively handle these unique linguistic phenomena, contributing to a deeper understanding of how mixed languages function in digital discourse.

3. Methodology

This section consists of an overview of data description, data preprocessing, feature extraction techniques, and the methods used to train models.

3.1. Dataset Description and Data Preprocessing:

For this task, we merged the HASOC 2021 [10] dataset with an open-source dataset from Kaggle [1] to conduct a comprehensive analysis of hate speech patterns. This integration provided us with a broader scope, enabling the exploration of diverse linguistic elements and social contexts, thus offering a deeper

[1]https://www.kaggle.com/datasets/bajpaipurva/hinglish-code-mixed-dataset

understanding of hate speech in various forms. However, one of the primary challenges we faced was the unclean nature of the initial dataset, which required extensive preprocessing. The dataset was structured in JSON format, with multiple comments nested within individual tweets, adding layers of complexity. Extracting each comment and removing sensitive information, such as individuals' names and tagged references, was essential to reducing bias. This step allowed us to focus solely on the language used in the dataset, ensuring a more objective analysis of hate speech. The detail description of dataset are presented in Table 1 and supplementary dataset, which we used for training the models are presented in 2. More explanation about tasks and dataset can be seen in the articles [11] [12].

To further refine the dataset, we employed meticulous preprocessing techniques, shown in Figure 2. These included converting all text to lowercase, removing hyperlinks, emojis and punctuation marks, and eliminating excessive white spaces. By cleaning and preparing the dataset in this manner, we could delve deeper into the nuances of language use in hate speech detection.

Table 1
Labels Distribution of Comments

Comments	Train Dataset	Validation Dataset
Hate/Offensive	7434	1866
Not Offensive	4394	1092

Table 2
Datasets Distribution

Datasets	Hate/Offensive	Not Offensive
Hasoc 2021 Dataset	2702	2730
Kaggle Dataset	2784	6570

Figure 1: Proposed Data Preprocessing Pipeline

3.2. Feature Engineering

Once the textual data was preprocessed—through operations such as lowercasing, removal of non-alphanumeric characters, and elimination of stop words—it was subsequently transformed into numerical representations suitable for model training. For deep learning-based approaches, we employed the TensorFlow framework to perform tokenization, wherein the text was segmented into individual words or subword units. To standardize input lengths across samples, we applied padding and truncation techniques, ensuring that all sequences conformed to a uniform size—a prerequisite for efficient model computation. Additionally, we utilized padding and truncating sequences to ensure uniform input sizes, which is vital for the model to process the data effectively. Furthermore, we delved into creating embeddings, which are essentially numerical representations of words, designed to encapsulate the semantic meaning of the words. In terms of methodology, while using machine learning classifiers, we opted for the TF-IDF vectorization method instead of using TensorFlow embeddings to convert text into numerical features. Hindi, being a morphologically rich language with complex linguistic structures,

can pose challenges for embedding techniques, especially when dealing with rare or out-of-vocabulary words.

The key observation was found to be sentences containing English abusive words are more lightly to be classified as HOF in code-mix rather than code-mix sentences with Hindi hate comments. Hence, TF-IDF that is well-suited to highlight frequently occurring terms, solves the problem to the limited semantic analysis of non-English hateful comments.It checks the frequency of hateful comments irrespective of its language.To further improve model performance, we stacked both character-level and word-level TF-IDF vectors, creating a comprehensive feature set [13]. These embeddings played a pivotal role in enabling the model to grasp the intricate nuances and contextual meaning of the text, thereby enhancing its ability to comprehend context and sentiment in a more nuanced and effective manner.

3.3. Model Construction

For the classification of hate speech in the Hinglish language, we explored ML algorithms as well as the Neural network approach. The neural network model architecture integrates a BiLSTM (Bidirectional Long-Short-Term Memory) network for classification tasks, optimized to capture both forward and backward contextual information from sequences. Starting with an embedding layer, the model maps tokens to dense vectors, effectively transforming the input text into meaningful numerical representations. This is followed by the BiLSTM layer, which processes the sequence bidirectionally to harness the context from both preceding and succeeding words, thereby enriching the model's understanding of sequential dependencies. The GlobalMaxPool1D layer is then applied to extract the most important features from the entire sequence, emphasizing the most critical information for classification. The architecture is further enhanced by dense layers, fully connected, which perform the bulk of the classification decision-making. To mitigate overfitting and improve generalization to unseen data, a dropout layer is incorporated, ensuring that the model does not simply memorize the training data. The final output layer features a single neuron with a sigmoid activation function, perfectly suited for binary classification tasks, such as distinguishing between categories like Highly Offensive (HOF) and Non-Offensive (NOT). This architecture leverages robust sequence modeling while remaining computationally efficient and effective for binary classification problems. After training the model on our labeled dataset, we evaluated its performance on the validation dataset using the macro F1 score. The stepwise model construction is shown in Figure 2.

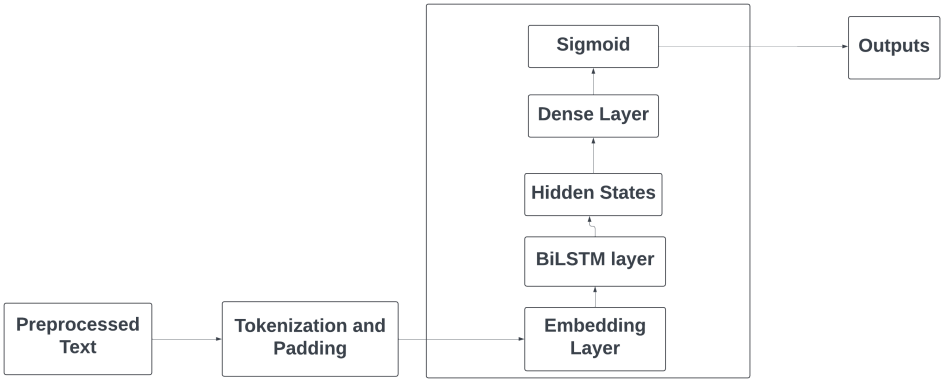

Figure 2: Model Construction Pipeline

4. Experiments and Results

For the Hasoc 2024 shared task, we splitted our dataset into training and validation sets, with 80% of the data used for training and 20% for validation, while the test dataset was provided by the HASOC

organizers. The primary task was to categorize text into two categories: *Hate Speech* or *Not-Hate Speech*, specifically for the code-mixed Hinglish dataset.

One of the main challenges we faced was the identification of nuanced expressions and slang that are commonly used in code-mixed languages.For example: “ab kaun paisa dega bhakwas karne ka” as NOF “rahul ka gulam chamcha gyan de rha hai ” as NOF “paid agent for anti india activities ban twitter in india lets all move to indian app koo” as HOF.

In the following, the labeled data classification of Hindi-English sentences has been identified as Hate/Offensive (HOF) whenever they consist of English slang. However, pure Hindi slang has been misclassified as NOF. These expressions are often complex due to the varying dialects of the Hindi language when mixed with English. The organizers evaluated and ranked the models' performance using macro F1-scores, which ensured balanced performance across both classes.

The results of these classifiers, evaluated in the validation set, were computed using the `Scikit-learn` library, and were presented in Table 3 for comparison.

Table 3
Results of different classification models

Classification Model	F1 Score	Recall	Precision
Logistic Regression	80	83	84
MultinomialNB	80	81	80
DecisionTreeClassifier	80	81	80
RandomForest	82	81	82
ExtraTreeClassifier	83	83	79
AdaBoostClassifier	80	81	82
BiLSTM Neural	85	83	84

5. Conclusion

In this paper, we present the comprehensive strategy developed by KK_IIIT_Research_Lab for the HASOC 2024 shared task. Our approach leverages a BiLSTM model to effectively classify code-mixed Hinglish text, achieving a macro F1 score of 84. While the model demonstrated strong performance, there are opportunities for further improvement.

To enhance the model's effectiveness and generalizability for real-world applications, future research could focus on expanding the dataset to include more diverse examples of code-mixed language, employing techniques to mitigate class imbalance, and exploring advanced model interpretability methods. These improvements will provide deeper insights into the model's decision-making process and help create a more robust system for hate speech detection.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. S. Varade, V. B. Pathak, Machine Learning and Information Processing, Springer Singapore, 2020, pp. 265–276. doi:10.1007/978-981-15-1884-3_25.

[2] A. Srivastava, M. Hasan, B. Yagnik1, R. Walambe, K. Kotecha, Lecture Notes in Electrical Engineering, Springer Singapore, 2021, pp. 83–95. doi:10.1007/978-981-16-3067-5_8.

[3] Rahul, V. Gupta, V. Sehra, Y. R. Vardhan, Ensemble based hinglish hate speech detection, in: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021. doi:10.1109/iciccs51141.2021.9432352.

[4] S. Bhaskara, S. P, S. Seth, S. Mohanty, P. Kanwal, Detection and comparison of abusive and hate speech in english and hinglish with emojis using deep learning and non-deep learning techniques, in: 2023 4th International Conference for Emerging Technology (INCET), 2023. doi:10.1109/incet57972.2023.10170633.

[5] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ hasoc 2020: Bert models for hate speech identification in indo-european languages., in: FIRE (Working Notes), 2020, pp. 319–324.

[6] R. K. Kaliyar, A. Goswami, U. Sharma, K. Kanojia, M. Agrawal, Hsdh: Detection of hate speech on social media with an effective deep neural network for code-mixed hinglish data, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023. doi:10.1109/ICCCNT56998.2023.10306709.

[7] S. Biradar, S. Saumya1, A. chauhan, Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach., Social Network Analysis and Mining (2022). doi:https://doi.org/10.1007/s13278-022-00920-w.

[8] S. Biradar, S. Saumya, A. Chauhan, Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set, in: 2021 IEEE International Conference on Big Data (Big Data), 2021. doi:10.1109/BigData52589.2021.9671526.

[9] U. Barman, A. Das, J. F. Joachim Wagner, Code mixing: A challenge for language identification in the language of social media, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, 2014. doi:10.13140/2.1.3385.6967.

[10] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hatespeech and offensive content identification in english and indo-aryan languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021, volume 3159 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1–19. URL: https://ceur-ws.org/Vol-3159/T1-1.pdf.

[11] K. Ghosh, N. Raihan, S. Modha, S. Satapara, T. Gaur, Y. Dave, M. Zampieri, S. Jaki, T. Mandl, Overview of the HASOC Track at FIRE 2024: Hate-Speech Identification in English and Bengali, in: FIRE '24: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation. December 9-13, Gandhinagar, India, Association for Computing Machinery (ACM), New York, NY, USA, 2024.

[12] N. Raihan, K. Ghosh, S. Modha, S. Satapara, T. Gaur, Y. Dave, M. Zampieri, S. Jaki, T. Mandl, Overview of the HASOC Track at FIRE 2024: Hate-Speech Identification in English and Bengali, in: K. Ghosh, T. Mandl, P. Majumder, D. Ganguly (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2024) December 9-13, Gandhinagar, India, CEUR-WS.org, 2024.

[13] K. Kumari, J. P. Singh, Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content, FIRE (working notes) 2517 (2019) 328–335.