A Multilingual BERT-Based Framework for Robust Online Hate Speech Detection

Yash Shukla
Information Technology
VPKBIE&T, Baramati
Baramati, India
studiyash@gmail.com

Mr. Rajkumar Vamanrao Panchal
Information Technology
VPKBIE&T, Baramati
Baramati, India
rajkumar.panchal@vpkbiet.org

Tanmay Nigade
Information Technology
VPKBIE&T, Baramati
Baramati, India
tanmaynigade916@gmail.com

Suyash Khodade
Information Technology
VPKBIE&T, Baramati
Baramati, India
suyashkhodade3331@gmail.com

Prathamesh Pimpalkar
Information Technology
VPKBIE&T, Baramati
Baramati, India
prathameshpimpalkar1@gmail.com

2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI) | 979-8-3503-7687-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVMI61877.2024.10782783

***Abstract*—Hate speech poses a significant threat to societal harmony and safety, often inciting violence and discrimination. Recent real-time incidents, such as the communal unrest in various parts of the world triggered by inflammatory online content, demonstrates the need for effective hate speech detection mechanisms. In this paper, we present a comprehensive hate speech detection model utilizing the Bidirectional Encoder Representations from Transformers (BERT) algorithm, leveraging advanced Deep Learning and Natural Language Processing (NLP) techniques. Our model is designed to address the challenges of detecting hate speech within the Hinglish language, a code-mixed variant combining Hindi and English. We used a pre-trained BERT model and tailored that model to function in real-time scenarios, offering robust and in-depth analysis capabilities. We adopted a multilingual and multimodal approach, enabling our system to detect hate speech. This task is accomplished across various content formats, including text, audio, video, images, GIFs, and YouTube video comments. Experimental results demonstrate that our model achieves an accuracy of 0.88. This underlines the model's efficacy in detecting hate speech, showcasing its potential for application in diverse real-world contexts.**

***Index Terms*—Hate Speech Detection, BERT, Deep Learning, Natural Language Processing, Multilingual, Multimodal, Hinglish, Real-Time Analysis.**

I. Introduction

In the digital age, the proliferation of social media and online communication has ushered to an enlarge use of multilingual language, particularly code-mixed languages such as Hinglish a blend of Hindi and English. It is estimated that approximately 60% of the global population engages in multilingual communication online. This trend is especially prominent in India, where Hinglish has become a ubiquitous mode of expression. However, this blending of languages presents unique challenges for automated systems designed to identify and alleviate hate speech.

Hate speech refers to verbal or written expressions that attack or demean individuals or groups based on attributes such as race, religion, ethnicity, gender, sexual orientation, etc. It often promotes hatred, discrimination, or violence against these targeted individuals or groups [1]. Several news articles and reports provide insights into the impact of hate speech on social media. In India, this issue is underscored by several incidents. For example, during the Delhi riots of 2020, incendiary posts on social media exacerbated communal tensions. Similarly, hate speech on platforms like Twitter and Facebook has been linked to the spread of false information during the COVID- 19 pandemic, leading to violent attacks on individuals from marginalized communities. Globally, events such as the Christchurch Mosque shootings in New Zealand were preceded by the perpetrator's hateful manifesto circulated online, illustrating the real-world consequences of unchecked hate speech. During the Rohingya crisis in Myanmar, social media platforms were leveraged to incite violence against minority communities. These instances highlight an urgent need for effective hate speech detection mechanisms.

The societal impact of hate speech is profound. It fosters an environment of fear, division, and hostility, undermining social cohesion and contributing to the marginalization of vulnerable groups. The psychological impact on victims, who often experience depression, anxiety, and a diminished sense of safety, cannot be overstated. Moreover, the perpetuation of hate speech can lead to real-world violence, as evidenced by numerous incidents worldwide [2].

To address this critical issue, researchers are increasingly turning to Deep Learning and other advanced machine learning techniques. Deep Learning models, particularly those utilizing neural networks, have shown great promise in understanding and processing complex language patterns inherent in code-mixed languages like Hinglish. Techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models (e.g., BERT, GPT) can be trained on large datasets to accurately identify hate speech, even in nuanced and context- dependent scenarios. By leveraging these technologies, it is possible to develop vigorous

This project is funded by Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India.

systems that can automatically identify and flag harmful content, thereby mitigating its spread and impact.

A. Our Contribution

In this work, we address the challenging problem of hate speech detection across multimodal and multilingual content. Our approach leverages a BERT-based Deep Learning model, enhanced by advanced NLP techniques. Unlike existing approaches that primarily focus on textual data, our model extends to various modalities including images, videos, emoticons, memes, and YouTube comments from live streaming and non-live video sources. The cornerstone of our research lies in the creation and utilization of an extensive codemixed dataset provided by the CONSTRAINT 2021 [3] and HHSD created by Prashant Kapil [4]. where comments are annotated as "Hate" or "Non-Hate". These datasets not only support multiple languages but also incorporates nuanced cultural and linguistic contexts, making it suitable for training our model to detect hate speech effectively in diverse settings.

BERT (Bidirectional Encoder Representations from Transformers) in hate speech detection operates by leveraging bidirectional context to understand the meaning of text comprehensively [2]. It is pre-trained on large text corpora, capturing deep contextual dependencies. Fine-tuned on hate speech datasets, BERT adjusts its parameters to accurately detect hate speech, including from diverse sources like images, videos, and social media content. Its multilingual capabilities and robust contextual understanding makes BERT a powerful tool for detecting hate speech across different languages and modalities.

To achieve robust performance, our model is trained to understand deep contextual meanings inherent in different forms of multimedia content. Specifically, it exhibits the ability to interpret and detect hate speech from non-textual elements such as images and videos, which are increasingly prevalent in online platforms. Furthermore, our approach incorporates a novel technique to decode hate speech from emoticons and memes, thus addressing a significant gap in existing hate speech detection systems. We also introduced optimizations that allow our model to achieve superior performance metrics using minimum computational resources, which makes it suitable for real-time applications and environments with restricted computing capabilities.

Overall, the contributions of this paper can be summarized as follows.

- A novel multimodal and multilingual hate speech detection model based on a Deep Learning approach using BERT and advanced NLP techniques.
- Our model's integration with BERT significantly boosts performance by comprehensively understanding contextual meanings and accurately predicting human intentions.
- Extension of our model's capability to detect hate speech from diverse sources including emoticons, memes, images, videos, and YouTube comments, both live and non-live, with superior performance metrics and efficiency.

Our work not only advances the state-of-the-art in hate speech detection but also provides a scalable solution that can be applied across various multimedia and multilingual contexts, thereby contributing significantly to the field of AI and NLP.

II. Related work

Hate speech detection has garnered significant attention in recent years, particularly with the rise of social media platforms where users frequently engage in the conversations that cross the linguistic boundaries. This has introduced unique challenges in detecting hate speech, especially in code-switched and multilingual contexts. Various studies have approached this problem using advanced machine learning and Deep Learning techniques, aiming to improve the accuracy and robustness of hate speech classifiers. This section reviews the notable contributions in this domain, focusing on methodologies and results that address the intricacies of hate speech detection in mixed-language data.

A. Machine Learning Based Approaches

Rahul et al. [5] in their paper, present a self-sufficient model for detecting hate speech in Hinglish texts using ensemble methods. They developed two ensembles: one is combining machine learning models (Logistic Regression, Random Forest, SVM, and Multinomial Naive Bayes) with a Voting Classifier, and another stacking Deep Learning model (CNN and Bi-LSTM) using FastText embeddings. Future work involves exploring character-level embeddings, implementing transformer models, and expanding the Hinglish dataset to enhance accuracy and robustness.

Hajime Watanabe et al. [6] provided a technique for detecting hate speech on Twitter by combining unigrams, patterns, and sentiment-based features. The approach involves manually annotating tweets into "Clean," "Offensive," and "Hateful" categories, extracting relevant features, and using these features to train a machine learning model. Future work includes enriching the hate speech pattern dictionary and studying the prevalence of hate speech.

B. Deep Learning Approaches

Sayani Ghosal et al. [7] in their paper, propose a hybrid approach for detecting Bengali hate speech. This approach combines ExtendedFuzzy SVM and mBERT text embeddings to handle imbalanced datasets and a Morphological Analysis with Hate Similarity (HS) Scheme for detecting implicit and explicit hate speech using lexical and semantic analyses. The algorithm preprocesses text, generates embeddings with mBERT, classifies text using an extended fuzzy SVM, and applies HS analysis with Word2Vec and a Bengali hate lexicon. Future work involves expanding the hate lexicon, detecting diverse emotional expressions, extending research to other languages, and improving model efficiency for handling misspelled words.

Shakir Khan et al. [8] presented a novel model combining Convolutional Neural Networks (CNN), Bi-Directional Gated

TABLE I
SUMMARY OF THE EXISTING RESEARCH

Approach	Paper Title	Authors	Methodology	Accuracy	F1-score	Findings
Machine Learning	Ensemble Based Hinglish Hate Speech Detection [5]	Rahul et al.	ensemble methods	0.873	0.873	Data scarcity and variations in codemixing syntax
	Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection [6]	Hajime Watanabe et al.	J48graft	78.4%	-	Distinguishing between "Offensive" and "Hateful" content not accurate and the limitations of training data size and scope
Deep Learning	Inculcating Context for Emoji Powered Bengali Hate Speech Detection using Extended Fuzzy SVM and Text Embedding Models [7]	Sayani Ghosal et al.	ExtendedFuzzy SVM and mBERT	85.00	0.88	Reliance on a comprehensive hate lexicon, computational time for long tweets, and challenges with dynamic slang and misspellings
	HCovBi-Caps: Hate Speech Detection using Convolutional and Bi-Directional Gated Recurrent Unit with Capsule Network [8]	Shakir Khan et al.	Combining CNN, Bi-GRU, and Capsule Networks	0.92	0.90	Lack of sentiment analysis, user profile features, and testing on multimodal and multilingual datasets
	Hate Speech Recognition in Multilingual Text: Hinglish Documents [9]	Arun Kumar Yadav et al.	CNN-BiLSTM model with word2vec embedding	0.876	0.835	Challenges with the small dataset size and the complexity of code-switching in Hinglish
	Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings [10]	Rahul et al.	Character level embedding with GRU and LSTM networks	0.86		Handling transliteration variations and the need for a larger corpus
	Deep Learning Based Fusion Approach for Hate Speech Detection [11]	Yanling Zhou et al.	Fusion-combining the outputs of ELMo, BERT, and CNN classifiers	0.746	0.712	Shallow integration among classifiers and high computational costs
Transformer - Based	Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model [12]	Hind Saleh et al.	Word Embedding with BiLSTM and a BERT-based Model	96%		It requires significant computational resources and extensive training data
	Hate Speech and Offensive Language Detection using an Emotion-aware Shared Encoder [13]	Khouloud Mnassri et al.	BERT-based ensemble models	0.9474	0.9288	Data imbalance and high computational costs
	Assessing the Impact of Contextual Information in Hate Speech Detection [14]	Juan Manuel P´erez et al	BETO	61.3%		Inconsistent context availability, and dataset constraints to media tweet replies

Recurrent Units (BiGRU), and Capsule Networks to enhance hate speech detection. The HCovBi-Caps model utilizes 1D convolutional layers to extract spatial features, BiGRU layers to capture contextual information, and Capsule Networks to preserve spatial hierarchies, leading to superior performance metrics on balanced (DS1) and unbalanced (DS2) datasets. Future work try to remove these limitations by incorporating sentiment and user profile analysis, handling multimodal data, and evaluating the model on diverse datasets.

Arun Kumar Yadav et al. [9] evaluated different ML and DL techniques to detect hate speech in Hindi English mixed text. They created a dataset of 20,600 instances by mixing three datasets and found that Deep Learning models, CNN-BiLSTM model with word2vec embedding, outperformed classical machine learning models. This paper combines CNNs and BiLSTMs to capture spatial features and temporal dependencies respectively. Future research directions include expanding datasets, enhancing models with advanced architectures and transfer learning, implementing real-time detection systems, and extending studies to other code-mixed languages.

Rahul et al. [10] in their paper, tackled the challenges of detecting hate speech in Hinglish tweets using Deep Learning models. They utilized character-level embeddings with GRU and LSTM networks incorporating attention mechanisms to manage vocabulary distortion and spelling variations in code-mixed language. Future work involves expanding the dataset, combining word and character embeddings, and exploring transformer models to enhance performance.

Yanling Zhou et al. [11] proposed a fusion based method to enhance hate speech detection by combining the outputs of ELMo, BERT, and CNN classifiers using algebraic rules like mean, max, and product. Future work suggests integrating ELMo and BERT embeddings within the CNN framework and exploring early cooperation strategies between classifiers for better performance.

C. Transformer-Based Approaches

Hind Saleh et al. [12] in their paper, investigated two primary approaches for hate speech detection: Domain-Specific Word Embedding with BiLSTM and a BERT-based Model. The first approach involves creating word embeddings from a hate speech corpus and using these embeddings in a BiLSTM

model for classification. The second approach fine-tunes a pre trained BERT model on hate speech data, utilizing the final hidden state of the [CLS] token for classification, enhanced with a softmax layer. Future work includes exploring larger datasets, other Deep Learning architectures, transfer learning models, and real-time detection in diverse languages and cultural contexts.

Khouloud Mnassri et al. [13] in their paper, used a multi-task learning (MTL) approach with BERT and mBERT models to enhance hate speech and offensive language detection by combining emotional features. This method trains models for both hate speech and emotion classification simultaneously, leveraging shared representations for improved efficiency. Future work suggests data augmentation, feature integration, cross-lingual generalization, and resource optimization.

Juan Manuel Pérez et al. [14] explored the effect of contextual information on hate speech detection using a dataset which uses user responses to posts from social media Twitter, focusing on the COVID-19 pandemic in Rioplatense Spanish dialect. They employed BETO, a Spanish BERT variant, to conduct binary and multi-label classification experiments, assessing the isolated comment, the comment with the tweet it responds to, and the comment with both the tweet and related news article. Results showed that contextual information significantly enhanced performance. Future work suggested incorporating real-world knowledge, addressing detection challenges in complex language, and extending methods to other conversational contexts and media types.

III. Proposed Approach

A. Input

Our model is designed to handle a diverse range of input types to detect hate speech effectively across multiple media formats:

1) **Text:** Plain text comments, reviews, posts, and messages from various social media platforms. This includes text written in Hinglish, a mix of Hindi and English commonly used in social media in India.
2) **Emoticons:** Graphical icons representing emotions and expressions, which can convey sentiment and be indicative of hate speech or offensive content.
3) **Images and GIFs:** Visual content is analyzed to detect hate symbols, offensive images, and other visual cues that may indicate hate speech. We employed Optical Character Recognition (OCR) techniques provided by the Google Vision API for character recognition in Image and GIF.
4) **Audio and Video:** Speech recognition is employed to transcribe spoken words from audio and video inputs, which are then analyzed for hate speech. We used OCR technique provided by Video Intelligence API for character recognition in video.
5) **YouTube Comments:** Both live stream and non-live video comments are collected and analyzed.

These various forms of input ensure a comprehensive approach to hate speech detection across multiple media types, providing a robust and versatile solution.

B. Text Preprocessing

Text preprocessing is a important step in preparing the data for further analysis and training. The steps involved are:

- **Data Loading:** The dataset is loaded from a local file, typically in CSV format, containing comments and their corresponding labels. The file is read into a pandas DataFrame for ease of manipulation.
- **Data Splitting:** The dataset is break into training and testing sets, This ensures that the model is evaluated on unseen data. Typically, an 70-30 split is used.
- **Label Encoding:** The labels in the dataset are converted into numerical values using LabelEncoder from scikit-learn.
- **Handling Missing Values:** Any missing or null values in the dataset are addressed by either filling them with appropriate placeholders or removing rows with missing data.
- **Text Normalization:** The text data is cleaned and normalized by converting it to lowercase.

C. Text Tokenization with BERT

The BERT tokenizer is used to preprocess the text data before feeding it into the BERT model:

- **Tokenization:** The text is tokenized into subwords using the BERT tokenizer.
- **Special Tokens:** Special tokens [CLS] and [SEP] are added to the tokenized text.
- **Padding and Truncation:** The tokenized sequences are padded or truncated to a fixed length of 128 tokens.
- **Attention Masks:** Attention masks are generated to indicate which tokens are actual input tokens and which are padding tokens.

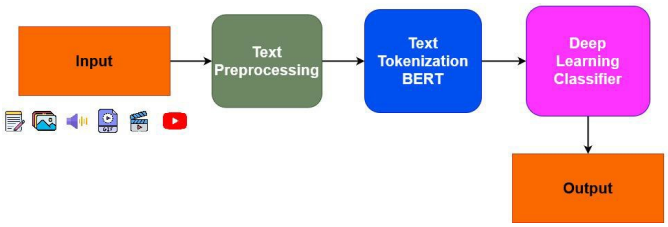

Fig. 1. System Architecture.

D. Deep Learning Classifier: BERT Model

We utilized a pre-trained BERT model, specifically **bert-base-multilingual-uncased**, which is fine-tuned for sequence classification. The BERT model consists of several layers:

- **Embedding Layer:** Converting input tokens into dense vectors.
- **Encoder Layers:** The model has 12 transformer blocks, each comprising:

1) **Multi-head Self-Attention Mechanism:** Model to concentrate on different parts of the input simultaneously.
2) **Feed-Forward Neural Network:** Adds non-linearity to the model.
3) **Layer Normalization and Residual Connections:** Enhance training stability and performance.

- **Output Layer:** The final representation of the [CLS] token is fed into a dense layer to produce logits for each class (hate or non-hate).

The BERT model is fine-tuned with the Adam optimizer and the Sparse Categorical Cross entropy loss function.

E. Output

The output from the classifier is represented in a structured format:

1) **Predicted Label:** The class with the highest logit is selected as the predicted label.
2) **Evaluation Metrics:** Comprehensive performance evaluation metrics such as accuracy, precision, recall, and F1-score are computed to assess the model's effectiveness.

In conclusion, our proposed approach effectively integrates advanced NLP and DL techniques to detect hate speech across different media types, ensuring robust and reliable performance in identifying harmful content in Hinglish comments. This detailed pipeline, utilizing the pre-trained BERT model, demonstrates significant potential in addressing the challenges of multilingual and code-mixed language processing in real-world applications.

IV. Technical detail

Here, we will delve into the technical aspects with a focus on key components such as tokenization, the attention mechanism in BERT, the optimizer used for training, and the loss function for model evaluation.

1) **Tokenization**
Tokenization in BERT involves converting text into tokens, which are then converted into token IDs. For instance, let's take the Hinglish comment "Tum bahut bure ho." The BERT tokenizer would process this text as follows:
   - **Text to Tokens:**
     - "Tum" → "tum"
     - "bahut" → "bahut"
     - "bure" → "bure"
     - "ho" → "ho"
   - **Add Special Tokens:**
     [CLS], "tum", "bahut", "bure", "ho",**[SEP]**
   - **Convert Tokens to IDs:**
     Assume the tokenizer maps these tokens to the following IDs:
     [CLS] → 101, "tum" → 456, "bahut" → 789, "bure"→123, "ho"→ 456, **[SEP]** → 102
     This results in the input sequence: **[101, 456, 789, 123, 456, 102]**.
2) **BERT's Self-Attention Mechanism**
The self-attention mechanism in BERT is crucial for understanding the relationships between different tokens in a sequence. The attention mechanism calculates attention scores using the following steps:
   - **Query, Key, and Value Matrices:** For each token, BERT generates query (Q), key (K), and value (V) vectors through linear transformations. These vectors capture different aspects of the token's relationship with other tokens.
   - **Attention Scores Calculation:**
     The Adam optimizer is used to updates the model weights in training, and sparse categorical crossentropy is employed to calculate the loss.

Example Execution: Hinglish Comment
Consider the Hinglish comment "Tum bahut bura ho":

1) **Tokenization:**
   - Tokens: **[CLS]**, "tum", "bahut", "bure", "ho", **[SEP]**
   - Token IDs: [101, 456, 789, 123, 456, 102]
2) **Self-Attention:**
Attention scores are calculated. Final embeddings are produced by weighted summation of vectors based on attention scores.
3) **Model Forward Pass:**
The token embeddings are processed through the BERT layers. The output logits for the [CLS] token are obtained.
4) **Prediction:**
The class with the highest logit is selected as the predicted label.
5) **Loss Calculation:**
If the true label is "hate speech", the loss is computed using the predicted probability for the true class.
6) **Output**
The model outputs logits. The predicted class is the one with the highest logit. Evaluation metrics assess the model's performance. For instance, if the model correctly predicts 80 out of 100 hate speech comments, the accuracy would be 80%. Detailed classification report provide insight knowledge about how good the model performs across different classes, ensuring a comprehensive evaluation of its effectiveness.

V. Experimental setup and result

1) **Datasets**
We used the dataset from CONSTRAINT 2021 [3] and HHSD, created by Prashant Kapil [4]. The combined dataset comprises a total of 22,977 comments. The label distribution is shown in Table II. In this table, label "0" represents hate, and "1" represents non-hate. The CONSTRAINT 2021 dataset includes labels such as non-hostile, fake, defamation, and offensive, all of which were converted into hate and non-hate classes.

TABLE II
LABEL DISTRIBUTION

Dataset	Label	Messages
CONSTRAINT 2021 [3]	0	2393
	1	5800
HHSD [4]	0	7312
	1	7472

2) **Experimental Setting**
For training the hate speech detection model, we utilized Google Colab with a T4 GPU, which offers high processing speed. Additionally, the model was tested on a Ubuntu system 16GB of RAM and 16GB GPU. TensorFlow library was employed to implement the hate speech detection model.
3) **Hyperparameter Setting**
In our experimental setup, the following hyper parameters were used:
- **Learning Rate**: 3e-5
- **Epochs**: 16
- **Batch Size**: 64
- **Maximum Sequence Length**: 128
- **Optimizer**: Adam
- **LossFunction**: SparseCategorical Crossentropy (from_logits=True)
- **EvaluationMetric**: SparseCategoricalAccuracy

The BERT tokenizer was configured with do_lower_case=True, and the BERT model was initialized with num_labels set to the number of unique labels in the dataset.

The performance of our model on the test set is summarized in the table below:

TABLE III
EVALUATION MATRICES

Class	Precision	Recall	F1-score	Support
no	0.85	0.88	0.86	2922
yes	0.91	0.88	0.90	3970

The overall accuracy of our model is 0.88, with a macro average of 0.59 and a weighted average of 0.88.

VI. EVALUATION RESULT & COMPARATIVE ANALYSIS

The performance of our hate speech detection Using BERT by implemented using Tensorflow model was thoroughly evaluated using a comprehensive dataset. Our model demonstrated a balanced performance, with a precision of 0.85 and a recall of 0.88 for the 'Yes' class, and a precision of 0.91 and a recall of 0.88 for the 'No' class. These results underscore the model's capability to effectively identify hate speech, making it a robust technique for monitoring social media.

Our study explore the complex problem of hate speech detection across multimodal and multilingual content, utilizing a BERT-based model augmented with advanced NLP techniques. Unlike traditional approaches that predominantly focus on text, our model extends its detection capabilities to various modalities including images, videos, emoticons, memes, and YouTube comments from both live and non-live sources.

TABLE IV
RESULT COMPARISON

Model	Accuracy
Logistic Regression (TF-IDF Vectorizer) [5]	0.731
Support Vector Machine (TF-IDF Vectorizer) [5]	0.741
Machine Learning Ensemble (TF-IDF Vectorizer) [5]	0.727
CNN with 100D embeddings [5]	0.827
BERT without Tensorflow	0.723
BERT with Tensorflow (proposed method)	**0.88**

BERT (Bidirectional Encoder Representations from Transformers) excels in understanding the bidirectional context of text, which enhances its performance for detecting hate speech effectively [9]. Fine-tuned on a hate speech dataset, BERT adapts its parameters to accurately identify hateful content across different media formats. This deep contextual understanding, combined with BERT's multilingual capabilities, positions it as a potent tool for identifying hate speech in several languages and modalities.

Our model not only deciphers textual hate speech but also interprets non-textual elements such as images and videos, which are prevalent on online platforms. Additionally, it incorporates a novel technique to decode hate speech embedded in emoticons and memes, filling a significant gap in existing detection systems. Furthermore, our model is optimized to perform efficiently with minimal computational resources, making it suitable for real-time applications and environments with limited computing capabilities.

- In comparison to other models detailed in the literature **Domain-Specific Word Embedding with BiLSTM** [12] and **BERT-based Model** both achieve high F1 scores (BiLSTM: 93%, BERT: 96%). However, these approaches are limited to textual data and require substantial computational resources. Our model, by contrast, is optimized for efficiency and extends beyond text to include multimedia content.
- **Emotion-aware Shared Encoder** [13] utilizes BERT and mBERT models within a multi-task learning framework to enhance detection through emotional features. Despite its high accuracy (BERT MTL: 93.85primarily focuses on text and does not extend to other modalities like images or videos.
- **Contextual Information Integration** [14] explores the effect of contextual information on hate speech detection using the BETO model, a Spanish variant of BERT.

While the inclusion of context improved performance, this method is confined to textual context within a specific language and dialect. Our model's capability to process multiple languages and media types offers a highly versatile solution.

In summary, our contribution marks a major advancement in the field of hate speech detection. By leveraging a multimodal and multilingual approach, our BERT-based model surpasses classical methods in comprehensiveness and efficiency. This innovation not only pushes the boundaries of current hate speech detection technologies but also offers a scalable solution applicable across various platforms and contexts, significantly enhancing the capabilities of AI and NLP in this critical area.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this paper presents multimodal and multilingual hate speech detection model using BERT and advanced NLP techniques. By integrating textual and non-textual data, including images, videos, emoticons, memes, and YouTube comments, our model demonstrates significant improvements in detecting hate speech across diverse contexts. Our findings highlight the model's scalability and applicability in real-time environments, contributing to enhanced detection capabilities and the mitigation of harmful online content. This work represents a significant advancement in hate speech detection by addressing the complexities of multilingual and multimodal inputs.

Future work will concentrating on expanding the dataset to incorporate a wider range of languages and cultural contexts, enhancing the model's generalizability. We also plan to incorporate sentiment analysis and user profile features to improve detection accuracy. Moreover, exploring the integration of transformer-based models with real-time detection systems will be a key area of development to ensure timely and effective mitigation of hate speech.

ACKNOWLEDGMENT

I am thankful to ChatGPT for insights, suggestions and provision of resources. It helped in refining ideas and aiding research throughout this work [15].

REFERENCES

[1] https://library.fiveable.me/key-terms/ap-gov/hate-speech
[2] Bansod, Pranjali Prakash, "Hate Speech Detection in Hindi" (2023). Master's Projects. 1265. DOI: https://doi.org/10.31979/etd.yc74-7qas, https://scholarworks.sjsu.edu/etd_projects/1265
[3] Mohit Bhardwaj and Md Shad Akhtar and Asif Ekbal and Amitava Das and Tanmoy Chakraborty, "Hostility Detection Dataset in Hindi," in arXiv, 2020, eprint-2011.03588.
[4] P. Kapil, G. Kumari, A. Ekbal, S. Pal, A. Chatterjee and B. N. Vinutha, "HHSD: Hindi Hate Speech Detection Leveraging Multi-Task Learning," in IEEE Access, vol. 11, pp. 101460-101473, 2023, doi: 10.1109/ACCESS.2023.3312993.
[5] V. Rahul, V. Gupta, V. Sehra, and Y. R. Vardhan, "Ensemble Based Hinglish Hate Speech Detection," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1800 1806, doi: 10.1109/ICICCS51141.2021.9432352.
[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
[7] S. Ghosal, A. Jain, D. K. Tayal, V. G. Menon, and A. Kumar, "Inculcating Context for Emoji Powered Bengali Hate Speech Detection using Extended Fuzzy SVM and Text Embedding Models," ACM Transactions on Asian and Low- Resource Language Information Processing, accepted March 2023, doi: 10.1145/3589001.
[8] S. Khan et al., "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network," IEEE Access, vol. 10, pp. 7881-7894, 2022, doi: 10.1109/ACCESS.2022.3143799.
[9] A. K. Yadav, A. Kumar, S. ., K. ., M. Kumar, and D. Yadav, "Hate Speech Recognition in multilingual text: Hinglish Documents," TechRxiv, Preprint, 2022, doi: 10.36227/techrxiv.19690177.v1.
[10] V. Rahul, V. Gupta, V. Sehra, and Y. R. Vardhan, "Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1112 1118, doi: 10.1109/ICCMC51019.2021.9418261.
[11] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," IEEE Access, vol. 8, pp. 128923-128929, 2020, doi: 10.1109/ACCESS.2020.3009244.
[12] H. Saleh, A. Alhothali, and K. Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," Applied Artificial Intelligence, vol. 37, no. 1, pp. 2166719, 2023, doi: 10.1080/08839514.2023.2166719.
[13] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Hate Speech and Offensive Language Detection using an Emotion-aware Shared Encoder," arXiv:2302.08777 [cs.CL],2023.[Online].https://doi.org/10.48550/arXiv.2302.0 8777
[14] J. M. Pérez, H. Saleh, A. Alhothali, and K. Moria, "Assessing the Impact of Contextual Information in Hate Speech Detection," IEEE Access, vol. 11, pp. 30575-30590, 2023, doi: 10.1109/ACCESS.2023.3258973.
[15] https://chat.openai.com.

Yash Shukla received his Bachelor of Engineering (B.E.) degree in Information Technology from Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India. His specialization includes Data Science, Machine Learning, Deep Learning and Natural Language Processing. He also has significant experience in Full Stack Web Development and dealing with REST APIs as well as integrating them with various machine learning and creating other various projects.

Mr. Rajkumar V. Panchal is currently an Assistant Professor in the Department of Information Technology at Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India. He received his M.Tech. in Computer Engineering and is currently pursuing his Ph.D from SRTMU, Nanded. His research interests include Natural Language Processing, Deep Learning, and Machine Learning. Mr. Rajkumar has 10 years of teaching experience and teaches a variety of subjects including Computer Graphics, Discrete Mathematics, Object-Oriented Programming, and Software Testing, Data Structure, Natural Language Processing. He is a life member of the Indian Society for Technical Education (ISTE) and the Computer Society of India (CSI).

Tanmay Nigade received his Bachelor of Engineering (B.E.) degree in Information Technology from Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India. He has expertise in Machine Learning, Deep Learning and Natural Language Processing, with extensive experience in big data handling. His work focuses on developing scalable machine learning models and optimizing Deep Learning algorithms for various applications.

Suyash Khodade received his Bachelor of Engineering (B.E.) degree in Information Technology from Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India. He is skilled in data analysis, GUI development, and backend operations. His professional interests include creating intuitive user interfaces and ensuring seamless integration between front-end and back-end systems for enhanced user experience.

Prathamesh Pimpalkar received his Bachelor of Engineering (B.E.) degree in Information Technology from Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India. His expertise lies in data analysis, GUI development, data management, and testing methodologies. He focuses on improving data accuracy and reliability, as well as developing efficient testing protocols to ensure high-quality software performance.