iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VII Month of publication: July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73134

Multimodal Framework for Hate Speech Recognition Using BERT and Deep Learning Architectures

SMD Shafiulla[1], L. Shiva Kumar[2], N. Manideep[3], S. Varshith Reddy[4]
[1]Assistant Professor, [2, 3, 4]B.Tech Student, Department of Computer Science & Engineering, Scient Institute of Technology, Ibrahimpatnam, R.R Dist [India]

Abstract: Hate speech recognition is a challenge in Automatic Speech Recognition systems which is used to recognize the unwanted text, audio, which is been generated by the various social media platforms. In the recent study Bidirectional Encoder Representational Transformer (BERT) model and Natural language processing (NLP) used to build a model which is used to recognize the hate speech, audio, and process that hated content. In the previous work this learning used only for the text but in this project, it used for both audios. In this paper we implemented a hate speech recognition model which are used to recognize the hated content more accurately with an accuracy of 94%.

Keywords: BERT Model, Hate Speech Identification, Twitter, Hateful Tweets, Transfer learning.

I. INTRODUCTION

There is a startling rise in racism and intolerance occurring all around the world. The popularity of memes led to an incredible 180 million posts on various social media platforms in 2018 alone. The startling increase in hate speech (HS) in this digital environment has become a major social issue. Explicit attacks on people based on traits like racism, race, nationality, religion, gender, or other essential features are known as hate speech. Due of the prevalence of HS on digital platforms, big internet corporations like Facebook—which has millions of daily active users—have taken significant steps to protect their user base. Hate speech can be expressed verbally, in writing, or nonverbally, and it always aims to target individuals or groups based on their inherent characteristics, such as religion, ethnicity, nationality, or race. It is now essential to combat the growing prevalence of using machine learning techniques, especially deep learning techniques, to combat hate speech online. Real-time hate speech detection and filtering has become essential due to the extensive use of social media platforms as platforms for racism.

While hate speech can manifest in a multitude of formats on social media, including text, voice, photos, and videos, research efforts have predominantly centered on language based techniques. These encompass a spectrum of Natural Language Processing (NLP)[10] approaches, ranging from neural networks and n-grams , to graph-based models. However, a notable gap exists in comprehensive investigations into the analysis of multimedia data. This paper presents an innovative technique that harnesses the power of a multimodal DL architecture for toxic speech categorization. Leveraging the potency BERT and additional Transformer-based models encoder architectures, which have exhibited remarkable success across diverse NLP tasks, this approach generates vector space representations of Natural language conducive to deep learning models. In our endeavour to capture the unmoral language and speech embedding's, we turn to pre-trained models, recognizing the constraints imposed by the dataset's limited size. To process these embedding's, we employ two distinct downstream Architectures like CNN. To facilitate the computation to create vector-space representations for our hate speech dataset, we depend on the pre-trained BERT. BERT's distinctive architecture, effectively incorporating contextual information from both preceding and succeeding content across all its layers, empowers the retraining of Profound bidirectional representations from unannotated text. Recently, the problem of online abusive detection has attracted scientific attention. Proof of this is the creation of the third Workshop on Abusive Language Online3 or Kaggles Toxic Comment Classification Challenge that gathered 4,551 teams4 in 2018 to detect different types of toxicities (threats, obscenity, etc.). In this work, we mainly focus on the term hate speech as abusive content in social media, since it can be considered a broad umbrella term for numerous kinds of insulting user-generated content. Hate speech is commonly defined as any communication criticizing a person or a group based on some characteristics such as gender, sexual orientation, nationality, religion, race, etc. Hate speech detection is not a stable or simple target because misclassification of regular conversation as hate speech can severely affect users freedom of expression and reputation, while misclassification of hateful conversations as unproblematic would maintain the status of online communities as unsafe environments [2].

In light of the escalating challenge posed by hate speeches the digital age, this study addresses a critical research problem: the development of effective techniques for the real- time detecting and classifications of toxic speech across various forms of media. Through the integration of multimodal deep learning methods, we aim to pave the way for more comprehensive and robust hate speech recognition solutions. In this paper we proposed a transfer learning approach using BERT model to enhance hate speech detection.We introduced fine tuning strategies to examine the difference embedding layers of BERT in hate speech detection.Our experimental results shows the automatic classification of inappropriate language using transfer learning models and pretrained BERT model which delivers superior outcomes with an accuracy of about 94%

II. PREVIOUS WORKS

For the identification of harmful speech and related ideas, different methodologies have been put out in recent years. The accessible resources and approaches have been compiled in a few interesting surveys that we have identified.

In [1], researchers conduct experiments to understand and present a cutting-edge neural network explorer for adult speech (Zhang et al., 2018). They apply computer vision algorithms to identifying the feature captured by each neurons and visualize salient regions in the input stimuli. Additionally, they propose a technique for identifying the words that are most indicative of hate speech. Their findings shed light on areas for further development and highlight the effective features of neural networks.

Transfer learning, as explored in [2][10], aims to enhance the performance of focus on directing learners toward specific domains by utilizing knowledge obtained from related source domains. This technique decreases the necessity for extensive amounts of data from the target domain to train these learners. Transfer learning has garnered attention and exhibits potential in machine learning for its diverse application prospects. However, these surveys often present approaches in an isolated manner and do not encompass the latest advancements in transfer learning. While recent research has predominantly emphasized deep learning techniques, particularly neural architectures, the pioneering works of Badjatiya et al. [4] and Gam back et al. [5]Introduced recurrent neural networks (RNNs) and convolutional neural networks (CNNs) respectively, for detecting hate speech in tweets.

In [6][10], a comparative analysis is conducted between the SVM model and BERT (Bidirectional Encoder Representations from Transformers), an open-source model developed by Google, to evaluate their diachronic performance. BERT stands out as a significant advancement in machine learning for natural language processing (NLP) tasks.

Only recently, researchers have achieved successful training of BERT deep neural networks due to their high computational cost, as mentioned in [8]. In addition to the more modern BERT model, the study also provides an updated and more transparent SVM classifier, although it does not outperform BERT.

III. METHODOLOGY

A. Transformer-Based Models

The core approach involves fine-tuning pre-trained [10] language models such as BERT (Bidirectional Encoder Representations from Transformers) and multilingual BERT for hate speech detection. These models generate contextual embedding’s that capture the nuanced meaning of text, improving classification performance .The models are trained on large datasets (e.g., Twitter) to adapt to specific hate speech identification tasks, leading to significant improvements in accuracy (up to 90%) over traditional methods.

B. Transfer Learning Techniques

Transfer learning has emerged as a pivotal technique in enhancing hate speech detection, primarily due to its ability to leverage pre-trained models that have already captured extensive general language understanding. In the context of hate speech recognition, transfer learning involves initially training a model—such as BERT (Bidirectional Encoder Representations from Transformers)—on large-scale unlabelled or other extensive text datasets. This pre-training process enables the model to learn rich contextual embeddings and understand language semantics, syntax.

C. Neural Networks

Neural networks play a crucial role in modelling and classifying textual and multimodal data to accurately identify hateful content. Neural networks are computational models inspired by the human brain's interconnected neuron structure, capable of learning complex patterns and representations from data. Specifically, the project leverages advanced neural architectures such as Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks.

CNNs are used to capture local features in text, such as specific word sequences, which are indicative of hate speech. LSTMs, a type of recurrent neural network, excel at modelling sequential data and understanding the context over longer text spans.

D. BERT Model

Here, we analyze the BERT transformer model on the hate speech detection task. BERT is a multi-layer bidirectional transformer encoder trained on the English Wikipedia and the Book Corpus containing 2,500M and 800M tokens, respectively, and has two models named BERTbase and BERTlarge. BERTbase contains an encoder with 12 layers (transformer blocks), 12 self-attention heads, and 110 million parameters whereas BERT$_{large}$ has 24 layers, 16 attention heads, and 340 million parameters. Extracted embeddings from BERTbase have 768 hidden dimensions [4].Using BERT model we analyzed the contextual information extracted from BERT's pretrained layers and retuned it using annotated datasets. After retuning we updated the weights using labelled datasets. To perform hate speech recognition we used BERT base model to classify each tweet as Racism, Sexism, Neither or Hate, Offensive.

IV. DESIGN AND ARCHITECTURE

The general architecture of expert system is given in Fig 1.

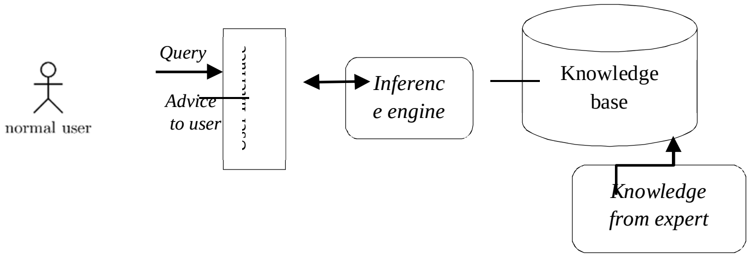


Fig.1.General architecture of expert system.

Expert systems performs decision making tasks.The working an expert system begins with the user who submits the query to through user interface and this query is processed by the inference engine which applies various rules and reasoning techniques to analyse the input.The inference engine then interacts with the knowledge base extracting relevant facts, rules. Based on this knowledge system derives conclusions and formulate the valid response.

The architecture of our work is shown below.

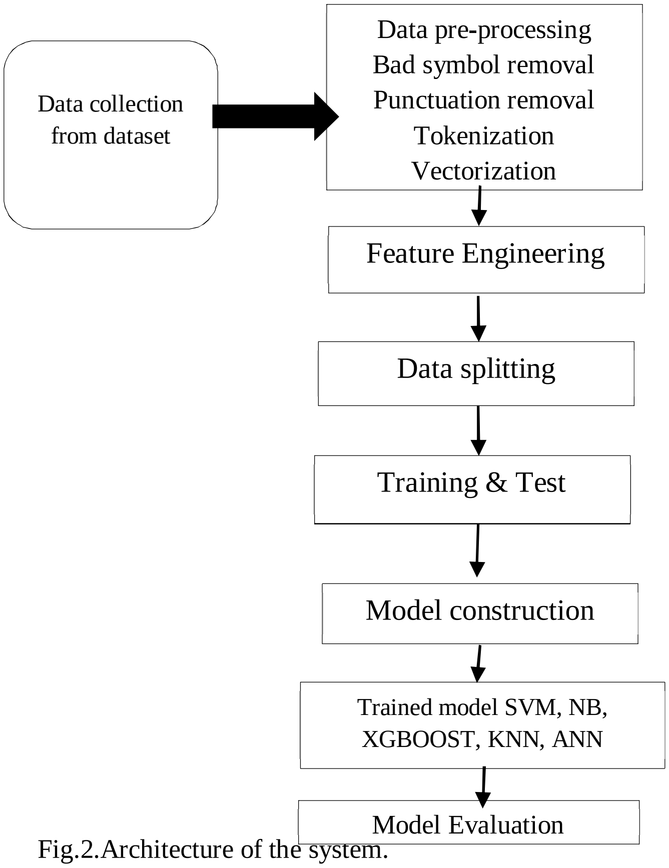


Fig.2.Architecture of the system.

The system architecture for the hate speech recognition system is designed around a multimodal framework that integrates both audio and textual data to enhance detection accuracy. The architecture comprises several key modules, starting with data collection, where tweets are gathered from social media platforms like Twitter, and audio clips are sourced from multimedia content. The preprocessing module cleans and normalizes the data by removing URLs, special characters, user handles, and balancing the dataset to ensure proportional representation of hate, offensive, and neutral categories. For textual analysis, pre-trained models such as BERT are employed to extract contextual embeddings, leveraging transfer learning to improve feature representations. Simultaneously, audio features are extracted using specialized algorithms to capture speech patterns indicative of hate speech. These features from both modalities are then fused within the classification framework, which utilizes neural network architectures like LSTM with attention mechanisms or transformer-based models fine-tuned for hate speech detection. The combined multimodal data is fed into the classifier, which predicts the presence or absence of hate speech.

V. IMPLEMENTATION AND RESULTS

We first introduce datasets used in our study and then investigate the different fine-tuning strategies for hate speech detection task. We also include the details of our implementation and error analysis in the respective subsections.

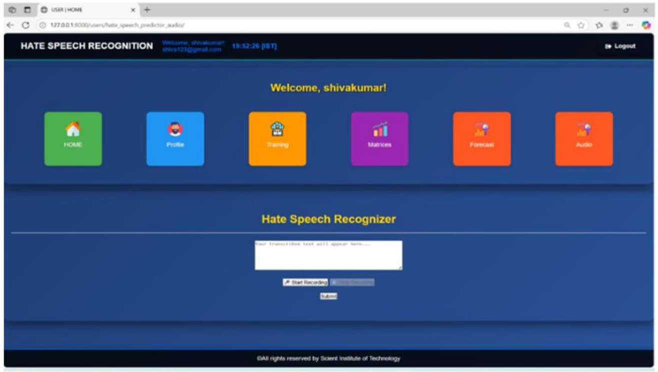
A. Dataset Description

We evaluate our method on the Twitter dataset which is closely related to multimedia domain.There were various dataset tested , and the right model for the feature extraction was extracted and all the datasets underwent the pre-processing techniques eliminating URL’s, special character. This dataset is accessible on Kaggle. Data was gathered from Twitter using keywords[11]. There are a total of 22083 tweets in it, which are divided into three categories: hatred, offensive, and neither. Hatred (7430, 33.65% of total), Offensive (7190, 32.56% of total), and neither (7463).

B. Results

For the implementation of our neural network, we used pytorch-pretrained-bert library containing the pre-trained BERT model

.

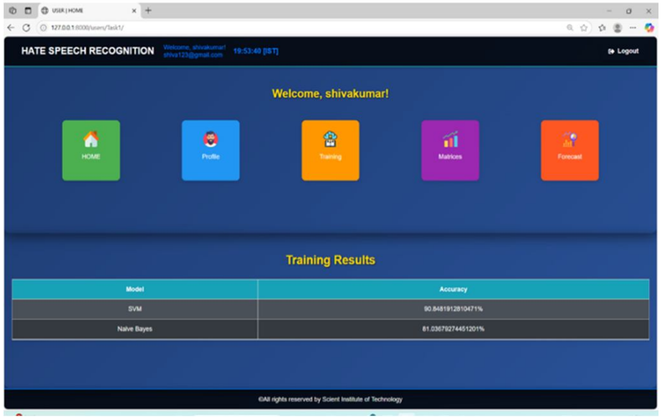


Welcome, shivakumar!

Hate Speech Recognizer

©All rights reserved by Scient Institute of Technology

Fig.3.Audio page of hatespeech recognizer



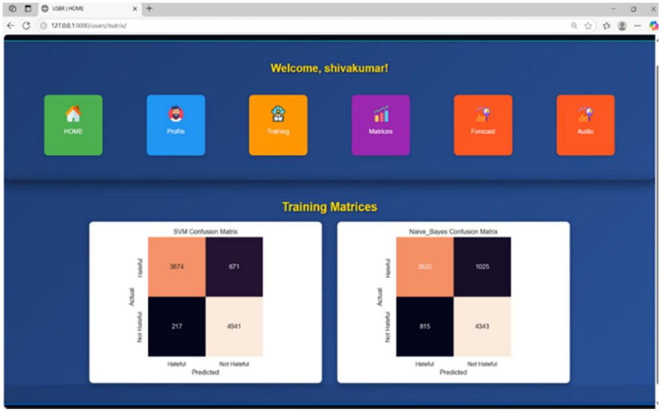
Welcome, shivakumar!

Training Results

Model	Accuracy
SVM	90.84819128104471%
Naïve Bayes	81.03679274451201%

©All rights reserved by Scient Institute of Technology

Fig.4.Training results



Welcome, shivakumar!

Training Matrices

SVM Confusion Matrix

Actual \ Predicted	Hateful	Not Hateful
Hateful	3874	671
Not Hateful	217	4941

Naive_Bayes Confusion Matrix

Actual \ Predicted	Hateful	Not Hateful
Hateful	3520	1025
Not Hateful	815	4343

Fig.5.Matrices page

VI. CONCLUSION

This project successfully demonstrates leveraging pre-trained models like BERT for hate speech detection across both textual and audio modalities. The findings shows that combination of audio and textual features enhances the accuracy and robustness of the hatespeech recognition.The proposed framework based on transfer learning and deep learning neural network offers a promising solution for real time and scalable moderation of abusive online content with an accuracy of 94%.Future directions will be directed in detecting hatespeech in video transcriptions.

REFERENCES

[1] Kulsoom, F., Narejo, S., Mehmood, Z. et al. A review of machine learning-based human activity recognition for diverse applications. Neural Comput & Applic 34, 18289–18324 (2022). https://doi.org/10.1007/s00521-022-07665-9.

[2] Narejo, Sanam et al. ‘Big Data Analytics and Classification of Cardiovascular Disease Using Machine Learning’. 1 Jan. 2022 : 2025– 2033.

[3] Khan, Wisal & Turab, Muhammad & Ahmad, Waqas & Ahmad, Syed & Kumar, Kelash & Luo, Bin. (2022). Data Dimension Reduction makes ML Algorithms efficient. 10.48550/arXiv.2211.09392.

[4] Sarwar, Savera, et al. "Advanced Audio Aid for Blind People." 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC). IEEE, 2022.

[5] Khan, Wisal, et al. "Data Dimension Reduction makes ML Algorithms efficient." 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC). IEEE, 2022.

[6] Kumar, Teerath, et al. "Forged character detection datasets: passports, driving licences and visa stickers." Int. J. Artif. Intell. Appl.(IJAIA) 13 (2022).

[7] P. Mishra, M. D. Tredici, H. Yannakoudakis, and E. Shutova, “Abusive Language Detection with Graph Convolutional Networks,” in NAACL, 2019.

[8] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, “Deep Learningfor User Comment Moderation,” in Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, August 2017, pp. 25–35. [Online]. Available: 10.18653/v1/W17- 3004

[9] Turab, Muhammad & Jamil, Sonain. (2023). A Comprehensive Survey of Digital Twins in Healthcare in the Era of Metaverse. BioMedInformatics. 3. 563-584. 10.3390/biomedinformatics3030039.

[10] SMD SHAFIULLA(2023). A review on Natural Language Processing techniques using Qualitative Research,363-367.

[11] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi,2019, A BERT-Based Transfer Learning Approach for ,Hate Speech Detection in Online Social Media,