IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

Hate Speech Detection using Deep Learning

Aswathy Biju[1] and Sanooja Beegam[2]
Student, Department of Computer Application[1]
Assistant Professor, Department of Computer Application[2]
Musaliar College of Engineering and Technology, Pathanamthitta, Kerala, India

Abstract: *Hate speech detection is a crucial task in natural language processing, aimed at identifying and mitigating offensive and harmful content online. In this study, we propose an innovative approach for hate speech detection that combines deep learning based Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. Our model is designed to make effective local and global dependencies in the document.This approach provides a framework for identifying hate speech by using the combined capabilities of deep CNNs and Bi-LSTMs, setting the groundwork for the creation of more advanced and precise detection systems to promote safer online environments.*

Keywords: CNN, Bi-LSTM, Deep learning, Natural language processing

I. INTRODUCTION

Hate speech detection is an important task in the field of natural language processing, aiming to identify and combat harmful and offensive language online. With the increasing prevalence of hate speech on social media platforms and online communities, there is a growing need for automated tools that can effectively detect and mitigate such content. Deep learning techniques, including Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) models, have shown promise in hate speech detection by leveraging their ability to capture intricate patterns and contextual dependencies within text data. This research focuses on the use of deep learning models, especially CNN and BiLSTM, to recognize hate speech. CNNs are good at extracting local features from input data, making them ideal for data analysis. On the other hand, BiLSTMs can capture progress and divergent content, enabling them to better understand the meaning and purpose behind the text.

II. PROPOSED SYSTEM

In this project, our goal is to classify content using a Convolutional Neural Network (CNN) model and Bidirectional Long Short Term Memory (Bi-LSTM). CNNs and Bi-LSTMs are neural network architectures known for their ability to process sequential data and capture patterns and dependencies in the data. CNNs excel at extracting local features from input data, making them well-suited for analysing textual information. On the other hand, BiLSTMs are capable of capturing sequential dependencies and contextual nuances, allowing them to effectively understand the meaning and intent behind the text. The main objective of this research is to develop a robust hate speech detection system that can accurately identify and classify hateful content in real-time.

III. METHODOLOGY

The methodology for hate speech detection utilizing CNN and BiLSTM models consists of several distinct steps. Firstly, a carefully labelled dataset containing instances of hate speech, offensive language, and non-offensive language is collected. Subsequently, the textual data undergoes pre-processing, which involves tokenization and stemming, to prepare it for the subsequent model input. The CNN component is responsible for capturing local features and patterns within the text, while the BiLSTM component focuses on capturing contextual information and dependencies. The outputs from both components are then combined and fed into a fully connected layer for the classification task. The model is trained using appropriate loss functions and optimized through the application of back propagation. To assess the model's performance, evaluation metrics such as accuracy and F1 score are computed. Finally, the trained model can be deployed to detect hate speech in real-time, promoting a safer online environment.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

IV. SYSTEM ARCHITECTURE

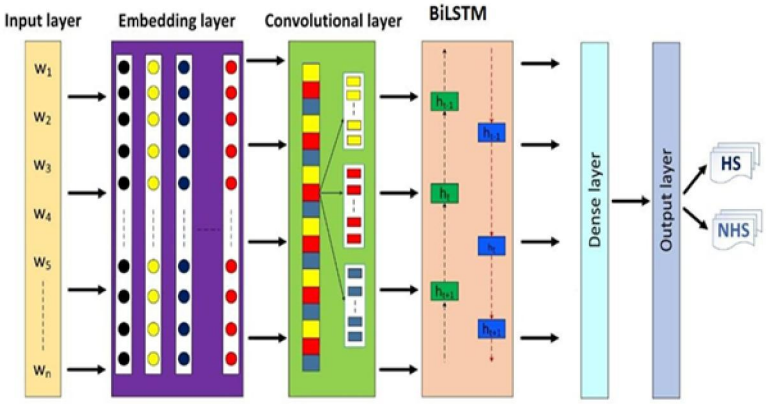

Fig 1.CNN and BiLSTM System Architecture

The proposed model for hate speech detection using CNN and BiLSTM consists of several layers, each serving a specific role in processing the input text data and producing a binary classification label.The hate speech detection model utilizing CNN and BiLSTM consists of an input layer for receiving and processing text data, an embedding layer to convert words into dense vectors, a convolutional layer for capturing local patterns, a max pooling layer for selecting salient features, a BiLSTM layer for contextual information processing, a dense layer for combining extracted features, and an output layer for binary classification of hate speech presence. .By utilizing these layers in CNN-BiLSTM architecture, the model can effectively process and analyse the input text data, capturing local patterns, contextual dependencies, and semantic information to make accurate predictions regarding hate speech classification**.**

V. COMPARISON AND RESULTS

The project "Hate Speech Detection using CNN and Bi-LSTM" presents an insightful comparison and analysis of the results obtained from the hate speech detection system. The evaluation metrics, including the confusion matrix, accuracy, and loss, provide valuable insights into the system's performance.The confusion matrix offers a comprehensive breakdown of the model's predictions. In this case, the confusion matrix reveals 563 true positives, 57 true negatives, 124 false positives, and 116 false negatives. These numbers indicate that the system correctly identified a significant number of hate speech instances while also correctly classifying a substantial portion of non-hate speech instances. However, there is still room for improvement, particularly in reducing false negatives to enhance the system's sensitivity in detecting hate speech.

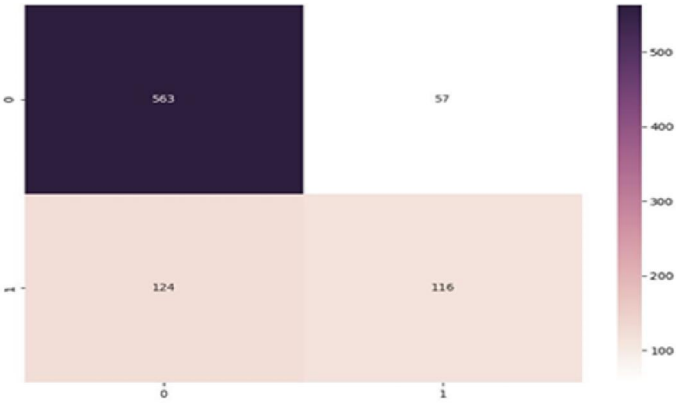

Fig 2.Confusion matrix

The accuracy of the hate speech detection system is reported at 72%. While this accuracy indicates a reasonably reliable performance, it is important to note that there is potential for further enhancements to increase the system's overall accuracy. Fine-tuning the model, exploring different architectural configurations, or incorporating additional data for training could help improve the accuracy of the system.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

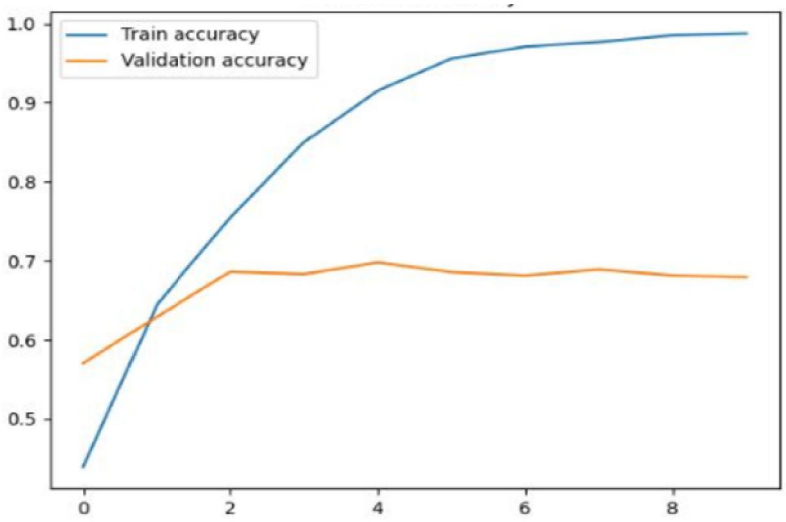

Fig 3.Model Accuracy curve

During the training process, the loss function steadily decreased, indicating that the model effectively learned the patterns and features associated with hate speech. This convergence of the loss function demonstrates successful training of the CNN and Bi-LSTM models.

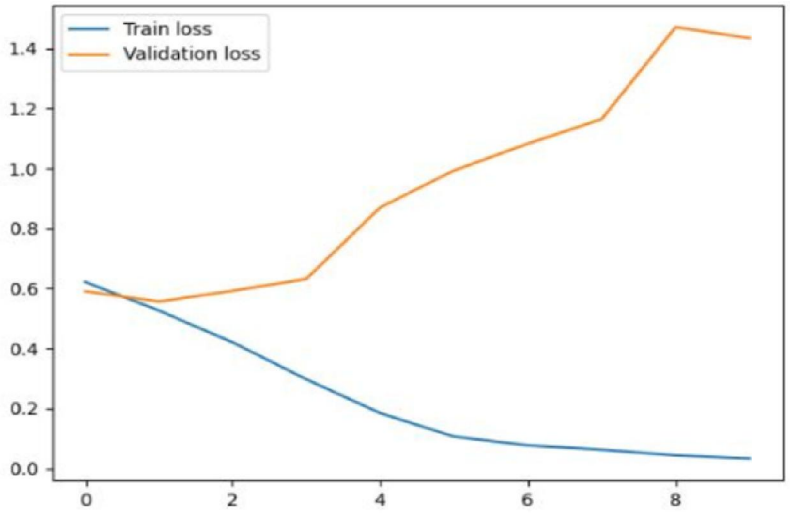

Fig 4.Model loss curve

VI. SAMPLE OUTPUT

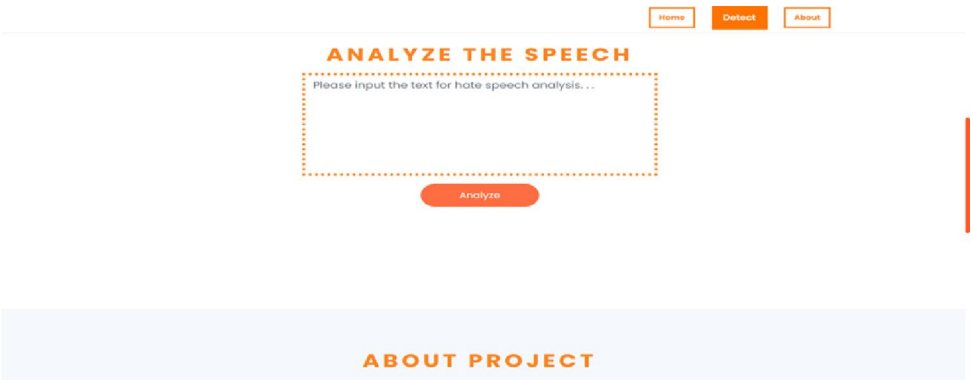

Fig 5.Sample output for hate speech analysis

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

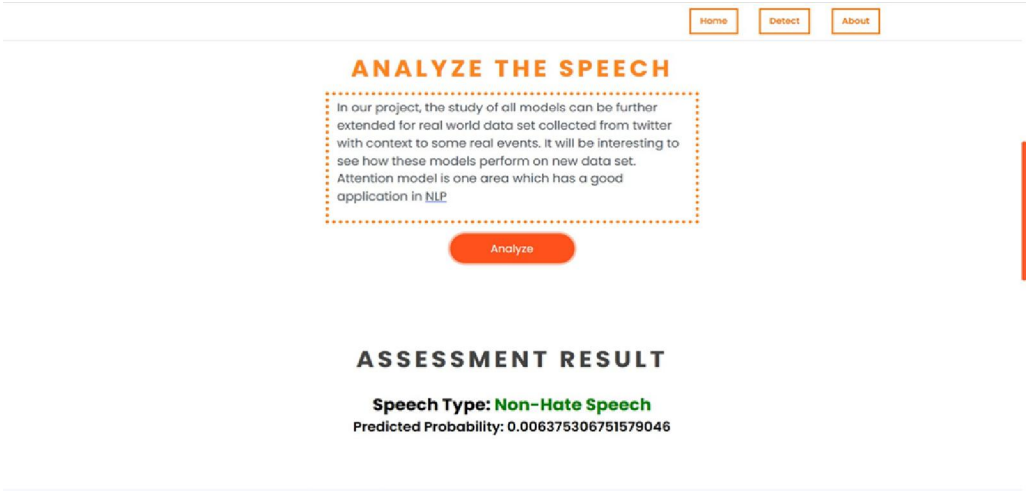

Fig 6.Sample output for non-hate speech detection

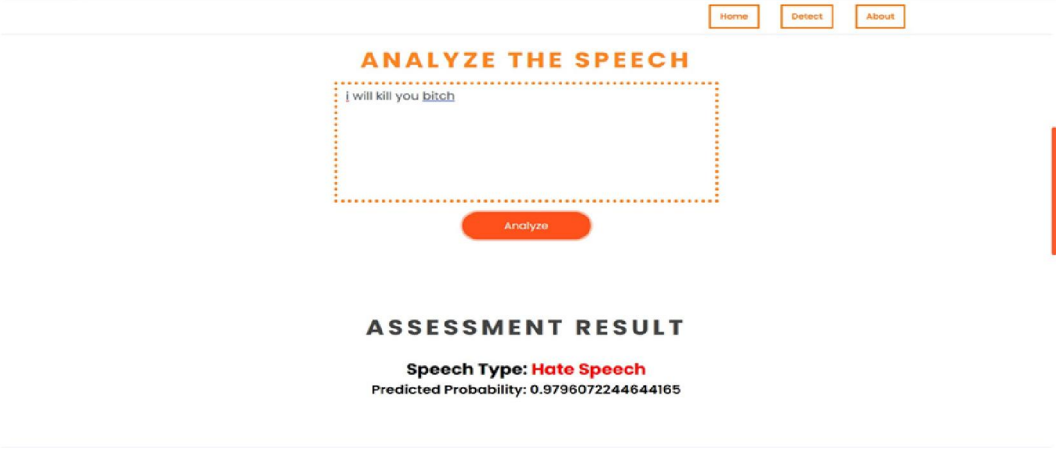

Fig 7.Sample output for hate speech detection

VII. CONCLUSION AND FUTURE SCOPE

Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) models, for hate speech detection has demonstrated promising results. The integration of CNNs and BiLSTMs enables the extraction of local features and contextual information, facilitating effective identification of hate speech and offensive language in online platforms.These deep learning models have shown proficiency in capturing patterns and representations from textual data, allowing for automated classification and flagging of harmful content. The reviewed literature emphasizes the superiority of hybrid CNN-BiLSTM architectures in accurately identifying instances of hate speech, surpassing the performance of traditional machine learning approaches.

In conclusion, the future scope of hate speech detection using deep learning in CNNs and BiLSTMs is promising. By focusing on dataset diversity, fine-grained classification, multi-modal approaches, and ethical considerations, researchers can advance the field and contribute to the development of more robust and inclusive hate speech detection systems.

REFERENCES

[1] C. Blaya, “Cyberhate: A review and content analysis of intervention strategies,” Aggress. Violent Behav., no. May, pp. 0–1, 2018.

[2] Z. Waseem and D. Hovy,“Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” Proc. NAACL Student Res. Work., pp. 88–93, 2016.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

[3] 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.
[4] F. Salem, "Arab Social Media Report 2017: Social Media and the Internet of Things: Towards DataDriven Policymaking in the Arab World," 2017.omputer Science & Information Technology (CS & IT) 97
[5] F. Miro-Llinares and J. J. Rodriguez-Sala, "Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy," Int. J. Des. Nat. Eco dynamics, vol. 11, no. 3, pp. 406–415, 2016.