IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

Ensemble Approach for Hostile Discourse Detection

J Sreedevi[1], M Srikanth Sagar[2], Joginipally Saanvi[3], Nannepamula Harshita Glady[4]

MTech, Department of Emerging Technologies, Mahatma Gandhi Institute of Technology, Hyderabad, India[1]

MTech, Department of Emerging Technologies, Mahatma Gandhi Institute of Technology, Hyderabad, India[2]

BTech, Department of Emerging Technologies, Mahatma Gandhi Institute of Technology, Hyderabad, India[3]

BTech, Department of Emerging Technology, Mahatma Gandhi Institute of Technology, Hyderabad, India[4]

Abstract: Hostile discourse, characterized by discriminatory language, expressions of hate, or overt aggression based on individual or group identity, presents a formidable challenge in online communication. This article is an in-depth study of hate speech research, specifically the definition and classification of hate speech in text. Through a comprehensive review, the research explores various techniques, ranging from classical machine learning algorithms to advanced deep learning models such as convolutional neural networks, short-term memory networks, gated recurrent units, and transformer-based architectures, with a special focus on bidirectional LSTM with self-generating mechanisms and feedforward neural networks. Moreover, the paper offers practical insights for effective model development, emphasizing the necessity of harnessing large-scale social media datasets, ensuring data balance for representative training, implementing regularization techniques for improved generalization, and incorporating a validation set for accurate performance evaluation. By combining theories from a variety of research methods and using an integrated approach from diverse models, this study aims to provide researchers and practitioners with a conceptual framework for developing powerful models that are effective. In summary, this article highlights the importance of adapting technology to the dynamic field of online communication, with the overall goal of promoting security and benefiting diverse communities..

Keywords: Hostile Discourse Detection, Hate Speech, Machine Learning Models, Data Preprocessing, Model Training, Ensemble Approach

I. INTRODUCTION

In natural language processing (NLP), the search for negative language has emerged as an important area of research aimed at identifying and analysing incidents of discrimination, hate speech, and violence in online text or multimedia content. As digital communication platforms have become ubiquitous, the proliferation of negative messages poses a serious challenge to online practice and security [1, 2, 8]. These messages often target attributes such as gender, religion, or other identifying characteristics, inciting hatred or violence towards a person or group. In addition to creating online hostility, such comments have divisive and destructive consequences in the real world [1, 2, 8]. Identifying and reducing harmful content plays a crucial role in maintaining a safe and inclusive online environment. It also promotes responsible and ethical digital participation by helping platforms comply with social rules and regulations.

Investigating complaints can help prevent cyberbullying, online harassment, and the spread of negative sentiments [1,2,8]. From classical machine learning algorithms such as support vector machines and logistic regression to deep learning methods like recurrent neural networks and transformer-based architectures [1, 6, 8], various approaches have been explored. This comprehensive analysis includes predefined features and complex relationships within the data. The dynamic nature of language, evolving linguistic tendencies, and situational variability add inherent complexity to the task. Efforts must also be made to ensure fairness and accuracy, as biases in the training data can affect the results [1, 6, 8].

Future research may focus on improving models to capture cultural nuances, address online toxicity, and enhance self-regulation capabilities [1, 6, 8]. Collaboration between researchers, industry stakeholders, and policymakers is essential to develop effective and ethical solutions that contribute to a safer and more inclusive digital environment. We propose an ensemble approach, leveraging various methods to improve the accuracy and potential of our system. By combining different techniques and models, we aim to address the dynamic nature of hostile discourse and ensure the reliable development of research in this field.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

II. RELATED WORK

Two significant studies on the spread of hate speech on social media platforms [1] and [4] provide insights into diagnosing and mitigating this issue. Study [1] emphasizes the importance of trust in online communication, using carefully selected data to provide a solid foundation for training learning models. Similarly, [4] addresses discrimination in the Malay language with a unique approach tailored to digital communication discourse. Both studies highlight the importance of integrating different cultural perspectives into discrimination research for a better understanding of the problem. Studies [2] and [5] tackle the limitations associated with manual annotation and content filtering techniques. Study [2] discusses the challenges in dealing with hate speech and cyberbullying in Taiwanese political discourse, while [5] analyzes anti-racism and hate speech in the South African digital environment, emphasizing the difficulty of measuring discrimination patterns' specificity and generality. Both studies underscore the need to understand context to effectively improve detection accuracy.

New approaches have been introduced to solve these problems. Study [3] combines various neural network architectures, recognizing the sensitivity of hyperparameters in functional models. In contrast, [6] evaluates the role of information content in improving detection accuracy, acknowledging the additional difficulty and resource consumption associated with the method. Despite these challenges, these studies aim to enhance the accuracy and effectiveness of hate speech detection, including in less commonly studied languages. Study [7] uses meta-learning to address micro-discrimination in different languages, achieving better performance compared to other classification algorithms on Bengali social media. These studies represent significant progress in promoting hate speech detection.

III. PROPOSED STYLE

The plan integrates several deep learning architectures, each carefully designed to address differences in information about attacks. By partnering with these standards in a unified way, we work to strengthen online security and improve the health of digital communities.

A. System Architecture:

Our design model is based on ensemble learning, a method of combining predictions from multiple frameworks to improve accuracy and robustness. At its core is a meta-model that manages the outputs of individual models to produce composite predictions. Through collaborative learning, our system delivers comprehensive data analysis and effectively detects subtle language patterns indicating abusive language and online toxicity.

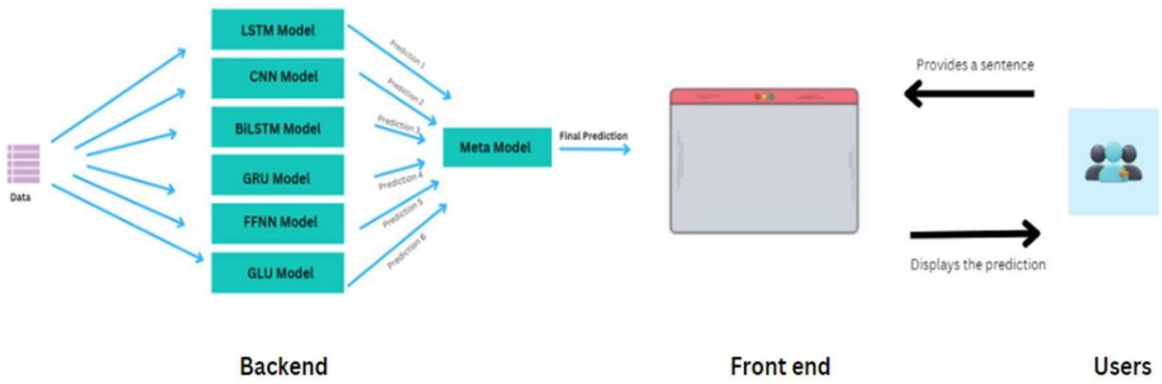

Figure 1: System Architecture of Hostile Discourse Detection

B. Model Architecture

Our system consists of a set of deep learning methods, each carefully designed to detect specific content related to defamation. LSTM models excel at capturing long-term dependencies, while CNN models are effective at extracting features from local patterns. The GRU model is adept at examining remote dependencies. Combining these models results in a comprehensive evaluation of data files, enhancing the system's efficiency in detecting and processing defamation cases and providing a safer online environment for all users.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

```
Meta-Model Architecture:
Model: "model_13"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_19 (InputLayer)       [(None, 3)]               0         []

 input_20 (InputLayer)       [(None, 3)]               0         []

 input_21 (InputLayer)       [(None, 3)]               0         []

 input_22 (InputLayer)       [(None, 3)]               0         []

 input_23 (InputLayer)       [(None, 3)]               0         []

 input_24 (InputLayer)       [(None, 3)]               0         []

 concatenate_1 (Concatenate  (None, 18)                0         ['input_19[0][0]',
 )                                                                 'input_20[0][0]',
                                                                   'input_21[0][0]',
                                                                   'input_22[0][0]',
                                                                   'input_23[0][0]',
                                                                   'input_24[0][0]']

 dense_18 (Dense)            (None, 64)                1216      ['concatenate_1[0][0]']

 dense_19 (Dense)            (None, 3)                 195       ['dense_18[0][0]']

=================================================================================================
Total params: 1411 (5.51 KB)
Trainable params: 1411 (5.51 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
None
```

Figure 2: Model Architecture of Meta-Model

C. Data Collection and Preprocessing:

To ensure rich and representative data, we use Kaggle's Hate Speech and Offensive Language Dataset (Twitter Profile), which contains 24,783 incidents. Strict data preprocessing, including case conversion, removal of URLs, numbers, mentions, hashtags, and tags, and tokenization using NLTK's word tokenization feature, ensures clean and well-organized data input.

D. Model Training and Evaluation:

Following standard procedures, we train each baseline model and carefully evaluate its performance based on various parameters such as accuracy, precision, recall, F1 score, and area under the ROC curve. The combined model achieved an accuracy of 93.36%, demonstrating its strong capability in detecting hostile discourse.

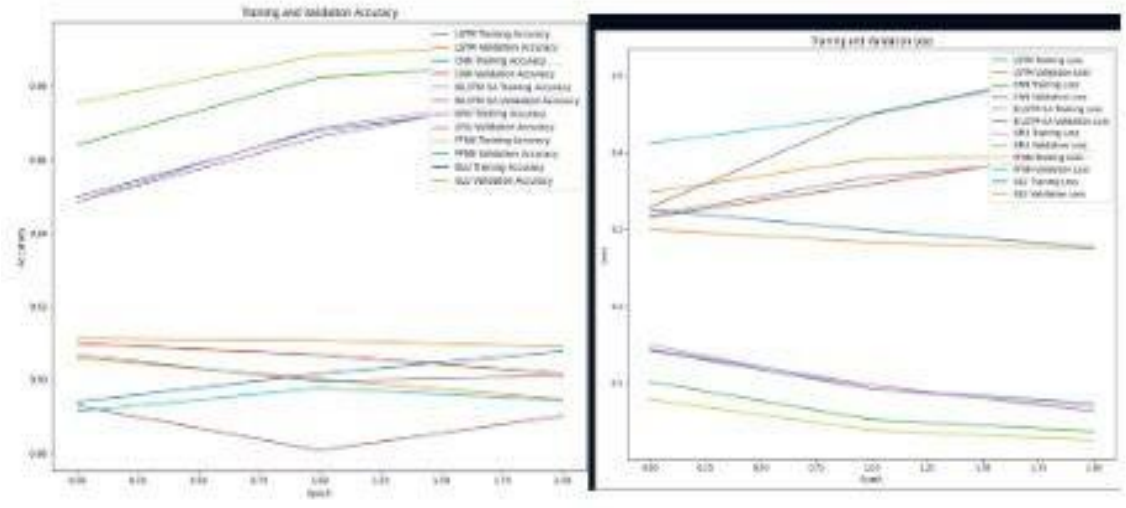

Figure 3: Training and Validation Graphs

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

E. Continuous Improvement and Adaptability:

Our approach emphasizes continuous monitoring and improvement to adapt to evolving languages and online communication patterns. Preliminary evaluations of new data yield promising results, showcasing the system's ability to expand to a wide range of topics and datasets.

F. Frontend User Interface and Web Service Deployment:

Our front-end user interface, built using the Streamlit library, is designed for simplicity and efficiency. This user-centered design supports seamless interaction with intuitive input fields for text entry. Using the FastAPI web framework, we deploy our model as a web service, extending its accessibility and making its services more visible to users. The statistics, including the accuracy of our combined model (93.36%) and the average accuracy of individual models, highlight the system's performance and reliability in combating online toxicity and fostering healthy digital communities.

IV. EXPERIMENTAL RESULTS

Our research on the effectiveness of hostile discourse detection aims to produce useful results across a range of performance measures. By utilizing its self-generating mechanism, the LSTM model integrates CNN, BiLSTM, GRU, FFNN, and GLU models into a unified framework, achieving main capabilities with accuracy rates ranging from 65% to 91%. The LSTM model, with an average accuracy of 83.6%, is followed by the CNN model, which achieved an accuracy of 88.7%. This showcases the fundamental robustness and adaptability of these models in capturing the complex nuances of hostile behavior in discussions across different contexts.

The analysis includes precision, recall, F1 score, and area under the ROC curve (AUC-ROC). These measurements provide a comprehensive understanding of the model's ability to accurately identify hate speech, insults, and negative rhetoric. This detailed analysis enhances our understanding of the system dynamics and its discrimination capabilities.

Figure 4: Predicts as Offensive Language

The analysis of training and validation curves indicates the model's capability to learn and adapt without signs of overfitting, suggesting the model's robustness. It exhibits minimal latency during reflection, making it suitable for real-time deployment on online platforms and social media. This quick response time enhances the system's efficiency in timely detection and mitigation of malicious content.

Figure 5: Predicts as Hate Speech

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

Positive feedback from users interacting with the front-end interface significantly improved our analysis. The interface allows users to input text, and the system provides predictions that classify the input as hateful, offensive, or neither. This real-world feedback supports our evaluation, reflecting the system's practical performance and effectiveness in identifying gender and speech conflicts. The user interface's ease of use and the system's accuracy in predictions underscore its feasibility and effectiveness.

By combining careful analysis of various performance parameters with advanced deep learning technologies, we have established a solid foundation for deploying the system online. These findings have significant implications for protecting digital environments and supporting safe online communities. Future work will focus on continuous improvement and adaptation of the model to evolving language patterns and contexts, ensuring its ongoing relevance and effectiveness in combating hostile discourse.

Figure 6: Predicts as Neither

V. CONCLUSION

Our research proposes unique solutions designed to identify and reduce negative speech online using advanced natural language processing (NLP) techniques within a user-friendly interface. Our ensemble models deliver high accuracy and efficiency across various media formats, including text, images, and audio, through careful analysis.

Our combined method demonstrates its effectiveness in discourse classification, achieving an overall performance of 93.36%, surpassing individual baseline models. It enhances system availability and user interaction by providing a seamless platform for monitoring and intervention. Continuous improvement, fine-tuning of datasets, and enhanced forecast accuracy are essential to our approach. The system's low latency is crucial for timely detection and mitigation of harmful content, promoting a safer and more effective digital environment.

Our program promises to increase accuracy, trust, and user engagement, effectively combating bad behavior while improving the health of online communities. Addressing hostile discourse is essential, highlighting the need for advanced NLP techniques and thoughtful user interface design to secure the digital environment.

REFERENCES

[1]. Qureshi, Khubaib Ahmed, and Muhammad Sabih. " Un-compromised credibility: Social Media based multi- class hate speech classification for text." IEEE Access 9 (2021): 109465109477.
[2]. J. Wang, Chih-Chien, Min-Yuh Day, and Chun-Lian Wu. ”Political Hate Speech Detection and Lexicon Building: A Study in Taiwan.” IEEE Access 10 (2022): 44337-44346.
[3]. Khan, Shakir, et al. ”HCovBi-caps: hate speech detection using convolutional and Bidirectional gated recurrent unit with Capsule network.” IEEE Access 10 (2022): 7881-7894.
[4]. Maity, Krishanu, et al. ”A deep learning framework for the detection of Malay hate speech.” IEEE Access (2023).
[5]. Oriola, Oluwafemi, and Eduan Kotze. ”Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets.” IEEE Access 8 (2020): 2149621509.
[6]. Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. ”Advances in machine learning algorithms for hate speech detection in social media: a review.” IEEE Access 9 (2021): 88364-88376.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13558

[7]. Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. ”Cross-lingual few-shot hate speech and offensive languagedetection using meta learning.” IEEE Access 10 (2022): 14880-14896.

[8]. Keya, Ashfia Jannat, et al. ”G-bert: an efficient method for identifying hate speech in Bengali texts on social media.” IEEE Access (2023)

[9]. Keya, Ashfia Jannat, et al. "G-bert: an efficient method for identifying hate speech in Bengali texts on social media." IEEE Access (2023).

[10]. Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources." SN Computer Science 2.2 (2021): 95.

[11]. Mohiyaddeen, Mr, and Sifatullah Siddiqi. "Automatic hate speech detection: A literature review." Available at SSRN 3887383 (2021).