iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66743

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Racism Detection through Sentiment Analysis of Tweets: A Review

Sathyanarayan S[1], Ruman Ahamed Baig[2], Sijan Khan[3], Syed Sadiq[4], Pramath S[5]
Department of CS&E, JNN College of Engineering, Shivamogga, Karnataka, India

Abstract: ***Racism detection in social media content has become a critical area of research, particularly in addressing the harmful impact of offensive language on online communities. This paper explores the application of sentiment analysis techniques in detecting racism in tweets, combining BERT (Bidirectional Encoder Representations from Transformers) for text classification and VADER (Valence Aware Dictionary and sentiment Reasoner) for sentiment analysis. The study demonstrates the potential of this hybrid approach to provide real-time classification, visualize sentiment distributions, and analyze the impact of user comments, contributing to efforts in fostering safer online environments.***

Keywords: Sentiment analysis, Machine learning, Hybrid approach, Online safety, Text classification, Hate speech detection, BERT, Tweet analysis.

I. INTRODUCTION

In the age of digital communication, detecting and combating online racism has become a pressing concern. Social media platforms, while fostering diverse conversations, also facilitate the spread of harmful and discriminatory content. This paper explores the use of sentiment analysis techniques, specifically combining BERT (Bidirectional Encoder Representations from Transformers) and VADER (Valence Aware Dictionary and sEntiment Reasoner), to accurately identify and flag racist content in tweets. The paper highlights the potential of this approach to enhance the detection of subtle racist language, Ensure accountability, and contribute to creating a more inclusive online environment.

II. LITERATURE SURVEY

In this section, various authors have presented various approaches and hate speech detection techniques.

In [1], The authors investigate the feasibility of using domain-specific word embeddings alongside BERT and Bidirectional LSTMs for hate speech detection. The model was trained on a labeled hate speech dataset, leveraging text preprocessing techniques like tokenization and embeddings. The study evaluates its models using accuracy, precision, recall, and F1-score. The research highlights that BERT-based methods significantly outperformed traditional ML models in capturing semantic nuances of hate speech. The integration of domainspecific embeddings improved accuracy by 8%. Furthermore, the study detailed how fine-tuning BERT with task-specific data enhances its capacity to understand contextual and implicit hate speech. However, the scope is limited to English datasets, and multilingual capabilities or cross-domain applicability were not explored.

In [2], the integration of BERT with neural networks like CNNs, MLPs, and LSTMs, combined with ensemble learning techniques, to enhance hate speech detection on Twitter. Leveraging transfer learning, the researchers fine-tuned BERT and implemented hybrid architectures, achieving superior classification performance compared to standalone models. Ensemble methods, including soft voting, hard voting, and stacking, were used to aggregate model outputs, yielding improved robustness and accuracy. The study also introduced the DHO dataset by merging multiple public datasets to address data imbalance issues, demonstrating resilience in multi-label classification tasks. Despite challenges like high resource demands, the hybrid and ensemble approaches achieved notable results, with F1 scores of up to 97% on the Davidson dataset and 77% on the DHO dataset, showcasing the effectiveness of combining BERT with deep learning and ensemble strategies.

In [3], exploration of the analysis of public sentiment on COVID-19 and its variant Omicron using sentiment analysis techniques. Tweets were collected from Twitter over seven days using hashtags like “covid-19” and “omicron,” forming two datasets. The study applied VADER and BERT for sentiment analysis and compared the performance of five supervised machine learning algorithms. Findings show that most tweets conveyed negative sentiments. In classification tasks, Support Vector Machine (SVM) achieved the highest accuracy of 92% using BERT on the Omicron dataset. BERT consistently improved performance across algorithms, except in one case where it underperformed compared to Naive Bayes. Random Forest (RF) also performed well, while XGBoost often gave the lowest accuracy. The study concludes that BERT, as a transformer-based model, excels in sentiment classification and suggests future exploration of other algorithms for improved accuracy. Potential applications include analyzing public sentiment on emerging COVID-19 variants, vaccinations, and broader social issues.

In [4], The focus is on the detection of hate speech on Twitter using a website-based simulation. Twitter allows users to freely exchange information and opinions, which can lead to exposure to hate speech. Victims of hate speech may experience mental health challenges due to verbal and emotional attacks. The research addresses the lack of effective countermeasures for detecting hate speech on Twitter. The study developed a website where users can input text to detect hate speech. The website preprocesses the input and analyzes it using the BERT algorithm, classifying whether the text constitutes hate speech. Results from the training phase revealed that the BERT-based hate speech detection achieved an accuracy of 78.69%, a precision of 78.90%, a recall of 78.69%, and an F1 score of 78.77%. The website provides an accessible tool for identifying hate speech on the Twitter platform.

In [5], Authors examine the effectiveness of various Natural Language Processing (NLP) models for classifying tweets into three categories: hate speech, offensive language, and neither. Four deep learning models were evaluated: a baseline CNN, BiDirectional LSTM with an attention layer, pre-trained BERT, and a fine-tuned RoBERTa transformer. These models were assessed using metrics such as precision, recall, F1-score, accuracy, and Matthews Correlation Coefficient (MCC). The baseline CNN models, utilizing Adam and RMSProp optimizers, showed moderate performance, with slightly better results for the Adam variant. The BiDirectional LSTM model with attention demonstrated excellent training performance but exhibited overfitting, as indicated by a decline in validation performance. The pre-trained BERT and fine-tuned RoBERTa models showed robust performance with high scores across all metrics. Among these, the fine-tuned RoBERTa transformer consistently achieved the best results, with MCC values of 0.84 on the training set and 0.82 on the test set.

In [6], exploration of the challenges and advancements in hate speech detection, with a particular focus on the complexities of multilingual settings, such as Hindi and Roman Hindi. The research evaluates various methods for detecting hate speech in online content, including normalization of lexical variations, traditional machine learning techniques, and more recent deep learning approaches. It highlights the growing problem of hate speech on social media platforms, which is exacerbated by the increasing use of code-mixed languages like Hinglish. Traditional methods such as lexicon-based and machine learning models have been employed, but these approaches often struggle with nuances, context, and multilingual datasets. The study shows that deep learning models like CNNs, BiLSTMs, and BERT-based transfer learning offer significant improvements in detecting hate speech, especially when trained on large datasets. Fine-tuning pre-trained models like BERT and RoBERTa has been shown to provide state-of-the-art results for detecting hate speech in various languages, including Hindi.

In [7], The effectiveness of the BERT-CNN model is demonstrated through experiments on a large movie review dataset, where the model achieves an accuracy of 86.67%. This performance significantly outperforms traditional text classification models like textCNN, highlighting the model's potential for better understanding and classifying the sentiment of textual data. The results indicate that the BERT-CNN approach holds promise for improving sentiment classification, showcasing its strong performance in capturing deeper semantic insights and offering more accurate sentiment predictions compared to conventional methods. While the BERT-CNN model demonstrates strong performance in sentiment classification, there are several limitations. One of the main challenges is the high computational cost associated with using BERT, which requires significant memory and processing power, especially when working with large datasets. This can limit the scalability of the model, particularly for applications with constrained resources. Additionally, the model's dependency on BERT's pre-trained embeddings may introduce biases from the pretraining data, potentially affecting its performance on specific domains or less-represented languages.

Table 1: Summarization of various authors

Authors	Title	Research Focus	Observations
Areej Alhothali, and Kawthar Moria (2023)	Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model	Utilizes BERT and domain-specific word embeddings with Bidirectional LSTMs for hate speech detection, emphasizing semantic understanding and task-specific fine-tuning.	Demonstrated an 8% accuracy improvement with domain-specific embeddings, outperforming traditional models in semantic nuance detection..
K. Mnassri, P. Rajapaksha, R. Farahbakhsh, N. Crespi (2022)	BERT-based Ensemble Approaches for Hate Speech Detection	Combines BERT with CNNs, LSTMs, and MLPs using ensemble techniques like voting and stacking to improve Twitter-based hate speech detection.	Achieved high robustness and classification accuracy, with F1 scores reaching 97% on the Davidson dataset.

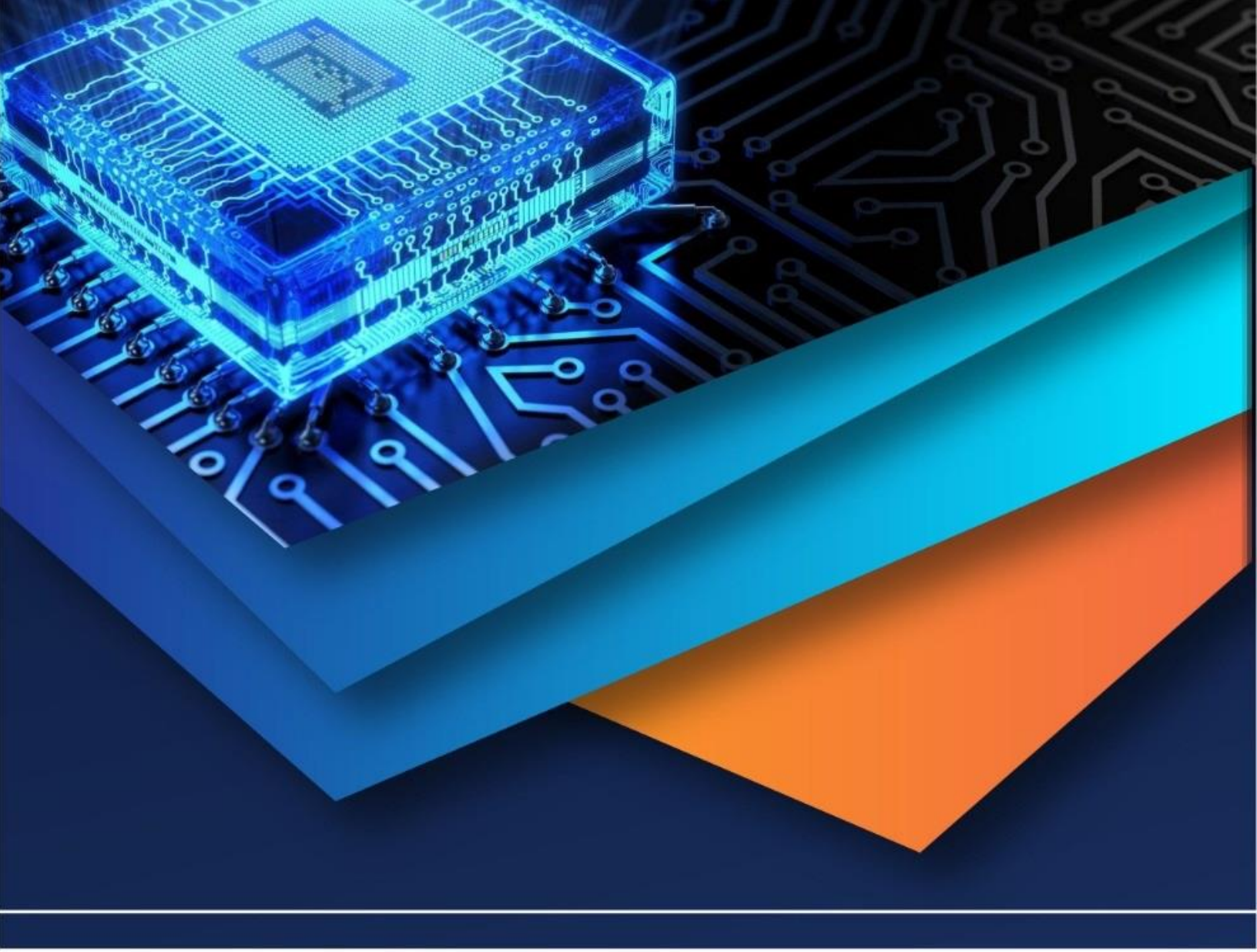
Subrata Saha, Md. Imran Hossain Showrov, Md. Rahman, Md. Majumder (2022)	VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Coronavirus Outbreak	Compares VADER and BERT for sentiment analysis on COVID-19 and Omicron-related tweets, leveraging supervised ML algorithms..	BERT consistently enhanced sentiment classification performance, with SVM achieving 92% accuracy on the Omicron dataset.
Nayla Adine, Casi Setianingsih, Burhanuddin Dirgantoro (2023)	Hate Speech Detection on Twitter Using BERT Algorithm	Implements a website-based hate speech detection tool using BERT, analyzing text input for classification accuracy and accessibility.	BERT-based detection achieved 78.69% accuracy with reliable performance metrics, providing an accessible solution for identifying hate speech.
U. Mittal (2023)	Detecting Hate Speech Utilizing Deep Convolutional Network and Transformer Models	Evaluates CNN, BiLSTM, BERT, and fine-tuned RoBERTa for classifying tweets into hate speech, offensive language, and neutral categories.	Fine-tuned RoBERTa achieved the best performance, consistently scoring high across metrics like accuracy, F1-score, and MCC.
Avishkar Gautam, Ayush Singh, Ayush Verma, Dr. Jaya Sinha (2023)	Hate Speech Detection Using Deep Learning	Investigates multilingual hate speech detection using deep learning models like CNNs, BiLSTMs, and fine-tuned transformers such as BERT and RoBERTa.	Demonstrated significant improvements in detection for code-mixed languages like Hinglish, highlighting the effectiveness of deep learning approaches in diverse linguistic settings.
R. Man, K. Lin (2021)	Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network	Combines BERT with CNN for sentiment analysis, leveraging pre-trained embeddings to enhance semantic understanding in textual data.	Achieved 86.67% accuracy on a large movie review dataset, outperforming traditional classification models with deeper semantic insights.

III. CONCLUSION

The growing prevalence of hate speech and racism on social media platforms underscores the critical need for advanced detection systems. This paper presents a comprehensive approach to racism detection using sentiment analysis, leveraging a BERT-based model for its superior ability to understand contextual and implicit nuances in textual data. By implementing a user-friendly Streamlit-based frontend, the system ensures accessibility and usability for a broad audience. The proposed system demonstrates significant potential for identifying racism in tweets, paving the way for real-world applications in monitoring and moderating harmful online content. Future research can build upon this foundation to further enhance the accuracy and effectiveness of racism detection systems.

REFERENCES

[1] Areej Alhothali, and Kawthar Moria, “Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model,” Applied Artificial Intelligence, 37(1), vol. 5, no. 2, pp. 112-123, 2023.
[2] K. Mnassri, P. Rajapaksha, R. Farahbakhsh and N. Crespi, “BERT-based Ensemble Approaches for Hate Speech Detection,” GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 4649-4654
[3] Saha, Subrata & Showrov, Md. Imran Hossain & Rahman, Md & Majumder, “VADER vs. BERT: A Comparative Performance Analysis for Sentiment on Coronavirus Outbreak,” Machine Intelligence and Emerging Technologies - First International Conference, MIET 2022, Proceedings (pp. 371-385)
[4] Nayla, Adine & Setianingsih, Casi & Dirgantoro, Burhanuddin, “Hate Speech Detection on Twitter Using BERT Algorithm,” International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), 2023, pp. 644-649
[5] U. Mittal, “Detecting Hate Speech Utilizing Deep Convolutional Network and Transformer Models,” 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM), Roorkee, India, 2023, pp. 1-4
[6] Avishkar Gautam, Ayush Singh, Ayush Verma, Dr. Jaya Sinha, “Hate Speech Detection Using Deep Learning,” ID: IJRASET61475.
[7] R. Man and K. Lin, “Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network”, 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2021, pp. 769-772.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY