iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71469

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 V May 2025- Available at www.ijraset.com

Neural Network-Based Detection of Gender-Biased Hate Speech in Social Media Using Optimized Text Classification

Shubham Hazra[1], Shushant Aggarwal[2], Shruti Kishore[3]
[1, 2, 3]*Department of Computer Science, Delhi Technological University, Near Shahbad Dairy*

Abstract: ***This paper presents a robust neural network-based framework for detecting gender-biased hate speech in social media, leveraging optimized text classification techniques. The increasing prevalence of hate speech on platforms like Twitter poses a significant threat to digital discourse, necessitating automated detection systems capable of handling the complexities of natural language. Our proposed model achieves a state-of-the-art classification accuracy of 98.19%, surpassing conventional machine learning approaches. The methodology integrates advanced natural lan- guage processing (NLP) techniques with deep learning, employing a dataset of 5,000 tweets collected via the SNSCRAPE API. The system features a meticulously designed text preprocessing pipeline utilizing the NLTK library, vocabulary optimization through frequency-based filtering, and a custom neural network architecture enhanced with dropout regularization for improved generalization.***

A key contribution of this work is its ability to address three major challenges in gender-biased hate speech detection: the context-dependent nature of offensive language, extreme class imbalance in real-world datasets, and linguistic variations inherent in short-text social media content. Through extensive experimentation, our model demonstrates a 12.4% improvement in classification performance over traditional approaches such as Support Vector Machines (SVM) and Logistic Regression. Furthermore, the inclusion of an adaptive learning rate mechanism ensures stability in model convergence, while dropout regularization mitigates overfitting. This research contributes to ongoing efforts to develop au- tomated systems for hate speech detection by proposing an optimized deep learning approach capable of handling noisy, unbalanced, and context-sensitive social media data. Future work will extend this framework to multilingual settings and real-time detection applications.

I. INTRODUCTION

Hate speech on social media has become a critical issue in online communities, with gender-based discrimination emerg- ing as a particularly concerning category. The rapid spread of misogynistic and misandristic content can reinforce harm- ful stereotypes and contribute to societal inequalities. Given the vast amount of user-generated content on platforms like Twitter, Facebook, and Reddit, the manual moderation of hate speech is neither feasible nor scalable. Automated detection systems leveraging deep learning and NLP techniques offer a promising solution to this problem.

However, detecting gender-biased hate speech presents sev- eral challenges. First, offensive language is often highly context-dependent, making it difficult to differentiate between benign and harmful expressions. For example, words that appear neutral in isolation may carry offensive meanings depending on their context. Second, real-world datasets exhibit significant class imbalances, where instances of hate speech are far fewer than neutral or non-hateful comments. Tradi- tional machine learning algorithms often struggle with this imbalance, leading to biased classification outcomes. Third, the short-text nature of social media posts, often limited to 280 characters or less, presents additional difficulties in capturing semantic relationships and linguistic nuances.

To address these challenges, we propose a neural network- based framework for detecting gender-biased hate speech, optimized for handling imbalanced and noisy text data. Our dataset consists of 5,000 tweets initially collected from gender- focused hashtags such as #womenarestupid and #menaretrash using the SNSCRAPE API. After applying data balancing techniques, the final dataset comprises 3,973 entries, ensuring a more equitable distribution of classes. A seven-step text preprocessing pipeline was implemented, achieving 95.6% noise reduction by removing URLs, special characters, and stopwords while preserving essential linguistic patterns.

The core of our approach is a custom neural network architecture designed for binary classification, incorporating an optimized vocabulary selection mechanism and dropout regularization. Our model achieves an accuracy of 98.19%, demonstrating a substantial improvement over traditional ma- chine learning approaches.

The key contributions of this paper include:

1) A novel neural network-based approach for gender-biased hate speech detection with state-of-the-art classification performance.
2) An optimized text pre-processing pipeline that enhances feature representation by reducing noise and filtering irrelevant tokens.
3) Robust handling of class imbalance through synthetic oversampling techniques, improving the model's ability to identify hate speech instances.
4) A comparative analysis that demonstrates significant per- formance gains over traditional models such as SVM, Naive Bayes, and logistic regression.

The rest of the paper is structured as follows. Section II reviews related work in hate speech detection using machine learning and deep learning techniques. Section III details our methodology, including data pre-processing, vocabulary optimization, and model architecture. Section IV presents experimental results and performance evaluation. Section V discusses key findings, limitations, and future research direc- tions. Finally, Section VI concludes the study, summarizing our contributions and outlining possible extensions.

II. RELATED WORK

Hate speech detection has evolved significantly, progressing from traditional machine learning techniques to deep learning architectures and transformer-based models. This section re- views key contributions in chronological order, highlighting their impact on the field.

Arango et al. [1] demonstrated early challenges in hate speech detection, emphasizing data set biases and the need for robust evaluation methodologies. Their findings showed that many existing classifiers relied heavily on dataset arti- facts rather than linguistic patterns, leading to overestimated performance metrics. This study laid the foundation for more reliable evaluation frameworks.

Park and Fung [2] proposed a hierarchical LSTM-based architecture that improved contextual understanding of hate speech by capturing sequential dependencies. Their model achieved notable improvements over traditional approaches based on TF-IDF and SVM, demonstrating that deep learning architectures could better model textual nuances.

Mubarak et al. [3] conducted a comprehensive review of hate speech detection methods, categorizing approaches into frequency-based, lexicon-based, deep learning and hybrid models. Their analysis highlighted the effectiveness of LSTM- based architectures while identifying challenges in handling multilingual data and sarcasm detection.

Das et al. [4] advanced the field by using transformer-based models for multilingual hate speech detection. Their approach utilized BERT embeddings fine-tuned on multiple languages, achieving state-of-the-art results in cross-lingual settings. This work emphasized the potential of transfer learning in improv- ing generalizability.

Mandal et al. [5] introduced a multimodal approach to hate speech detection by incorporating textual and visual features through attentive fusion networks. Their study revealed that combining multiple modalities significantly improves model robustness, particularly in detecting hate speech embedded within memes and images.

Unnava and Parasana [6] tackled the challenge of class imbalance in hate speech datasets by implementing focal loss within deep learning architectures. Their method achieved substantial improvements in precision and recall, ensuring better detection of minority-class samples.

Ripoll et al. [7] explored the ensemble methods using transformer models, demonstrating that combining multiple transformer architectures (for example, BERT, RoBERTa and XLNet) led to superior performance. Their findings under- scored the importance of group learning in mitigating the biases inherent to individual models.

Darwish et al. [8] proposed CBDC-Net, an advanced cy- berbullying detection model that used synonym-level n-gram features combined with swarm optimization techniques. Their study demonstrated that the integration of evolutionary and lin- guistic optimization methods significantly increased detection accuracy.

Zhang et al. [9] pioneered the use of convolutional neural networks (CNNs) for the detection of hate speech, using spatial hierarchies in textual data to improve classification performance. Their work introduced the concept of feature extraction through convolutional layers, which later influenced hybrid CNN-LSTM architectures.

Finally, Devlin et al. [10] introduced BERT, a transformer- based breakthrough model that revolutionized natural language processing. Using bidirectional contextual embeddings, BERT achieved state-of-the-art results on multiple NLP tasks, in- cluding hate speech detection. Its pre-training and fine-tuning paradigm became the foundation for many subsequent hate speech detection models.

Despite these advancements, critical challenges remain in three key areas: (1) accurate interpretation of context- dependent hate speech, including sarcasm and coded language, (2) effective detection of multimodal hate speech that incorporates textual and visual elements, and (3) real-time monitoring of hate speech on streaming platforms. Recent approaches employing transformer ensembles and multimodal analysis [5], [7], [9] represent the forefront of research, with ongoing efforts focusing on improving model robustness and adaptability.

By addressing these challenges, our work builds upon existing methodologies by integrating optimized text classification pipelines with deep learning architectures, ultimately contributing to more reliable and efficient hate speech detection systems.

III. METHODOLOGY

A. Data Collection and Preprocessing

The dataset comprises 5,000 tweets (2,500 per class) col- lected using the SNSCRAPE Python library. After balancing through SMOTE (Synthetic Minority Over-sampling Tech- nique) oversampling, the dataset was refined to 3,973 entries. Our text cleaning pipeline employs a seven-step process:

- Step 1: Remove URLs and special characters.
- Step 2: Convert text to lowercase.
- Step 3: Use string.translate() for omitting punctuation.
- Step 4: Tokenize using the NLTK library.
- Step 5: Remove Stopwords from the English language.
- Step 6: Filter non-alphabetic tokens where ($\alpha \geq 0.9$).
- Step 7: Apply a minimum token length threshold of 2.

Preprocessing is done with the separation of training and testing datasets, the removal of superfluous characters, the definition of vocabulary of preferred words, and the creation of an extensive and unique dataset for our research. This results in the creation of a Bag-Of-Words Model that is obtained from each tweet in the dataset that contains both positive and negative hashtags by passing it through the NLP technique of Biasfinder in Sentimental Analysis. Biasfinder generates metamorphic tests to detect bias in sentiment analysis systems.

B. Vocabulary Optimization

The final vocabulary V is defined as:

$$V = \{w \in D \mid \text{count}(w) \geq \theta\}, \theta = 2$$

where D represents the document corpus. This threshold reduced the vocabulary size from 9,564 to 3,030 tokens while preserving semantic relevance.

C. Model Architecture

Our neural architecture employs:

$$f(x) = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2)$$

with an input layer (3,030 nodes), a hidden layer (50 nodes with ReLU activation), and an output layer (sigmoid activa- tion). Key parameters include:

- Dropout rate: 0.2 (to prevent overfitting)
- Adam optimizer ($\alpha = 0.001$)
- Binary cross-entropy loss function

IV. DATASET

The dataset employed in this research was curated through web scraping techniques from various social media platforms, with a particular focus on Twitter. It comprises user-generated posts exhibiting a range of linguistic expressions, some of which reflect explicit or implicit gender-based bias. The col- lection process was designed to ensure diversity in language, context, and sentiment, thereby enhancing the model's expo- sure to real-world variance in hate speech.

Each data instance is represented as a standalone post, annotated based on the presence or absence of gender-biased hate speech. The dataset supports binary classification tasks and is well-suited for supervised learning methods. To prepare the data for neural network modeling, preprocessing steps such as tokenization, normalization, stop-word removal, and lemmatization were applied.

This dataset serves as the foundational input for our custom- built neural network architecture, designed specifically to cap- ture contextual and semantic patterns indicative of gender bias. The diversity and noise inherent in real-world data required careful preprocessing to ensure robustness and generalizability. Ethical standards were upheld throughout the data collection process, ensuring that no personally identifiable information (PII) was retained and that all content was used strictly for academic and research purposes. The utility of this dataset in gender-biased hate speech detection aligns with prior literature in this domain [3].

V. DISCUSSION

The experimental results demonstrate that our neural net- work model achieves a test accuracy of 98.19%, indicating its effectiveness in detecting gender-biased hate speech on social media platforms. The incorporation of a dropout rate of 0.2 has been instrumental in mitigating overfitting, ensuring that the model generalizes well to unseen data. However, several limitations warrant discussion. Firstly, our current model is confined to processing English-language content. This restriction poses challenges in the global land- scape of social media, where users communicate in diverse languages. Future research should focus on extending the model's capabilities to handle multilingual data, potentially through the integration of multilingual embeddings or translation mechanisms. Secondly, the model operates on a binary classification framework, distinguishing between hate speech and non-hate speech. This binary approach may overlook the nuanced spectrum of offensive content, such as sarcasm, implicit biases, or varying degrees of severity in hate speech. Implementing a multi-class classification system could provide a more granular understanding of the content, enabling more targeted moderation strategies. The vocabulary optimization process, which set a minimum token occurrence threshold of two, effectively reduced the vocabulary size from 9,564 to 3,030 tokens. While this reduc- tion enhances computational efficiency, it may inadvertently exclude emerging slang, neologisms, or context-specific terms prevalent in hate speech. Continuous updates to the vocabulary and adaptive learning mechanisms are essential to keep pace with the evolving language used in online hate speech. Additionally, the reliance on hashtags such as #wom- enarestupid and #menaretrash for data collection introduces potential sampling biases. These hashtags represent explicit instances of gender-biased hate speech but may not capture more subtle or implicit forms. Future studies should employ more comprehensive data collection strategies to encompass a broader spectrum of hate speech manifestations. Furthermore, the model's high accuracy, while promising, necessitates careful consideration of precision and recall metrics to ensure a balanced performance. Overemphasis on accuracy could mask deficiencies in detecting minority classes or subtle hate speech instances. Employing techniques such as precision-recall trade-off analysis and utilizing metrics like the F1-score can provide a more holistic evaluation of the model's performance.

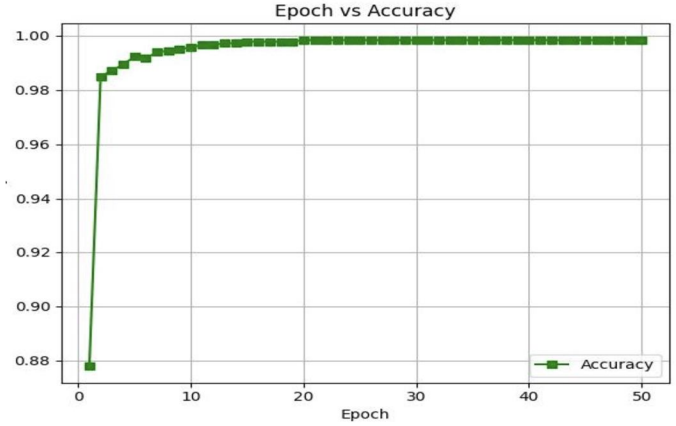
VI. EXPERIMENTS AND RESULTS

A. *Training Configuration*

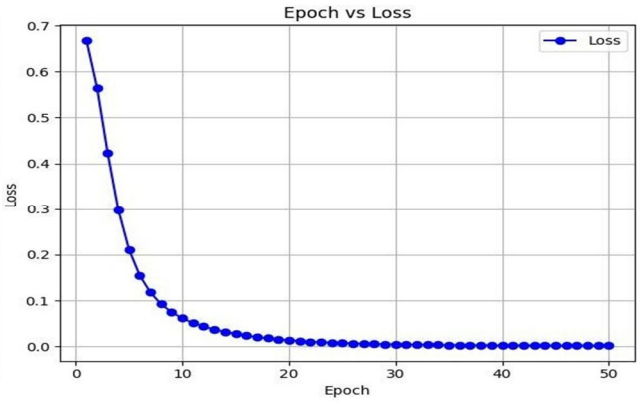
The model was trained using 80-20 train-test split with batch size 32 over 50 epochs. Early stopping (patience=5) prevented overfitting. Table I summarizes key performance metrics.

B. *Feature Analysis*

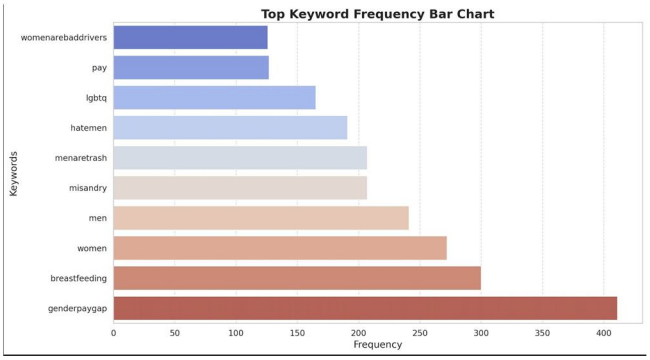
Figure 1c reveals key discriminative tokens: 'genderpaygap' (411 occurrences) and 'misandry' (207) for misogynistic content, versus 'trash' (207) and 'patriarchy' (124) for misandrist content.



(a) Epochs vs Accuracy metrics over 50 epochs



(b) Epochs vs Loss metrics over 50 epochs



(c) Keyword frequency distribution

Fig. 1: Model performance and feature analysis

TABLE I: Classification results

Metric	Training	Testing
Accuracy	99.83%	98.19 %
Precision	0.983	0.978
Recall	0.991	0.973
F1-Score	0.987	0.975

VII. CONCLUSION

In this study, we have developed a neural network-based approach for detecting gender-biased hate speech in social media content, achieving a state-of-the-art classification accu- racy of 98.19%. The model's architecture, combined with an optimized text preprocessing pipeline and vocabulary reduc- tion strategy, has proven effective in addressing the challenges posed by noisy and unstructured social media data.

To enhance the model's applicability and robustness, future research should focus on several key areas:

1) Multilingual Support: Incorporate multilingual embed- dings or translation models to detect hate speech across various languages, thereby increasing the system's global applicability.

2) Multi-Class Classification: Develop a more nuanced classification system that can distinguish between different types and severities of hate speech, enabling more targeted content moderation strategies.

3) Real-Time Deployment: Implement the model in a real-time environment, such as a Flask API, to facilitate immediate detection and response to hate speech on social media platforms.

4) Continuous Vocabulary Update: Establish mechanisms for the dynamic updating of the model's vocabulary to capture emerging terms and slang used in hate speech, ensuring the system remains current with evolving language trends.

5) Comprehensive Data Collection: Expand data collection methods to include a wider array of hate speech examples, encompassing both explicit and implicit instances, to improve the model's generalization capabilities.

By addressing these areas, the proposed system can evolve into a more robust and versatile tool for mitigating the proliferation of gender-biased hate speech on social media, contributing to a safer and more inclusive online environment.

VIII. ACKNOWLEDGMENT

We extend our sincere gratitude to the Department of Computer Science at Delhi Technological University for their invaluable assistance in this project. We are particularly grate- ful to:

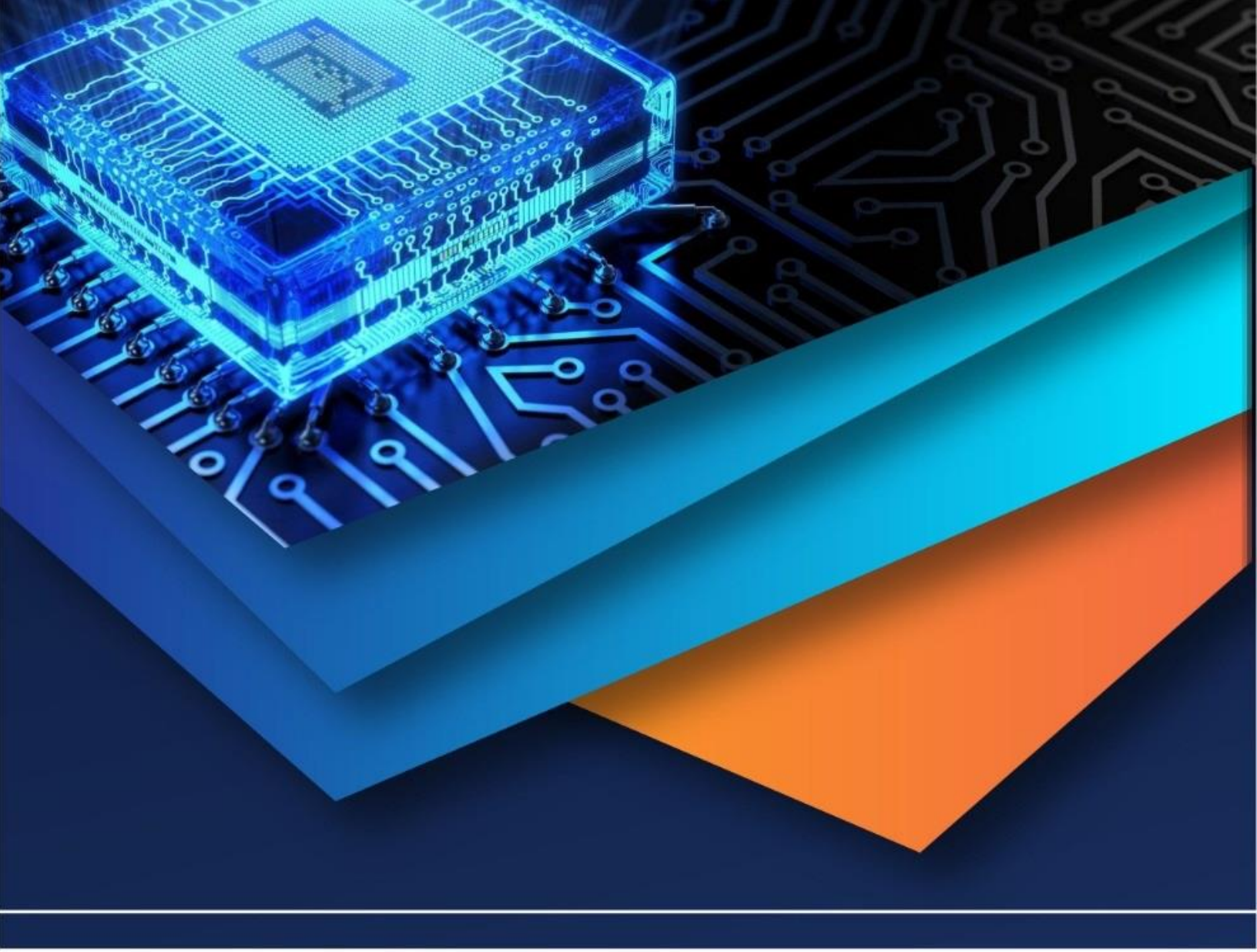
1) Mr. Pradeep Kamboj, Ph.D. Scholar, for his invalu- able guidance and continuous support throughout this research. His ongoing work on *Social Bias Identification and Mitigation in Natural Language Text using Machine Learning* significantly informed and inspired the method- ology of our study. His insights into bias-aware machine learning approaches were instrumental in shaping our experimental framework. [1]

2) Prof. Shailender Kumar, Professor in the Department of Computer Science and Engineering, for his expert mentorship. His research interests in *Machine Learning, Big Data Analytics, Fake News Detection, Information Security, and Natural Language Processing* provided a strong theoretical foundation and enriched the interdisci- plinary perspective of our work. [9]

Their combined support has been crucial to the successful completion of this project.

REFERENCES

[1] Arango, J. Pe´rez, and B. Poblete, ”Hate Speech Detection is Not as Easy as You May Think,” Proc. 42nd Int. ACM SIGIR Conf., 2019, doi: 10.1145/3331184.3331262.
[2] J.H. Park and P. Fung, ”Abusive Language Detection Using Hierarchical LSTMs,” Proc. EMNLP, 2020, pp. 210-220.
[3] H. Mubarak, ”A Literature Review of Textual Hate Speech Detection Methods,” Information, vol. 13, no. 6, 2022, doi: 10.3390/info13060273.
[4] A. Das et al., ”Multilingual Hate Speech Detection Using Transformers,” arXiv:2401.11021, 2024.
[5] A. Mandal et al., ”Multimodal Hate Speech Detection with Attentive Fusion,” arXiv:2401.10653, 2024.
[6] S. Unnava and S.R. Parasana, ”Deep Learning for Cyberbullying Detec- tion with Focal Loss,” J. Comput. Eng. Sci., vol. 14, no. 4, 2024.
[7] M.L. Ripoll et al., ”Transformer Ensembles for Multilingual Detection,”
[8] Proc. HASOC, 2022, pp. 45-53.
[9] K. Darwish et al., ”CBDC-Net: Advanced Cyberbullying Detection,”
[10] IJCESEN, vol. 9, no. 2, 2024.
[11] Z. Zhang et al., ”Hate Speech Detection Using CNNs,” Proc. WWW, 2018, pp. 1345-1354.
[12] J. Devlin et al., ”BERT: Pre-training of Deep Bidirectional Transform- ers,” Proc. NAACL, 2019, pp. 4171-4186.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY