iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82595

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 14 Issue V May 2026- Available at www.ijraset.com

Design and Implementation of a Comprehensive AI-Driven NLP Framework for Classifying and Preventing Hate Speech in Social Media

V Vani Tejaswini[1], Dr. K Srinivas[2], G Phani Durga Anusha[3]
[1, 2, 3]*Computer Science Engineering, Bonam Venkata Chalamayya Institue of Technology and Science*

Abstract: ***The rapid growth of social media has increased the spread of hate speech, creating challenges for online safety and communication. This research proposes an Artificial Intelligence (AI) and Natural Language Processing (NLP) based system to automatically detect and classify hate speech in social media posts. The system uses deep learning and transformer-based models to improve detection accuracy. It also applies text preprocessing techniques to handle informal social media language. The proposed model can identify different types of hate speech such as racism, sexism, religious hatred, and cyberbullying. Experimental results show better accuracy and performance compared to traditional machine learning methods for hate speech detection.***

Keywords: Hate Speech Detection, Artificial Intelligence, Natural Language Processing, Deep Learning

I. INTRODUCTION

Social media platforms such as Twitter, Facebook, Instagram, and Reddit have become major communication channels for sharing information and opinions. Although these platforms improve global connectivity, they also increase the spread of harmful online content including hate speech, abusive language, and cyberbullying. Hate speech targets individuals or communities based on religion, gender, race, ethnicity, nationality, or social identity, negatively affecting social harmony and user safety. Traditional moderation systems mainly rely on manual monitoring and keyword-based filtering approaches. However, these methods are inefficient because they cannot process the massive volume of data generated every second on social media platforms. In addition, hate speech often appears in implicit, sarcastic, or context-dependent forms, making detection difficult for conventional rule-based systems.

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have provided more effective approaches for analyzing textual data and detecting harmful online behavior. NLP techniques enable computers to understand semantic relationships and contextual meaning in language, while machine learning algorithms identify linguistic patterns and automatically classify harmful content. This paper proposes an AI-driven NLP framework for detecting hate speech in social media text. The framework incorporates text preprocessing techniques, TF-IDF feature extraction, and machine learning classifiers such as Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest to classify text into hate speech, offensive language, and neutral content categories with improved accuracy and efficiency.

II. RELATED WORK

AI-based hate speech detection has become an important research area in Natural Language Processing (NLP) and machine learning. Researchers have applied various techniques such as SVM, Naive Bayes, Logistic Regression, Random Forest, and transformer-based models like BERT, LLaMA 2, and XLM-R to improve hate speech classification. Recent studies show that transformer and emotion-aware models provide better contextual understanding, multilingual support, and higher classification accuracy compared to traditional methods. These approaches demonstrate that AI and NLP techniques are highly effective for automatic hate speech detection on social media platforms.

III.EXISTING SYSTEM

Existing hate speech detection systems mainly focus on keyword-based filtering, manual moderation, and traditional machine learning approaches to identify harmful content on social media platforms. These systems use conventional Natural Language Processing (NLP) techniques and basic classification algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Decision Trees for detecting offensive or abusive text.

Rule-based systems analyse predefined offensive keywords, while machine learning-based systems classify text using handcrafted features such as TF-IDF and Bag-of-Words (BoW). Manual moderation systems rely on human reviewers to identify and remove harmful content from online platforms.

Although these systems provide moderate detection performance, they have several limitations:

1) They lack contextual understanding of language.
2) Accuracy decreases for implicit or sarcastic hate speech.
3) Traditional models produce high false positive and false negative rates.
4) Informal social media language is difficult to process effectively.
5) Real-time large-scale hate speech detection is not efficiently supported.

Due to these limitations, an advanced AI and deep learning-based hate speech detection framework is required to improve accuracy, contextual understanding, and real-time performance.

IV. PROPOSED METHODOLOGY

The proposed system is designed to detect and classify hate speech in social media content using Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques. Initially, textual data is collected from various social media platforms and preprocessed by removing noise, stop words, and unwanted characters, followed by tokenization and lemmatization to improve data quality.

The processed data is then transformed using feature extraction techniques such as word embeddings and contextual representations. Advanced deep learning and transformer-based models such as BERT are used to analyze linguistic patterns and contextual information for accurate hate speech detection. The trained model classifies text into categories such as hate speech, offensive language, and neutral content.

The proposed framework improves detection accuracy, enhances contextual understanding, and efficiently identifies both explicit and implicit hate speech expressions. It also provides a scalable and efficient solution for monitoring harmful content in large-scale social media environments.

V. SYSTEM ARCHITECTURE

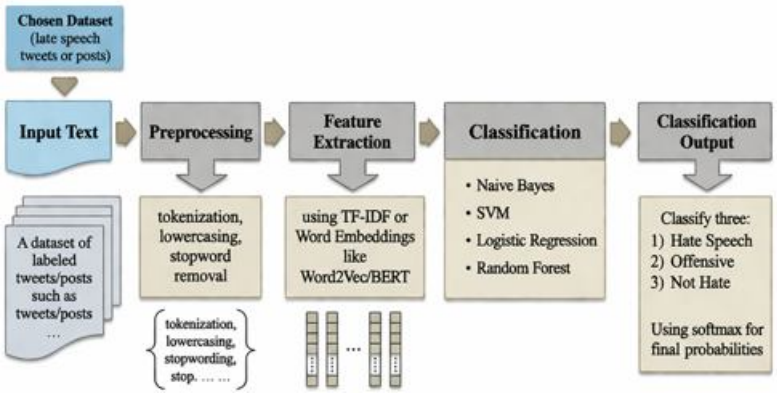


Fig: System Architecture

The system consists of the following layers:

1) Dataset Layer – Stores labeled social media datasets containing hate speech, offensive language, and neutral text for model training and evaluation.
2) Input Layer – Accepts textual data such as tweets, posts, and comments collected from social media platforms.
3) Preprocessing Layer – Performs text cleaning operations including tokenization, lowercasing, stop word removal, and normalization to improve text quality.
4) Feature Extraction Layer – Extracts meaningful textual features using techniques such as TF-IDF, Word Embeddings, Word2Vec, and BERT representations.
5) Machine Learning Layer – Utilizes classification algorithms such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Random Forest for hate speech detection.

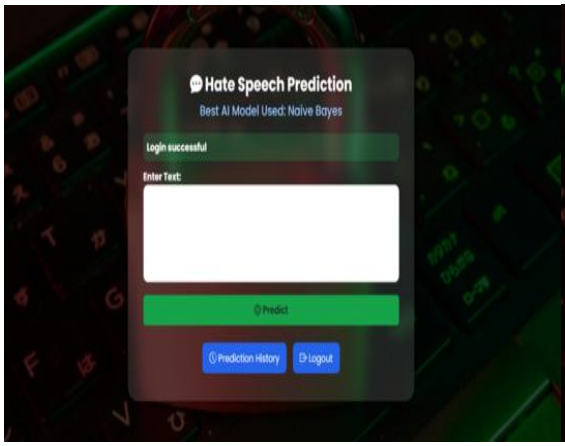
6) Classification Layer – Classifies the input text into categories such as Hate Speech, Offensive Content, or Non-Hate Content.
7) Probability Analysis Layer – Uses Softmax or similar probability functions to determine the final prediction confidence.
8) Output Layer – Displays the final classification result and prediction status for the given social media text.

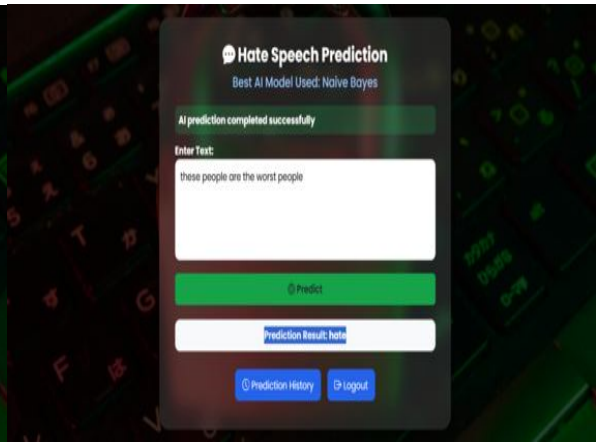
VI. IMPLEMENTATION

The Hate Speech Detection System is implemented using Python and advanced Natural Language Processing (NLP) and machine learning techniques to identify harmful and offensive content in social media text. The collected textual dataset is preprocessed using Pandas and NLP libraries, where noise such as URLs, special characters, and stop words are removed, and text normalization techniques such as tokenization and lemmatization are applied to improve data quality. Feature extraction methods such as TF-IDF, word embeddings, and transformer-based contextual representations are used to convert textual data into meaningful numerical features for model training. Machine learning and deep learning algorithms such as Naive Bayes, Support Vector Machine (SVM), and BERT are implemented and evaluated, with transformer-based models achieving higher classification accuracy and better contextual understanding. The system is integrated with the Streamlit framework to provide a scalable and user-friendly web application that allows users to input social media text and obtain instant hate speech predictions. The trained models and preprocessing components are stored for efficient deployment, while prediction results are maintained for analysis, monitoring, and future improvement of the hate speech detection framework.

VII. RESULTS AND DISCUSSION

The proposed Hate Speech Detection System was developed using the Naive Bayes machine learning algorithm to classify hateful and non-hateful text content. The system was trained using textual datasets containing various online comments and messages. Experimental results showed that the model achieved accurate and fast prediction performance for real-time text analysis. The application was successfully integrated into a web-based platform with secure login, text prediction, and prediction history features. Users can enter text and instantly receive prediction results indicating whether the content contains hate speech. The system effectively identified offensive statements and displayed accurate classification results. In addition, the Naive Bayes model provided efficient processing with low computational complexity. The obtained results confirm that the proposed system is reliable, efficient, and suitable for detecting harmful online content and promoting safer digital communication.





VIII. CONCLUSION

This paper presented a comprehensive AI-driven NLP framework for hate speech detection in social media. The proposed system uses text preprocessing, TF-IDF feature extraction, and machine learning algorithms including Naive Bayes, Logistic Regression, SVM, and Random Forest for classification.

The framework effectively identifies hate speech, offensive language, and neutral content with improved contextual understanding and higher accuracy compared to traditional moderation systems. The proposed solution supports automated moderation and contributes to safer digital communication platforms.

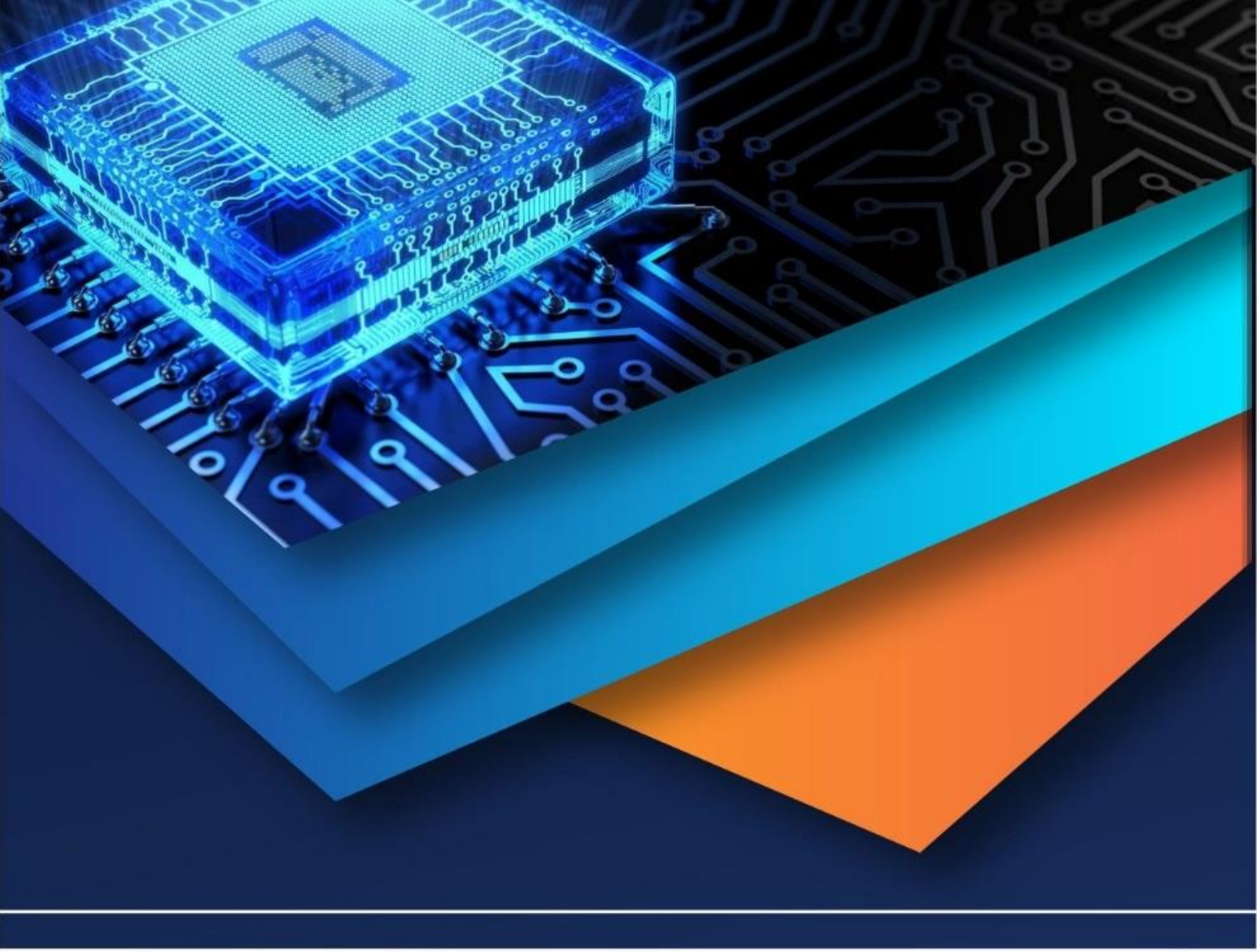
IX. FUTURE WORK

The proposed Hate Speech Detection System can be further enhanced in several ways:

1) Support real-time hate speech detection from live social media streams and online chat platforms.
2) Improve classification accuracy using advanced deep learning and transformer-based models such as LSTM, GRU, BERT, and RoBERTa.
3) Integrate advanced multimodal techniques for detecting hate speech from text, audio, and video content simultaneously.
4) Support multilingual hate speech detection for regional languages such as Telugu, Hindi, and other local languages.
5) Develop mobile and cloud-based deployment solutions for scalable and real-time monitoring applications.
6) Add sentiment analysis and emotion-aware hate speech classification for better contextual understanding.
7) Implement explainable AI techniques to visualize and interpret prediction results for improved transparency.
8) Enhance system scalability and performance for handling large-scale social media and multimedia datasets.
9) Improve robustness under noisy text, slang words, abbreviations, and mixed-language social media content.
10) Integrate hybrid feature extraction and multi-input models to improve prediction accuracy and overall system efficiency.

REFERENCES

[1] Davidson T. et al., “Automated Hate Speech Detection and the Problem of Offensive Language,” ICWSM, 2017.
[2] Devlin J. et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.
[3] Fortuna P., Nunes S., “A Survey on Automatic Detection of Hate Speech in Text,” ACM Computing Surveys, 2018.
[4] Asogwa D.C. et al., “Hate Speech Classification Using SVM and Naive Bayes,” IOSR Journal, 2022.
[5] Kumar S. et al., “Hate Speech Detection Using Logistic Regression and Deep Learning Hybrid Models,” IEEE Access, 2023.
[6] Sharma R. et al., “Hate Speech Classification Using Random Forest and Ensemble Learning Techniques,” Expert Systems with Applications, 2024.
[7] Gil Ramos et al., “Automatic Hate Speech Detection in the Age of the Transformer,” 2024.
[8] Md Saroar Jahan et al., “A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection,” 2025.
[9] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY