iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80361

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 IV Apr 2026- Available at www.ijraset.com

A Cross Platform Multilingual Moderation Framework for Real-Time Hate Speech Detection and Masking using Contextual NLP Techniques

Sairaj Alave[1], Tanish Bodekar[2], Sami Fodkar[3], Omkar Kadam[4], Prof. Nilesh Patil[5]
Information Technology Department, Saraswati College of Engineering, Kharghar, India

Abstract: ***In the digital era, social media platforms, online forums, and communication channels have become integral to human interaction. This project focuses on the design and development of an intelligent system capable of detecting and masking hate speech in online textual content using machine learning and natural language processing techniques. With the rapid growth of social media platforms and user-generated content, there has been a significant rise in the spread of offensive, abusive, and harmful language. Manual moderation is not only inefficient but also impractical due to the sheer volume of data generated every second. Therefore, this project aims to provide an automated, scalable, and efficient solution to identify and control hate speech in real time. The proposed system utilizes NLP techniques such as tokenization, stop-word removal, and vectorization (TF-IDF) to preprocess textual data. A supervised machine learning model is trained on labelled datasets to classify content into categories such as hate speech, offensive language, or neutral text. Once hate speech is detected, the system masks or replaces harmful words to ensure a safer user experience. The system is further integrated into a Chrome extension to provide real-time filtering of content on web pages. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score. This project contributes to building a more inclusive and secure digital environment.***

Keywords: Hate Speech Detection, Natural Language Processing (NLP), Machine Learning, Text Classification,Content Moderation, Chrome Extension, Real-time Filtering, Data Preprocessing, Artificial Intelligence, Web-based Application, Automated Moderation

I. INTRODUCTION

In recent years, the exponential growth of social media platforms such as Twitter, Facebook, Instagram, and YouTube has transformed the way people communicate and share information. While these platforms have enabled global connectivity and freedom of expression, they have also given rise to serious issues such as cyberbullying, harassment, and hate speech. Hate speech refers to any form of communication that demeans, insults, threatens, or targets individuals or groups based on attributes such as race, religion, gender, ethnicity, or nationality. The presence of such content not only harms individuals psychologically but also create a toxic online environment. Traditional methods of controlling such content rely heavily on manual moderation, which is both time-consuming and inconsistent. With millions of posts generated every minute, it becomes nearly impossible for human moderators to effectively monitor all content. This has led to the need for automated systems that can efficiently detect and filter harmful content. Machine learning and natural language processing provide powerful tools to address this challenge by enabling systems to understand and analyse human language.

This project aims to develop a hate speech detection and masking system that can automatically identify offensive content and take appropriate action. By integrating this system into a browser extension, users can experience real-time filtering of harmful content while browsing online platforms.

II. PROBLEM STATEMENT

The rapid increase in online communication has resulted in a significant rise in hate speech and offensive content. Existing moderation systems are either manual, inefficient, or lack real-time capabilities. Many automated systems fail to accurately detect contextual meaning, sarcasm, or variations in language. Therefore, there is a need for an intelligent, accurate, and real-time system that can detect and mask hate speech effectively.

The main aim of this project is to develop an intelligent system that can detect and mask hate speech in real-time using machine learning techniques. Design and implement a Machine Learning–based system that can automatically detect and mask hate speech from textual content to promote safe and inclusive digital communication.

III. LITERATURE SURVEY

Early approaches to hate speech detection relied on keyword-based filtering, where predefined lists of offensive words were used to identify harmful content. Although simple and easy to implement, these methods lacked the ability to understand context and often resulted in false positives and false negatives. With the advancement of machine learning, models such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression were introduced. These models improved accuracy by learning patterns from labelled datasets. However, they still struggled with understanding complex linguistic structures and context. Recent studies have focused on deep learning techniques such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and transformer-based models like BERT. These models are capable of capturing contextual information and semantic meaning, resulting in significantly higher accuracy. However, they require large datasets and computational resources.

1) Davidson et al. (2017)

Davidson and colleagues introduced one of the most influential datasets for hate speech detection in their paper "Automated Hate Speech Detection and the Problem of Offensive Language".They collected over 24,000 tweets and categorized them into hate speech, offensive but not hateful, and neutral. Using TF-IDF features and Logistic Regression, they demonstrated that traditional ML algorithms can effectively classify hate speech but also highlighted the challenge of distinguishing hate speech from general profanity. This study forms a foundational reference for many subsequent works.

2) Mozafari et al. (2020)

Mozafari and co-researchers proposed the use of BERT fine-tuning for hate speech detection, achieving state-of-the-art performance. Their study, titled "Hate Speech Detection Using BERT," highlighted how contextual embeddings help differentiate between offensive and neutral statements. They also stressed the necessity of addressing data imbalance and annotation quality in hate speech datasets.

3) Mishra et al. (2021)

In their work"Online Hate Speech Detection: A Survey and Research directions,"Mishra and colleagues reviewed over 80 research papers, identifying gaps such as limited multilingual support, lack of domain adaptation, and poor real-time deployment. They suggested integrating AI models into user applications to provide real-time hate speech filtering, a concept that directly aligns with the objectives of the current study.

4) Multimodal hate-speech detection (CASE / EACL 2024 and related).

Shared tasks and papers from 2024–2025 show meaningful progress on multimodal hate detection (text + images/video). Transformer-based fusion strategies and cross-modal attention architectures were shown to be effective in shared tasks, and several systems submitted to CASE/EACL 2024 use attentive fusion to detect hateful content that only appears when text and image are combined. This matters if you plan future work that includes screenshots, memes, or video comments.

5) Multilingual and low-resource approaches (MultiFED / NAACL 2024)

Newer work focuses on federated and transfer-learning solutions that generalize across languages and protect user data. For example, MultiFED uses continuous adaptation and federated training across many Indic-language datasets to improve low-resource detection while preserving privacy — an important advance for deploying hate-detection in linguistically diverse settings. If you plan multilingual support (e.g., Hindi, Marathi, Hinglish), these strategies are highly applicable.

IV. PROPOSED SYSTEM

The proposed system for hate speech detection and masking follows a structured methodology that involves multiple stages, starting from data collection to real-time deployment. The system is designed to efficiently process textual data, classify it using machine learning techniques, and take appropriate action based on the classification results.

The first phase of the methodology is **data collection**, where labelled datasets containing examples of hate speech, offensive language, and neutral text are gathered from reliable sources such as Kaggle. These datasets form the foundation for training the machine learning model.The next phase is **data preprocessing**, which is crucial for improving model performance. In this stage, the raw text data is cleaned by converting all characters to lowercase, removing punctuation, special characters, and stopwords. Tokenization is applied to break the text into individual words or tokens. This step ensures that the data is in a standardized format suitable for further processing.

Following preprocessing, **feature extraction** is performed to convert textual data into numerical form. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) are used to represent text as vectors. This allows the machine learning model to understand and process the input data effectively.The next step involves **model training**, where supervised machine learning algorithms such as Logistic Regression or Naive Bayes are used. The model learns patterns from the labelled dataset and builds a classification function that can distinguish between hate speech, offensive, and neutral content.

After training, the model undergoes **evaluation** using metrics such as accuracy, precision, recall, and F1-score. These metrics help in determining the effectiveness and reliability of the model. If the performance is not satisfactory, the model is fine-tuned by adjusting parameters or improving preprocessing techniques.

Once the model is finalized, it is integrated into a **real-time system** using a Chrome extension. The extension captures text from web pages and sends it to the trained model for analysis. If the text is classified as hate speech, a **masking mechanism** is applied, where offensive words are replaced with symbols (e.g., ****) or modified into safer alternatives. If the text is non-offensive, it is displayed as it is.

This methodology ensures that the system operates efficiently in real time while maintaining high accuracy in detecting harmful content. The combination of machine learning and browser integration makes the system practical and scalable for real-world applications.

The process begins with user input text such as comments or messages.

The input text undergoes data preprocessing, including cleaning, lowercasing, tokenization, and normalization.

The system performs language detection to identify the input language.

A check is performed to verify if the language is supported:

- If supported → proceed further
- If not supported → handle using error message or manual review

The cleaned text is converted into numerical form using **feature extraction techniques** such as TF-IDF or embeddings.

The processed data is passed to a **machine learning model** for prediction.

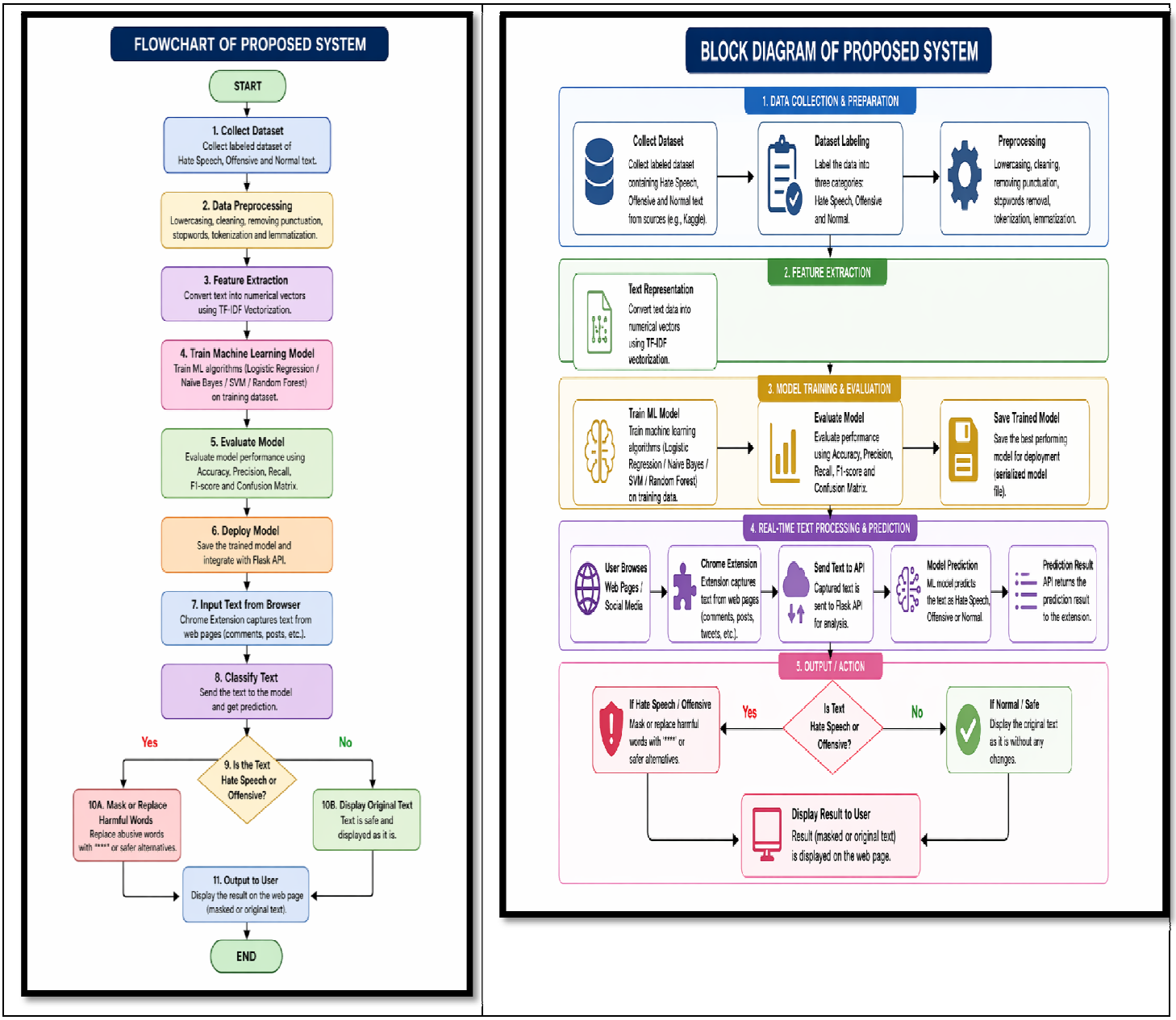
The model classifies the text as **hate speech or non-hate speech**.

If hate speech is detected:

- Offensive words are **masked or replaced**
- Content may be **flagged for manual review**

Flagged content is sent to a **moderation system** for further action if required.

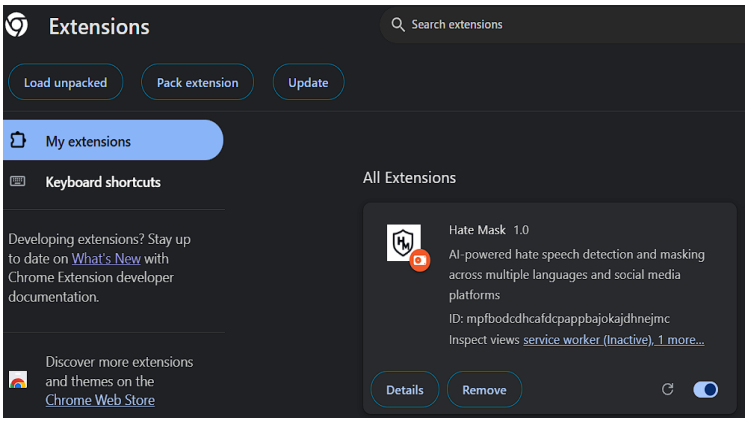
Finally, the system **displays the output text** (masked or original) to the user in real time.



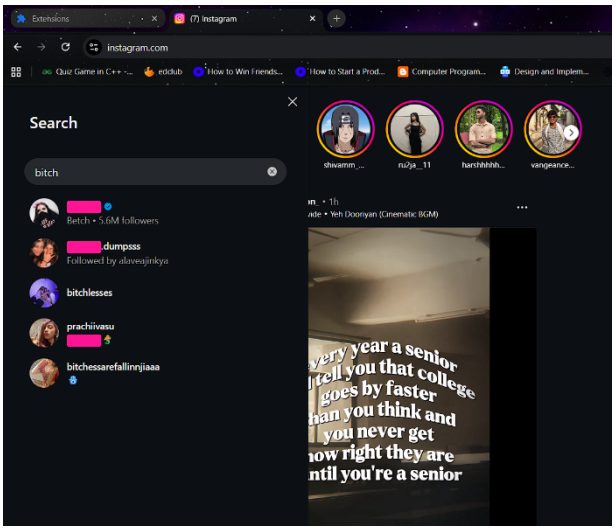
V. IMPLEMENTATION

The implementation of the Hate Speech Detection and Masking System is carried out through a systematic and structured approach that connects the theoretical concepts discussed in previous chapters with practical execution. This chapter focuses on the step-by-step development of the system, including data handling, model creation, and integration into a real-time environment.The implementation phase begins after understanding the problem statement and reviewing existing techniques for hate speech detection. Based on this understanding, an appropriate machine learning approach is selected. The system is designed to process textual data efficiently and provide accurate classification results in real time. The implementation also ensures that the system is scalable and can be enhanced in the future with advanced models and additional features. This chapter highlights how different components such as data preprocessing, model training, backend integration, and browser extension development are combined to form a complete working system.

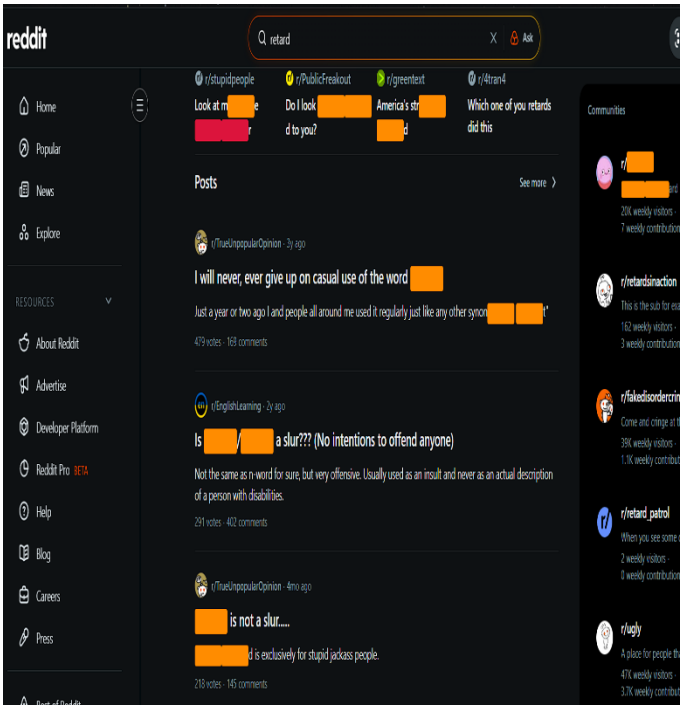
International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com



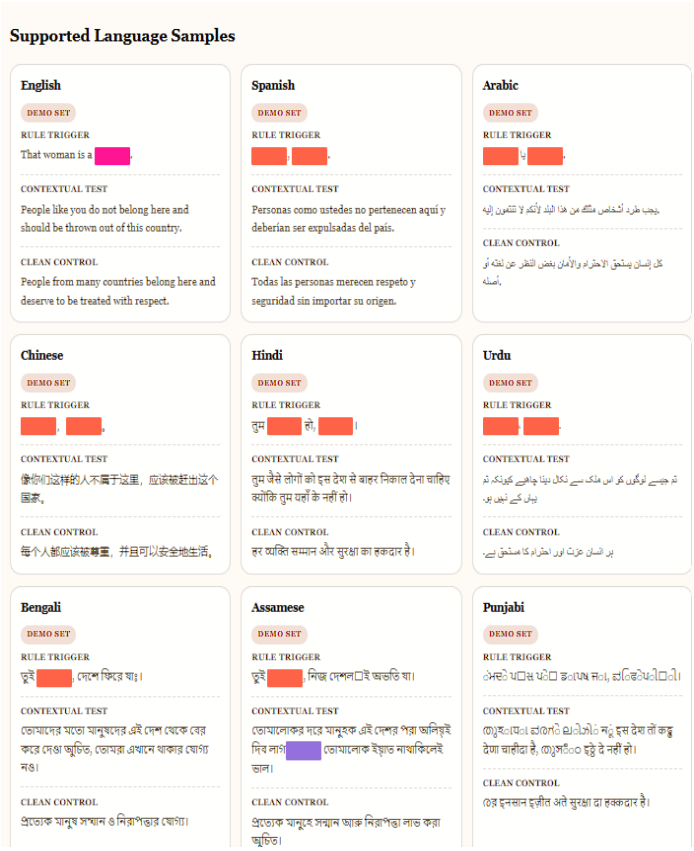
4.1 Chrome Extension



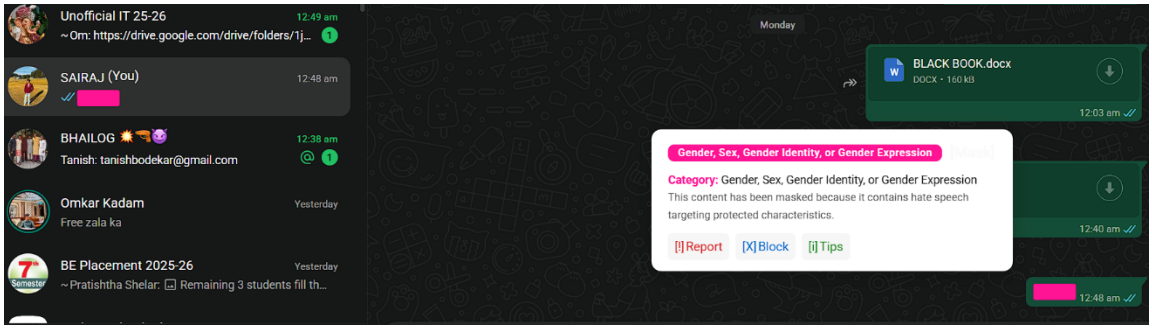
4.2 Hate Masking on Instagram



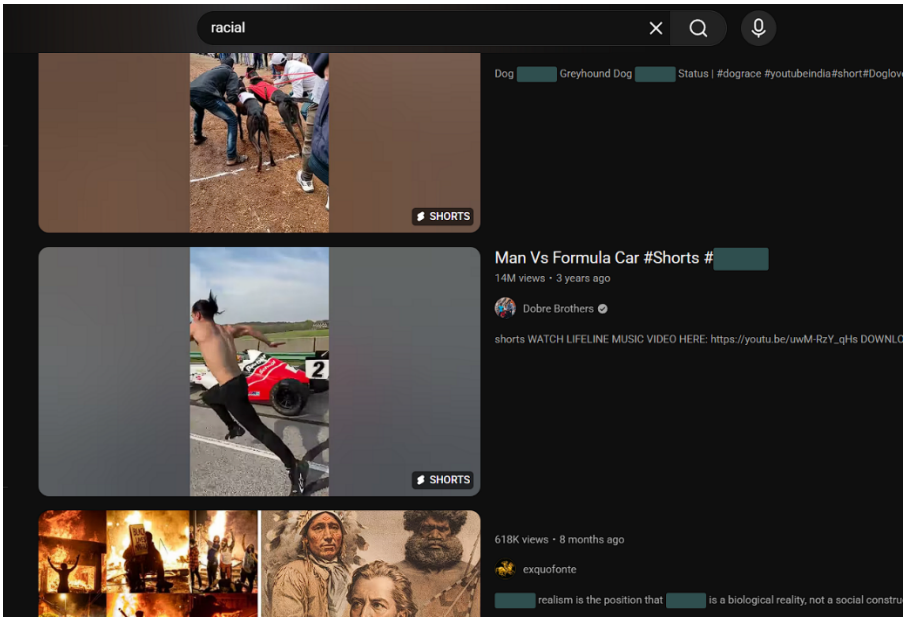
4.3 Hate Masking on Reddit



4.4 Sample Hate Speech Masking Page (Multilingual)



4.5 Hate Masking on WhatsApp



4.2 Hate Masking on YouTube

VI. CONCLUSION

The increasing use of social media and online platforms has significantly transformed communication, but it has also led to the rapid spread of harmful content such as hate speech and offensive language. This project was undertaken to address this issue by developing an automated system capable of detecting and masking such content in real time. Throughout the previous chapters, the study covered the problem definition, literature review, methodology, system design, and implementation of a machine learning-based solution.

The project successfully integrates natural language processing and machine learning techniques with web technologies to create a practical and efficient system. By combining data preprocessing, feature extraction, model training, and real-time deployment through a Chrome extension, the system demonstrates how modern technologies can be used to solve real-world problems. This chapter summarizes the outcomes of the project and highlights future possibilities for further enhancement. This project demonstrates an efficient method for detecting and masking hate speech using machine learning and NLP. The integration into Chrome and mobile platforms ensures accessibility and real-time moderation.

VII. FUTURE SCOPE

Although the current system performs effectively, there are several opportunities for further improvement and expansion. One major enhancement is the inclusion of multi-language support, which would allow the system to detect hate speech in regional and international languages. This would make the system more versatile and applicable to a wider audience.

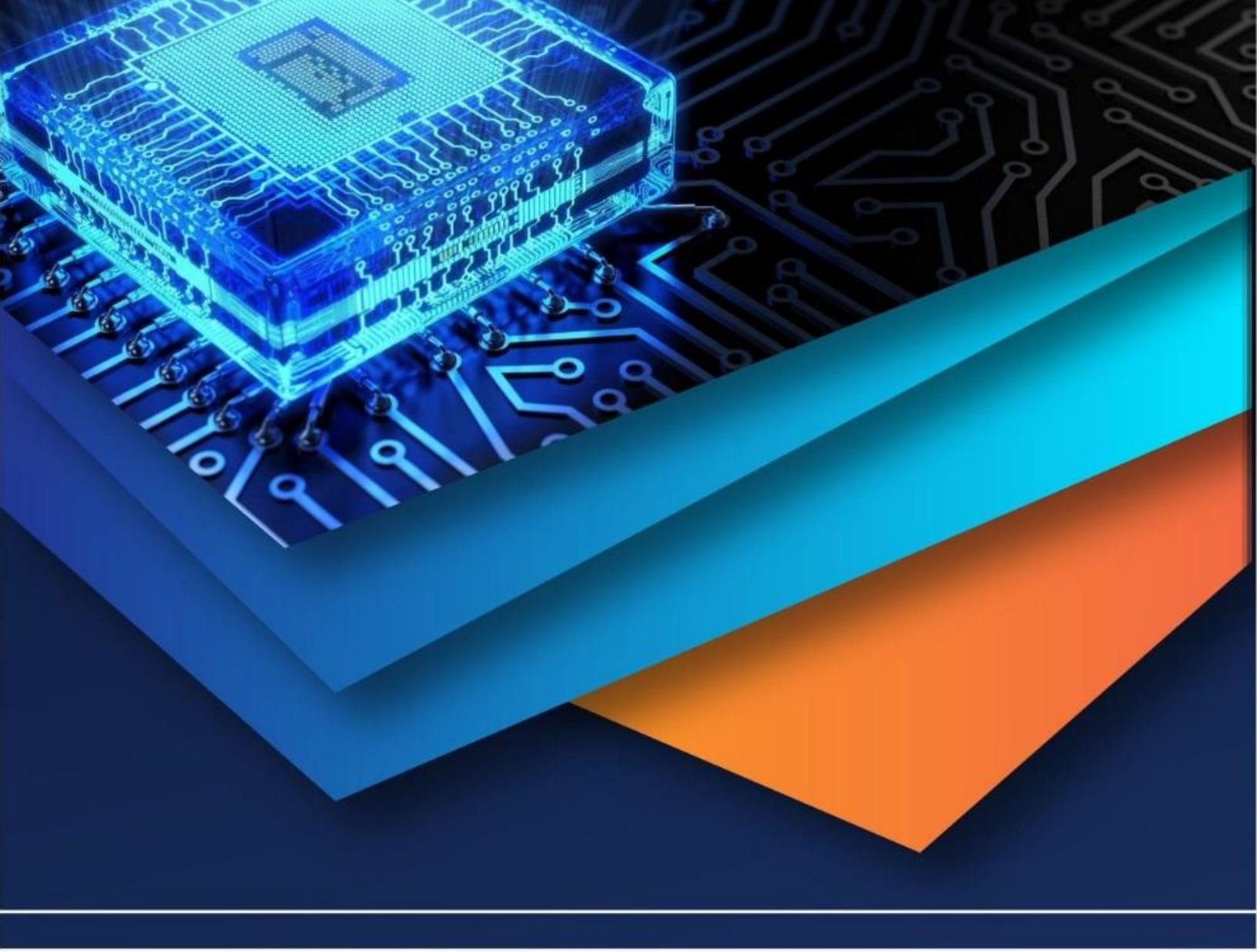
Another area of improvement is the use of advanced deep learning models such as LSTM, BERT, or transformer-based architectures, which can better understand context, sarcasm, and complex language patterns, thereby improving accuracy.

The system can also be extended to support voice-based hate speech detection, where audio inputs can be converted to text and analysed. Additionally, instead of simply masking offensive words, the system can be enhanced to rewrite sentences into polite or neutral language using AI-based text generation techniques.The developed system's modular structure—comprising data preprocessing, model training, prediction, and masking—allows for flexibility, scalability, and easy deployment on multiple platforms, including Chrome extensions. The achieved accuracy and performance metrics from the trained model demonstrate the potential of artificial intelligence in promoting a respectful and inclusive digital environment. Moreover, the integration of reporting and moderation features further enhances transparency and accountability, creating a balanced mechanism between automation and human judgment.

Overall, the project highlights how AI can be leveraged responsibly to minimize digital toxicity, protect users from harmful content, and contribute to the creation of safer online communities.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 512–515, 2017.
[2] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," Proceedings of the NAACL Student Research Workshop, pp. 88–93, 2016.
[3] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759–760, 2017.
[4] Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1621–1622, 2013.
[5] Ramos, G., Batista, F., Ribeiro, R., et al. (2024). A comprehensive review on au-tomatic hate speech detection in the age of the transformer. Social Network Anal-ysis and Mining, 14, 204. https://doi.org/10.1007/s13278-024-01361-3 Spring-erLink
[6] Singh, A., & Thakur, R. (2024). Generalizable Multilingual Hate Speech Detec-tion on Low Resource Indian Languages using Fair Selection in Federated Learn-ing. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol-ume 1: Long Papers) (pp. 7211–7221). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.400 ACL Anthology+1
[7] Fetahi, E., Susuri, A., Hamiti, M., et al. (2025). Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI. Social Network Analysis and Mining, 15, 82. https://doi.org/10.1007/s13278-025-01497-w SpringerLink
[8] Ahmad, M., Waqas, M., Hamza, A., Usman, S., Batyrshin, I., & Sidorov, G. (2025). UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers. Computers, 14(6), 239. https://doi.org/10.3390/computers14060239 MDPI
[9] Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Multilingual Hate Speech De-tection: A Semi-Supervised Generative Adversarial Approach. Entropy, 26(4), 344. https://doi.org/10.3390/e26040344 MDPI



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY