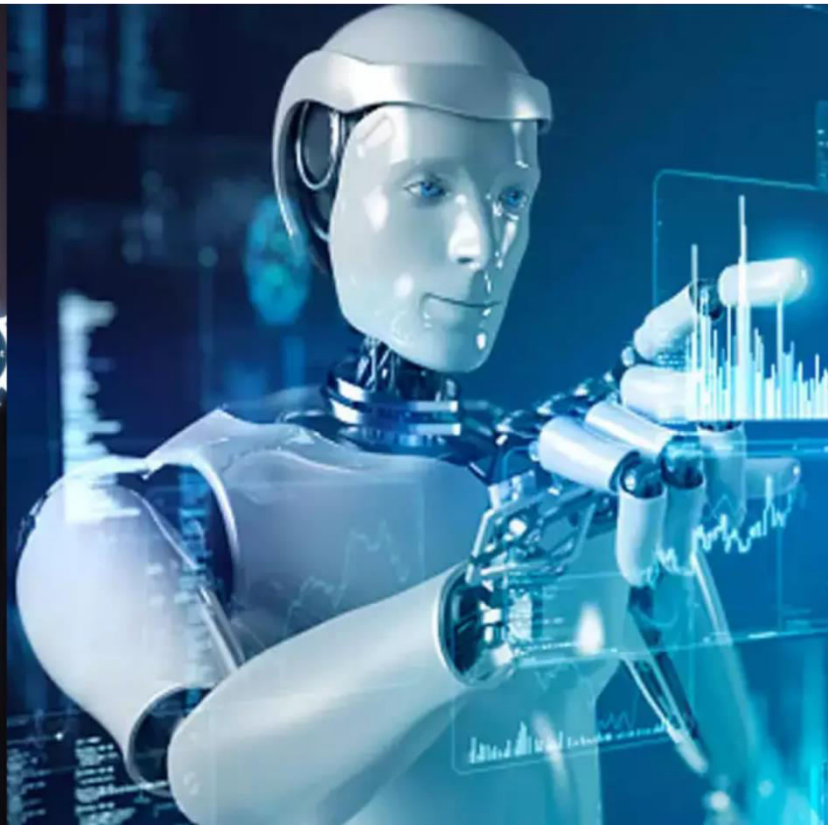
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404625|

Hate Speech Detection Model for Social Media Applications

Dr. Ravi Prakash, Dhruv Dornal, Shivam Kapoor, Yash Koli, Aditya Mane

Professor, Department of Computer Engineering, K.C. College of Engineering, Thane, Maharashtra, India

Researcher, Department of Computer Engineering, K.C. College of Engineering, Thane, Maharashtra, India

Researcher, Department of Computer Engineering, K.C. College of Engineering, Thane, Maharashtra, India

Researcher, Department of Computer Engineering, K.C. College of Engineering, Thane, Maharashtra, India

ABSTRACT: This paper suggests a machine learning-based method of automated hate speech detection using Natural Language Processing (NLP) techniques in processing text data. The model based on deep learning integrates conventional NLP methods to accurately label speech as hateful or non-hateful, with greater precision. Integrating this system into social media moderation software, we seek to reduce the dissemination of toxic discourse. The suggested framework offers encouraging outcomes, emphasizing the possibility of artificial intelligence in rendering the digital realm safer and addressing the complexities involved with linguistic nuances and inherent prejudice.

KEYWORDS: Pretrained BERT model, Natural Language Processing (NLP), Keyword extraction, Named entity recognition, Intent recognition

I. INTRODUCTION

The unparalleled growth of social media has transformed individual communication, increasing the world's connectivity. However, with its benefits, social media has also become fertile ground for hate speech, leading to real-world outcomes like psychological harm, social disturbance, and even violence. Resolving such an issue through manual processes is not scalable and efficient, given the vast quantity of content generated per second. The use of automated hate speech identification by the power of machine learning (ML) and Natural Language Processing (NLP) offers a sustainable solution to identify and counteract harmful rhetoric without compromising freedom of speech.

Traditional rule-based methods of content moderation are greatly challenged by the nuances of human language, given that hate speech typically contains insidious meanings, sarcasm, and changing slang. Contemporary machine learning algorithms, especially those that are founded on deep learning architectures, allow for the understanding of contextual signals and enhanced accuracy of classification. Through training on large volumes of labeled text, these algorithms can be capable of distinguishing between offensive language, hate speech, and neutral communication, thereby minimizing false positives and guaranteeing fairness. Nonetheless, issues such as dataset bias, identification of multiple languages, and adversarial attacks are still areas of concern.

This paper presents an end-to-end hate speech detection system that incorporates the most up-to-date natural language processing techniques with robust machine learning models. The proposed system here is structured to enhance precision, flexibility, and usability for real-world deployments, and therefore is appropriately positioned for deployment within social media moderation systems. With the resolution of ethical concerns and improved detection efficacy, this effort is part of the continued progress toward making the online space more secure and welcoming for all.

II. LITERATURE SURVEY

Increasing instances of hate speech and indecent content on social media platforms have raised great concerns regarding user safety, psychological well-being, and site integrity. Although filtering is implemented in platforms such as Twitter and Instagram, there are still loopholes, particularly regarding real-time client-side filtering. This has spurred research into AI-based, client-side solutions for automatic detection and censorship of abusive content.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404625|

Hate Speech Detection Methods
Machine Learning (ML)-driven techniques:

- Davidson et al. (2017) used logistic regression, SVM, and random forest classifiers learned on social media corpora to distinguish hate speech, offensive language, and innocuous content.
- ML models tend to use bag-of-words, TF-IDF, and n-grams for feature extraction.

Deep Learning Methods:

- Zhang et al. (2018) applied CNNs and RNNs in text classification and found better performance than the baseline models, particularly with difficult and context-dependent data.
- Pretrained models like RoBERTa and BERT have become ubiquitous contextual hate speech detectors.

Vulgarity and Profanity Filtering:

- Rule-based systems (for instance, blacklists) can efficiently detect and censore predetermined offensive words but are not very adaptive.
- Hybrid systems employ ML for context and regex/wordlists for censorship with performance. They are best suited for client-side extensions with performance requirements.

Content Moderation Browser Extensions:

- Client-side moderation software like Tweak New Twitter, UndeadBlocks Censor demonstrate the practicability of altering DOM elements to achieve real-time filtering.
- Injecting detection logic onto web pages using JavaScript content scripts is a common practice.
- Chrome's WebExtensions API offers real-time DOM access, making it ideal for moderation tools.

III. METHODOLOGY

This project will develop a Chrome extension which can identify and censor hate speech and obscene material in real-time across different social media platforms like Instagram, Twitter, and Reddit. The extension enables a safer surfing experience by allowing users to block objectionable material without requiring platform-level moderation. The extension is client-side based, intercepting text content during rendering inside the browser and processing it through embedded NLP and machine learning algorithms. The project is user-control, privacy, and dynamic website integration focused.

Workflow and Architecture
The addon consists of three basic modules: a content script, a background script, and an ML inference engine. The content script is injected into the DOM during page load and starts monitoring text elements. MutationObserver is used for dynamic content updates (e.g., infinite scroll posts). The text is fed into the ML model for a classification. The extension masks/removes offending content based on prediction. The user can interact with the addon interface to enable/disable filtering, adjust sensitivity, or customize censorship modes.

Hate speech detection model
A lightweight machine learning model forms the backbone of the extension's content classification. It is trained on well-curated datasets like Davidson's Hate Speech Dataset and HateXplain, which contain labeled social media posts. Preprocessing includes tokenization, lowercasing, stopword removal, and lemmatization. Feature vectors are generated using TF-IDF or word embeddings, based on performance requirements. For browser support, models are converted using TensorFlow.js or ONNX.js, enabling real-time inference directly in the browser. This eliminates latency and offers end-to-end user-side processing, ensuring privacy and speed.

Profanity Filtering and Hybrid Approach
While the ML model is responsible for contextual hate speech detection, there is an ongoing profanity filter that detects and hides well-known vulgar words from an existing pre-curated blacklist. The hybrid approach enhances detection speed and accuracy, especially for short or highly slang-based messages. Regex and keyword filtering facilitate real-time filtering of identified toxic words, with the ML model doing contextual analysis. The segregation also ensures lower computational demands and response times during continuous browsing sessions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404625|

DOM Manipulation for Censorship

The add-on applies JavaScript DOM manipulation to modify or censor content in real time. On detection of offensive content, operations like blurring text, symbol masking of keywords, or full post replacement with a warning overlay are dynamically performed. Changes are reversible so that users can expose content if desired. Platform-specific selectors are applied to support compatibility with different HTML structures of Instagram, Twitter, and Reddit. Updates are modular, and thus future site modifications can be supported with minimal code modifications.

User Interface and Interaction

The user interface remains minimal and in user control. Toggle switches to enable censorship, dropdowns to choose filtering modes (e.g., mild, strict), and indicators of the number of filtered posts per page are provided. User settings are stored using Chrome's local storage API so that they remain between sessions. The interface also provides transparency with whitelisting pages, reporting false positives, and model sensitivity adjustments. The goal is to give users control over their online experiences without compromising functionality or speed.

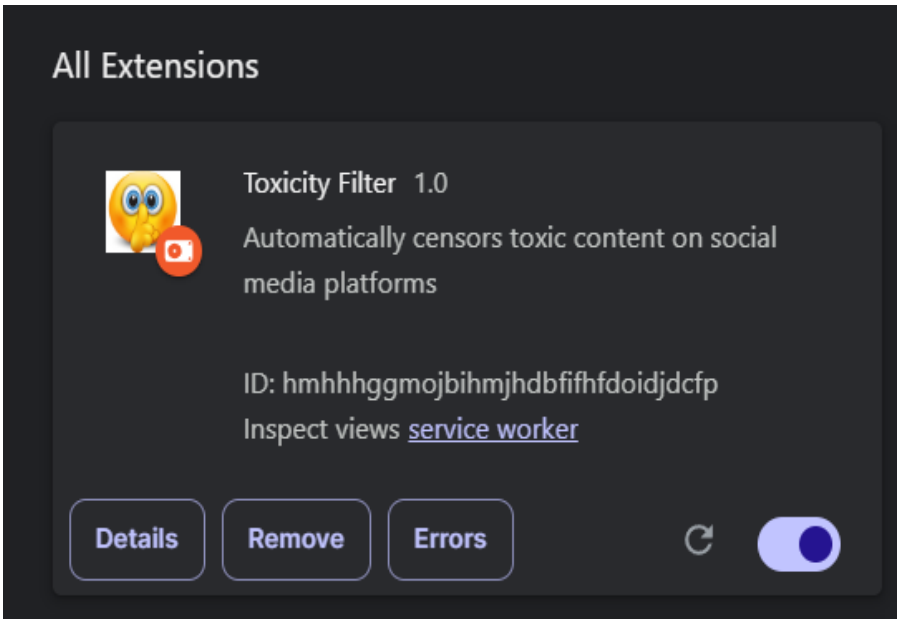
Ethical Design and Privacy Principles

One of the most important aspects of the approach is the ethical use of content filtering. All processing is locally performed within the browser, meaning that no content or user data is sent to third-party servers. This architecture is respectful of user privacy and adheres to data protection policies. Furthermore, users can opt to personalize their experience, including turning off the extension altogether. Educational messages are also integrated in order to create awareness on hate speech. By placing control within the user's hands and concentrating on ethical filtering, the extension aims to enhance safety as well as freedom of expression.

IV. EXPERIMENTAL RESULTS

Figures shows the extension for censorship (a) Shows the extension application. (b) Shows the toggle button of the extension.

The Hate Speech Filter is applied as an extension to your respective browser and toggled for use. After the toggle we can go to posts of various social media platforms to see the implementation.



(a)

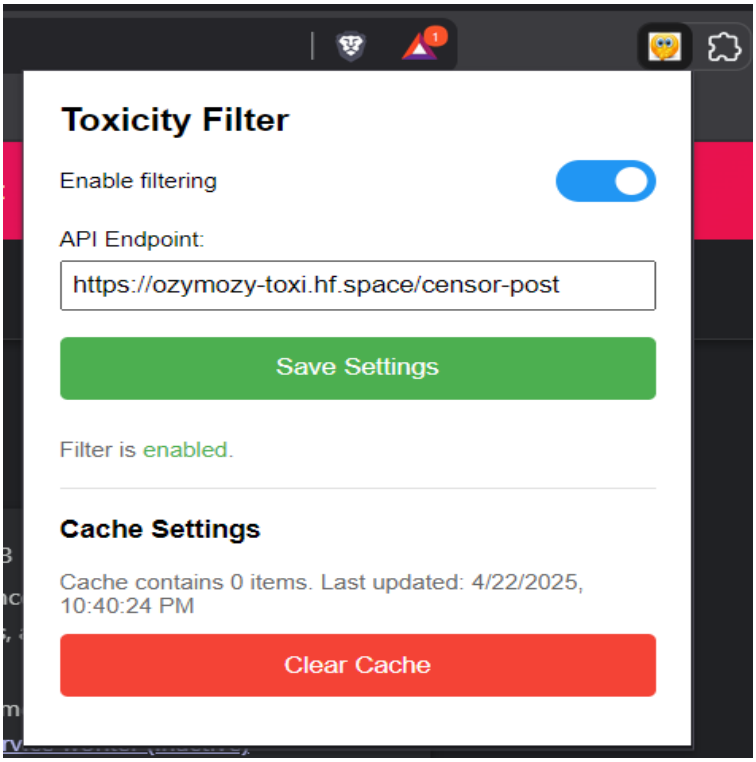
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404625|



(b)

Figures shows the results of censorship of a twitter. (a) Shows the posting window. (b) Shows the previous posts. All posts with hate speech will be censored basis on the degree of the vulgarity.

(a)

(b)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

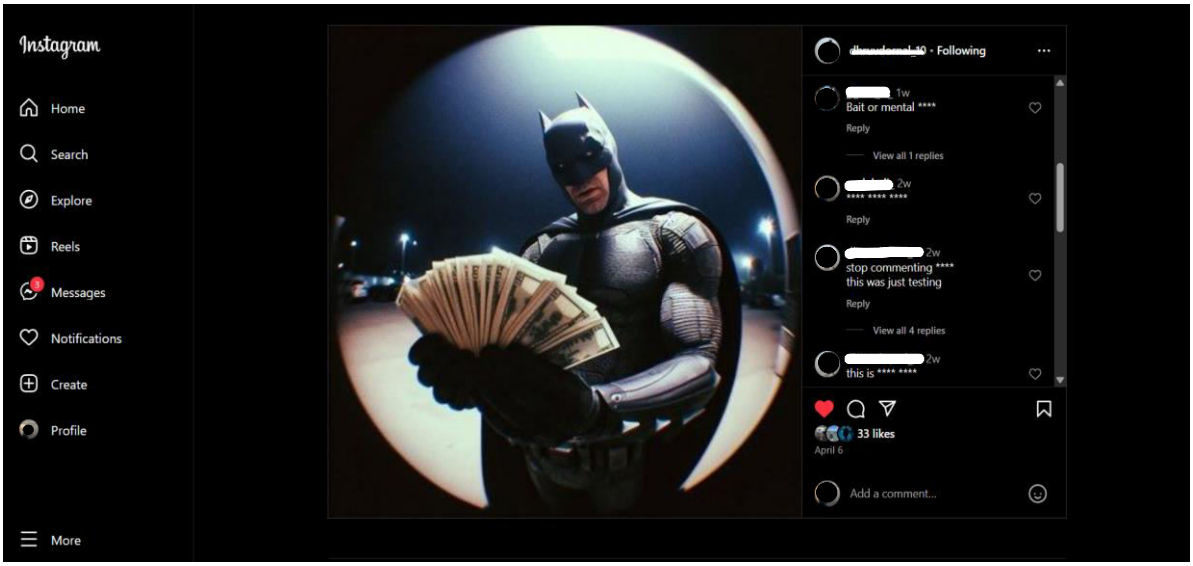
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

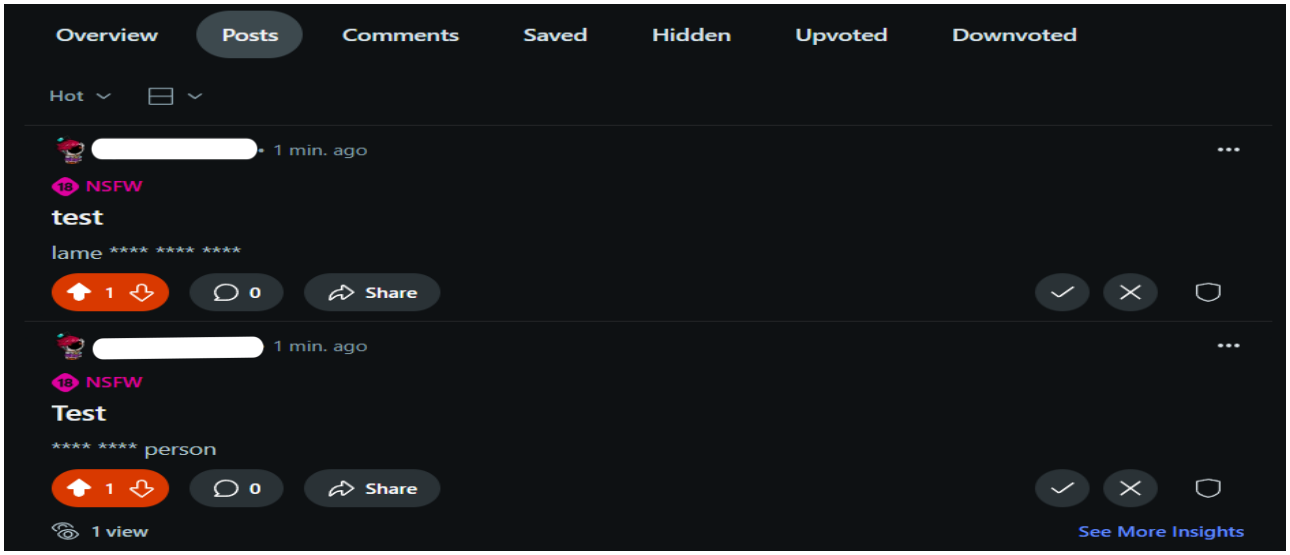
|DOI: 10.15680/IJIRSET.2025.1404625|

Figures show the results of censorship of a Instagram (a) Image of an Instagram post.
All posts with hate speech will be censored basis on the degree of the vulgarity.



(a)

Figures show the results of censorship of a Reddit (a) Image of Reddit posts.
All posts with hate speech will be censored basis on the degree of the vulgarity.



(a)

V. CONCLUSION

The proposed distributed machine learning-based hate speech detection system addresses the key issues of traditional centralized models, including scalability, accuracy, and real-time processing on multiple social media platforms. Based on the principles of distributed computing, the system offers efficient use of resources, fault tolerance, and consistency,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404625|

making the system a viable candidate for large-scale deployment. Integration with a Twitter-like social media platform, the project demonstrates the feasibility of real-time content moderation with minimal bias and ethical concerns. Additionally, the system's flexibility allows it to support emerging language styles and multimodal hate speech, making it resilient. Future enhancements may involve cross-platform, more sophisticated contextual understanding through deep learning, as well as multimodal detection in images, videos, and memes.

REFERENCES

1. Badjatiya, P., Gupta, M., Gupta, S., & Varma, V. "Deep Learning for Hate Speech Detection in Tweets." IEEE, 2017 .
2. Zhang, L., & Liu, R. "Detecting Hate Speech in Social Media."IEEE, 2018
3. Davidson, T., Warmsley, S., Macy, M., & Weber, I. "Automated Hate Speech Detection and the Problem of Offensive Language." IEEE, 2017
4. Fortuna, P., & Nunes, S. "A Survey on Automatic Detection of Hate Speech in Text. IEEE, 2018
5. Anupam Bhardwaj, Pooja Khanna, Sachin Kumar, Pragya, “Generative Model for NLP Applications based on Component Extraction”, IEEE, 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details