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.2026.82039

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

Automated Toxicity Detection in Online platform

Reshma E[1], Shalini J[2], Bharathy G[3]
[1, 2]*Department of Computer Science and Engineering, Anand Institute of Higher Technology, India*
[3]*Assistant Professor, Department of Computer Science and Engineering, Anand Institute of Higher Technology, India*

Abstract: ***The growth of online communication tools has caused a rise in the number of instances of toxic, abusive and harmful content is now a real issue for the safety of users and their digital well-being. Old School moderation practices based on the use of keyword-based filtering are generally not able to identify context and therefore lead to poor identification of content. This paper discusses an automated AI-based toxicity detection system which will be able to accurately identify, in real-time, the nature of what someone is posting, taking context into account and making use of various forms of data to help make this determination possible. The proposed system uses a hybrid model of Natural Language Processing and Deep Learning techniques to analyze multi-modal sources of data; including text, image and voice. Image data and voice data are converted to text format through Optical Character Recognition (OCR) and Speech-to-Text methods. The converted data then undergoes "preprocessing" before being evaluated using multiple techniques for text analysis, i.e. REGEX rule based filtering, TextBlob (for sentiment analysis), and BERT (for context analysis). The ability to evaluate the data at multiple layers of evaluation allows for very precise toxic scoring of the data which then allows for the automated moderation of that data through content filtering, warning generation, and user feedback. The results of the experimental analysis indicate that the proposed system has improved accuracy in detecting toxicity, reduced false positives, and improved the quality of user interactions. Therefore, it can be concluded that the proposed system is a scalable, intelligent and efficient way to create safe digital environments while also addressing the shortcomings of previous moderation techniques.***

Keywords: Toxicity Detection, Natural Language Processing, BERT, Sentiment Analysis, Content Moderation, Artificial Intelligence

I. INTRODUCTION

The huge increase in the use of things like Social Networks (Facebook, Twitter, Instagram, etc.), Instant Messenger (WhatsApp, Messenger, etc.) and Discussion Forums (Reddit, Quora, etc.) has dramatically changed how we communicate and exchange information. However, this explosive growth has also created a significant increase in the amount of toxic, harassing and dangerous content that people experience. This type of content has a negative effect on users' experience with, and mental health from, digital communication as a whole, and ultimately decreases the quality of digital communications overall.

Currently, most content moderation solutions rely heavily on either keyword-based filtering through automated systems (keyword filters) or by having people manually review the content. Because of this, most current systems fail to correctly identify the subtler types of poisonous content, such as sarcasm, implicit hate speech, and other forms of offensive/objectionable language that have context-sensitive or indirect meanings.

To keep up with the complexity and diversity of today's digital communications, traditional moderation systems are inadequate. Users' opinions can be expressed in many forms: text, images, and voice. Therefore, it is difficult for a traditional moderation system to accurately analyse all of these mediums of expression. The sheer volume of user-generated content being created makes it also very difficult for a moderator to manually moderate content and create a consistent experience for users in terms of how they will see their opinions moderated. Additionally, many existing automated moderation systems suffer from high rates of false positives and false negatives due to a limited understanding of the context of user-generated content as well as a lack of multi-modal processing capabilities. Therefore, there is a significant need for an intelligent, scalable solution that can analyse user-generated content more accurately and quickly adapt to different forms of digital communication in real-time.

A project that's trying to build an AI-Based Automated Toxicity Detection System. This is meant to be used to provide an efficient, accurate, and context-relevant way to moderate Online Content. The system is multi-modal, meaning it can accept multiple types of input by processing Text, Text that's Extracted from Images (using OCR) and transforming Voice input into Text (using Speech Recognition). The system also has a Hybrid approach; using Rules-based Methods (Regex), Sentiment Analysis (TextBlob), and Deep Learning models (BERT) to Perform Comprehensive Toxicity Detection. After calculating the Toxicity Score, the System will apply automated Moderation actions; Filtering content out, Creating Warnings and/or Blocking content.

An Integrated AI-Based Suggestion Module will also be provided, allowing Users to see possible Non-Toxic alternative expressions to create Global Positive and Respectful Communication. The goal of this system is to Create a Scalable, Intelligent and Efficient Solution to Improve Online Safety, Enhance User Experience, and Overcome the Challenges Associated with Traditional Moderation Methods.

II. LITERATURE SURVEY AND RELATED WORK

Computers have become more capable of detecting hate speech through using old-style computer methods— traditional machine-learning techniques such as "Support Vector Machines" (SVM) — but have not learned how people use the words to communicate online effectively[1]. These techniques also lacked the ability to understand the different meanings or implications in language. Nonverbal signals often convey the true meaning of an individual's intentions, such as sarcasm. Thus, the traditional keyword/system approach will not adequately detect all forms of hate speech[1]. In short, although the researchers provided evidence of using old methods to analyze online content, they did not demonstrate an accurate measure of having accomplished anything significant in providing new methodologies or evidence of the methodology shown to be effective for future users of the method.

According to Badjatiya et al. [2] their deep learning model using LSTM networks and embedding techniques is effective for improving the classification of toxic content and demonstrates that the use of deep learning models allow for the capture of sequential dependency structures within the text and improve the performance of toxicity detection models. The study finds that while the model exhibited increased accuracy with improvements in the detection of toxicity, it still requires large amounts of labelled data and computationally intensive resources thus restricting its scalability and applicability in real-time environments.

The study "Devlin et al." [3] revealed the "Bidirectional Encoder Representations from Transformers (BERT)," which significantly advanced understanding contextual information. Through analyzing how two words relate to each other in both directions; therefore, the ability to identify subtle but relevant terms increased. One of the drawbacks of BERT is that it may require a high level of computation due to its complexity. Thus, this can be an obstacle for deployment in real time systems without sufficient resources available..

Using convolutional neural networks for detecting hate speech on social media is the focus of Zhang et al.'s work [4]. Convolutional neural networks (CNNs) efficiently process local features and patterns to increase the likelihood of accurate classification above non-CNN models. In addition to this gain, CNNs have limitations in detecting long-range dependencies and deeper contextual information, both of which are critical for recognizing the many variations of toxic language.

Together, these Together, these studies illustrate how toxic detection technologies have transitioned from traditional (machine learning) techniques to now incorporating more advanced methods (primarily the use of deep learning and transformers). Many of these systems ultimately lack multi-modality (e.g., images and text), as well as the ability to adapt to real time, and not being able to use integrated moderation as part of their overall systems. The proposed Automated Toxicity Detection System builds on the advancements made with current systems by using a hybrid (i.e. rule-based filters, sentiment analysis and contextual deep learning) approach. Therefore, the Proposed System is a complete, accurate, scalable and real-time method of detecting various forms of toxic content from multiple channels on the internet and is a fully comprehensive technical solution for any modern form of web-based platform.

III. PROPOSED SYSTEM

An AI-powered Automated Toxicity Detection Platform is an information technology (IT) application that automatically detects harmful, abusive, and toxic content on social media sites through the use of Natural Language Processing (NLP) and deep learning technologies. Content analysis may include the detection of text, images, and audio by the Automated Toxicity Detection Platform. Text will be analyzed directly by the Automated Toxicity Detection Platform as it appears on the social media platform; images that are provided will be converted into text through OCR prior to being processed by the Automated Toxicity Detection Platform; and audio will be processed using a speech-to-text process in order for the audio to be processed by the Automated Toxicity Detection Platform. The Automated Toxicity Detection Platform will allow organizations to detect various forms of antisocial communications that have been common practice in today's society.

At the heart of the Automated Toxicity Detection Platform is a hybrid toxicity detection engine. This engine utilizes various analytical methodologies to increase the performance and reliability of the toxicity detection process. The toxicity detection engine employs a filtering process based on regular expressions (regex) to identify and detect explicit offensive keywords and keyword patterns.

In addition, the Automated Toxicity Detection Platform will use sentiment analysis as a method for analyzing the overall emotional tone of the content being analyzed. Furthermore, the Automated Toxicity Detection Platform has a deep learning (DL) model built on BERT (Bidirectional Encoder Representations from Transformers) that allows for capturing the contextual meaning of the individual's communication and will assist in detecting subtle forms of toxicity, such as impersonating other people; sarcasm; and implicit hate speech

The combination of the above analytic methodologies will assist in producing a final toxicity score that will enable the Automated Toxicity Detection Platform to classify content as being either toxic or not toxic, and will provide greater accuracy for all classifications.

This system offers a scalable, efficient and intelligent approach to contemporary Content Moderation. It uses a combination of technique including rules-based decisions with machine learning through sentiment analysis (to help determine which content should be approved vs approved), as well as deep learning methods. By doing so, it increases detection accuracy and reduces false positives as well as providing a better environment for safe digital interactions.

IV. SYSTEM ARCHITECTURE

The proposed AI-Powered Automated Toxicity Detection System has been developed, featuring a multi-layer architecture to ensure the monitoring of online content at scale, in real-time. Consisting of five distinct but connected layers (input, preprocessing, detection, application and data), the system will process and analyze multi-modal (textual, photographic and audio) user inputs. Audio and photo input files will be converted to text using Optical Character Recognition (OCR) and Speech-to-Text technology, followed by preprocessing steps to clean/normalize/tokenize the textual data. The core detection layer will deliver a toxicity score and classification of the content employing a hybrid approach using combining Regex filtering, TextBlob sentiment-based analysis, and BERT contextual understanding. Having received a toxicity classification from the detection layer, the application layer will deliver automated moderation actions to warn, filter and block, along with delivering non-toxic alternative suggestions using AI-based techniques. The data layer will have a secure store for all inputs, output results and activity logs for monitoring and analysis. This multi-layer architecture facilitates fluid data flow, intelligent decision-making processes, effective toxicity determination, and enables the system to be implemented on a large scale and in real-time.

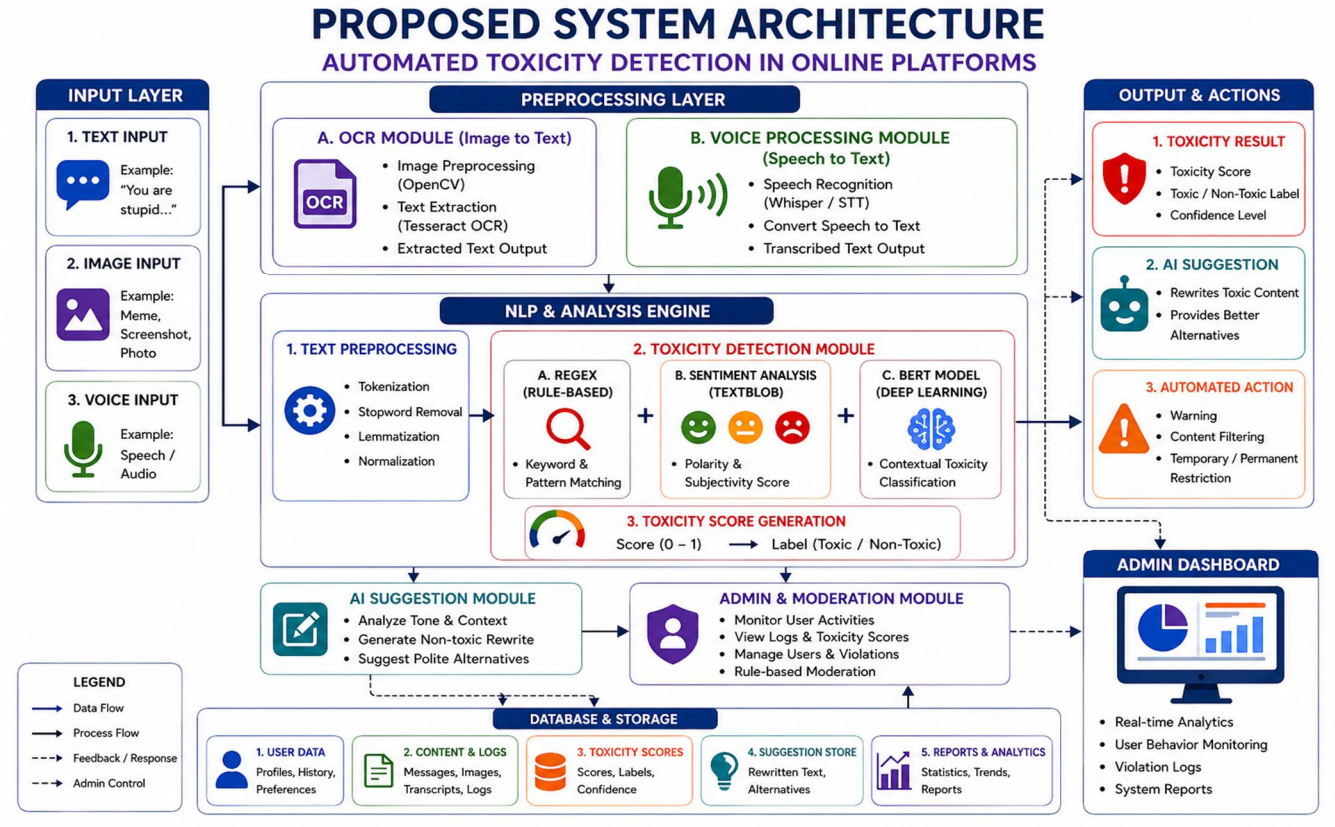


Fig 1: System Architecture

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

V. IMPLEMENTATION

The AI-Powered Automated Toxicity Detection System is designed using an organized, pencil-based structure. This will maximize its operational efficiency, capability for expansion, and ability to provide real-time content moderation for the entire cyberspace environment.

The solution follows a structured sequential method of workflow beginning with user input received then followed by the first time toxicity detection, what's included; toxicity analysis and automated moderation of content. Finally, there will be four separate and distinct modules that comprise the overall system as outlined in more further detail below..

A. User Input and Data Acquisition Module

This portion of the application is used to collect user-generated content from multiple sources of content creation. It allows for multi-modal user input in the form of text, images, and audio (i.e., voice) content.

Users will submit content through the Application User Interface (API), which has been integrated with various online platforms. Text data will be processed directly, while image data will be converted to text format using Optical Character Recognition (OCR) technology; audio data will be converted to text using speech recognition technology. Via the submission processing layer of the User Input and Data Acquisition Component, all user inputs will be supplied to the platform in a standard text format to enable uniform processing across the platform.

This component of the application is also the first point of contact for users of the application and facilitates seamless interaction (defined as an uninterrupted flow of information) between users and the platform.

B. Preprocessing and Feature Extraction Module

This Preprocessing & Feature Extraction module prepares the input data for further analysis by performing necessary steps to clean up any unwanted characters/symbols/noise from the extracted text (i.e., removing stop words). It also tokenizes and normalizes (i.e., structures) the text so that it will be more consistently formed and improve the accuracy of feature extraction through Tokenization & Normalization techniques along with the removal of stop words, in order to create more meaningful features for improving accuracy of analysis.

By using these techniques during preprocessing, you are ensuring that your input data will be in an appropriate format to detect toxicity effectively and reduce errors in later stages of processing..

C. Toxicity Detection and Classification Module

The main function of this module is to perform the analysis of toxicity through hybrid methods. That means that we use several different approaches/rules to analyze toxicity. These methods will use a combination of techniques including Regex rule filtering, TextBlob sentiment analysis, and BERT for contextual understanding.

Regex will allow us to detect explicit offensive words/patterns through textual matching. TextBlob will allow us to measure whether there is a positive vs negative sentiment polarity associated with the text. BERT will enable us to understand the contextual meaning of the text and determine if there are any implicit or complex forms of toxicity, such as sarcasm or indirect abuse. These outputs will be combined to produce a toxicity score (i.e., indication of toxicity) that can then be used to classify the content as either toxic or non-toxic. This multi-faceted approach ensures that our methods are both accurate and reliable.

D. Moderation, Feedback, and Data Management Module

The Moderation, Feedback, and Data Management (MH&M) module is designed to support the implementation of moderation actions and manage the outputs generated by the system. First, the system executes an action based on the classification of toxicity in each entry into the system (i.e., issue warning; filter out objectionable content; block objectionable messages).

A supplemental AI-based suggestion mechanism is provided to promote the use of alternative non-toxic expressions thereby enhancing positive communication among people that use the system.

All moderation actions performed will have their associated data (e.g., moderation log, toxicity scores, etc.) stored securely for future monitoring and analysis by the system to improve the overall effectiveness of the moderation efforts. As well, continuous feedback via data collection, analysis, and reporting will lead to improvements in both user experience as well as the ability of the digital arena to provide a safe environment for users.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

VI. RESULTS AND DISCUSSION

An Automated Toxicity Detection System capable of detecting and moderating harmful content on online platforms has been developed and evaluated using AI. This system incorporates the ability to process multiple inputs, from both audio and visual sources, combine AI methods of detecting and moderating harmful content, and provide automatic moderation capabilities, thus creating a more safe and user-friendly online environment. The main goal of this project had been to increase the accuracy of detecting harmful content, dealing with complex patterns of language, and providing moderation in real-time.

A. Analysis of Results

The system was very effective at detecting toxic content using a combination of different techniques including Regex Filtering, sentiment analysis and BERT for understanding the context of information. By using more than one technique, the accuracy of detection improved over just using one method. The preprocessing module prepared input data, making it clean and structured which improved how the detection models were able to perform. Because of the ability to process text, images, and voice input formats, this system is also versatile enough to be used on many different online platforms.

The key to being able to understand the context of the information was the BERT based model, which allowed the system to identify more complex forms of toxicity like sarcasm, indirect abuse, and more subtle forms of offensive language. On the other hand, Regex Filtering allowed for quick detection of clearly damaging content by searching for and capturing only clearly defined toxic words.

B. System Performance Comparison

A comparative analysis has been conducted comparing existing toxicity detection systems with the model proposed use of key performance metrics (accuracy, efficiency and adaptability), and the results are as follows…

1.) Accuracy: Because the new system uses automated detection and AI-analytical techniques to deliver relevant content for learners, it was able to achieve a greater overall rate of accuracy than traditional systems..

2.) Efficiency: Because the new system automates the detection and moderation of toxicity in its content, there is a significant reduction in time and manpower used to perform this task..

3.) Usability: The new system's user-friendly interface and structured workflow helped to improve accessibility and ease of use for both admins and students as well as all users.No one measured in the graphical comparison chart above that was measured on an existing traditional system performed worse than the new system on any of the performance measures defined. Thus we may conclude the new system was a more effective and efficient means to provide a personalized and intelligent educational experience.

C. Overall Outcome

The system effectively detects and reduces toxic content, improving online safety. It provides stable, scalable, and real-time performance across platforms.

.

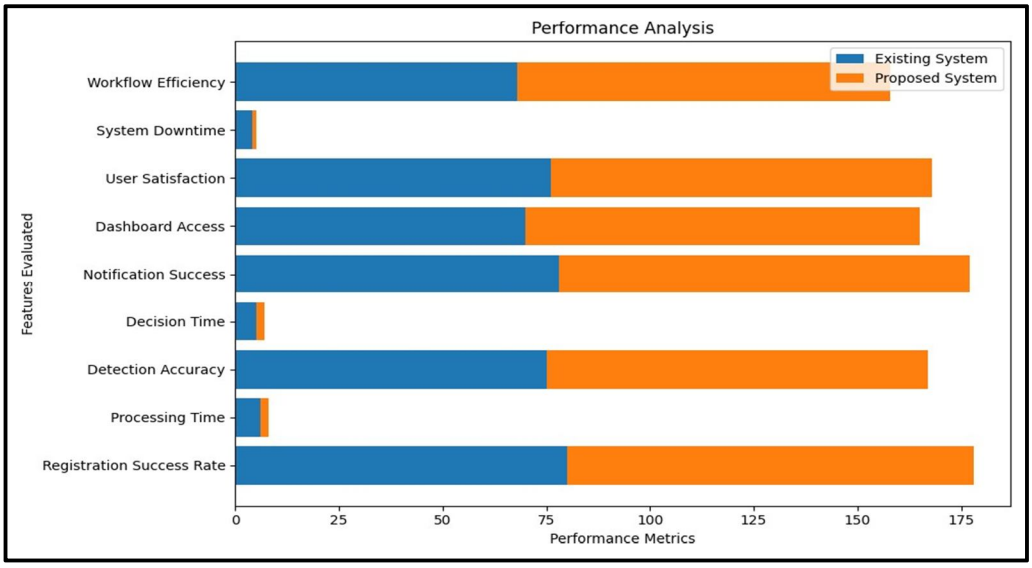


Fig 2: System performance Comparision

VII. CONCLUSION

The presented paper describes an AI-based automated toxicity detection system intended to enhance the quality and safety of online communication through intelligent moderation. The system integrates the processing of multi-modal inputs, pre-processing of text, and a hybrid detection framework combining Regex-based filtering with sentiment analysis and BERT-based contextual understanding to facilitate accurate detection of both explicit and implicit toxicity, including contextually dependent and nuanced expressions.

In conclusion, the proposed system offers an effective, scalable, and intelligent method for addressing the increasingly complex issue of toxic content in digital environments and emphasizes the value of AI-based monitoring systems in improving user experiences and providing safe online interactions.

VIII. FUTURE WORK

Future enhancement potential exists within our existing Automated Toxicity Detection System. For example, incorporating more advanced low-latency transformer models will allow us to achieve both faster processing and increased accuracy for our computation model's performance.

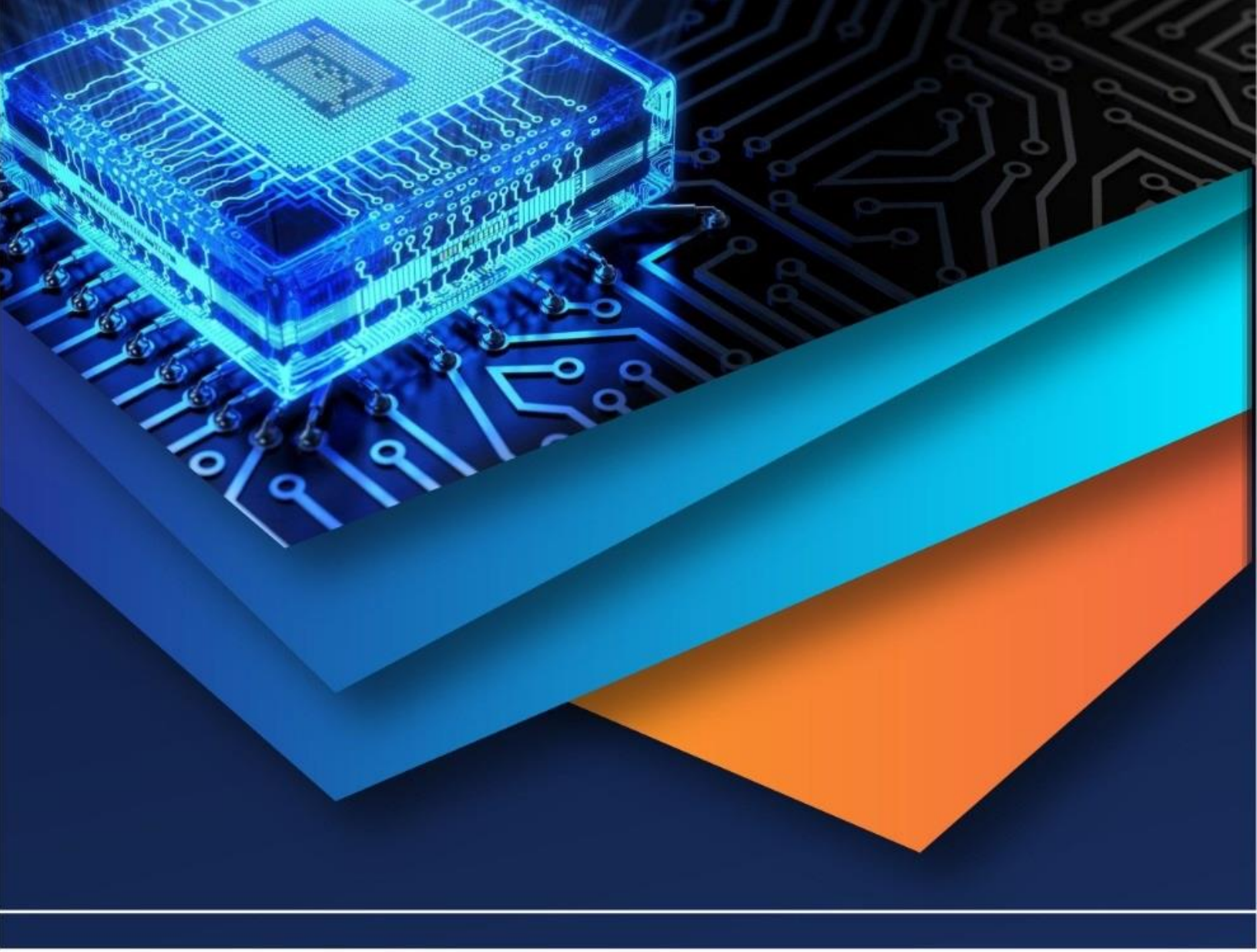
Mobile-based moderation systems would also provide enhanced accessibility to our system. By providing stream-based moderation for live chat rooms and other forms of social media, the response and functionality of our system may significantly increase.

Another possible future improvement would be to implement support for multiple languages. Doing so would allow us to detect toxic activity in all languages and cultures, thereby providing a more inclusive experience for all users. The ability to use voice-based moderation would give us additional insight into user intent and communication patterns, thereby providing deeper analysis of toxicity within online environments.

Finally, implementing explainable AI techniques would provide transparency to users by helping to explain why decisions were made regarding whether or not toxicity exists. Predictive analytics would help prevent toxic interactions by helping identify potential behaviors prior to them occurring. Together with these improvements, we would have created an automated solution that is more holistic, streamlined, and globally-oriented with respect to creating, maintaining, and promoting safe and respectful online communities.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” Proc. International AAAI Conference on Web and Social Media (ICWSM), 2017.
[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” Proc. World Wide Web Companion (WWW), pp. 759–760, 2017
[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. NAACL-HLT, pp. 4171–4186, 2019.
[4] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on social media using convolutional neural networks,” Proc. European Semantic Web Conference (ESWC), 2018.M
[5] M. Mozafari, R. Farahbakhsh, and N. Crespi, “Hate speech detection and racial bias mitigation in social media based on BERT model,” Proc. EMNLP Workshops, 2019.
[6] B. Mathew, P. Saha, S. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A benchmark dataset for explainable hate speech detection,” Proc. AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 14867–14875, 2020.
[7] D. Noever, “Machine learning suites for online toxicity detection,” IEEE Access, vol. 6, pp. 1–10, 2018.
[8] F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on Facebook,” Proc. Italian Conference on Computational Linguistics, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY