iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68277

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 IV Apr 2025- Available at www.ijraset.com

Automatic Detection of Cyberbullying using Machine Learning

Prof. Manali Patil[1], Sawmya Pandey[2], Samruddhi Yadav[3], Shreya Deshmukh[4], Bhavana Jagdale[5]
[1]*Assist. Professor, Dept. of Computer Engg., Alard College of Engineering, Pune, Maharashtra*
[2, 3, 4, 5]*BE students, Dept. of Computer Engg., Alard College of Engineering, Pune, Maharashtra*

Abstract: ***Cyberbullying is one of the most recent evils of social media. With a boom in the usage of social media, the freedom of expression is being exploited. The program presents a multi-group chat application developed using Python sockets and a Tkinter-based graphical user interface (GUI). The application allows users to create and join chat rooms, send text messages, and share files within rooms. Additionally, an AI-based message filtering system is integrated to detect and hide messages containing bullying content using a pre-trained Linear SVC model. The system enhances real-time communication while ensuring a safe chat environment by preventing offensive or harmful conversations.***
Index Terms: Cyber bullying, Machine Learning, Natural Language Processing (NLP), Hinglish Languages.

I. INTRODUCTION

In recent years, instant messaging applications have evolved significantly, enabling users to communicate across various devices in real-time. These applications are widely used in businesses, educational institutions, and social networking. However, with this increasing connectivity comes the risk of cyberbullying, harassment, and inappropriate content, particularly in public or semi-public chat environments. Many existing chat applications either lack content moderation or rely on manual intervention, which is time-consuming and ineffective. There is a growing need for automated, AI-powered moderation tools that can identify and filter harmful messages in real-time without affecting the flow of conversation.

With the increasing reliance on digital communication, online chat applications have become an essential part of social and professional interactions. However, these platforms often suffer from issues such as cyberbullying, spam, and misuse of language. This project aims to address such concerns by implementing a real-time, room-based chat system that not only facilitates communication but also integrates AI-driven message filtering. The chat system is built using Python sockets for networking, Tkinter for GUI-based interaction, and machine learning (ML) models for detecting inappropriate messages.

II. MOTIVATION AND RELATED WORKS

A. *Motivation*

The primary motivation for this research comes from the increasing incidents of cyberbullying and online harassment in digital communication platforms. Studies indicate that a large percentage of users—especially students and teenagers—experience harmful or offensive messages in group chats. Some key concerns that motivated this project include:

1) Cyberbullying in Group Chats

- Users often engage in harassment, hate speech, or personal attacks.
- Victims may feel uncomfortable or even leave chat platforms due to offensive content.

2) Lack of automated moderation

- Most platforms use keyword-based filtering, which is ineffective.
- Human moderators can only monitor a limited number of conversations.

3) Need for a real-time AI-based solution

- By using machine learning models, harmful messages can be detected and hidden instantly.
- This ensures a safe, non-toxic chat environment without manual intervention.

4) Expected Impact

- Ensuring a positive online space where users feel safe to interact.
- Reducing the psychological harm caused by offensive messages.
- Encouraging responsible communication through automated intervention.

B. Related Works

In the past work, authors and researchers have experimented with various machine learning models (such as logistic Regression, Naive Bias, Random Forest, Support Vector Machine) for the detection of cyberbullying. For this, they have taken datasets from a variety of social media platforms (like FormSpring, Twitter, Wikipedia etc.). They have also proposed deep learning models and claimed that deep learning based models outperformed machine learning based models for this classification problem.

Rui Zhao and Kezhi Mao [3] used a new representation learning technique. In this method, they used Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) developed via semantic extension of the deep learning model stacked denoising autoencoder. P. Zhou, et. al. [4] proposed attention-based B-LSTM technique, S. Bhoir, et. al. [5] presented research of various word embedding techniques based on various parameters. Banerjee [6] used CNN with GLoVe embeddings to achieve a higher accuracy.

Sweta Agrawal and Amit Awekar [7] presented a systematic work in which they have experimented with four different DNN models: Convolutional Neural Network (CNN), Long Short-Term Memory Neural Network (LSTM), Bi-directional LSTM and BLSTM with attention. In these Deep learning models, they have also used some different word embedding techniques-random, GloVe, and SSWE.

III. METHODOLOGY

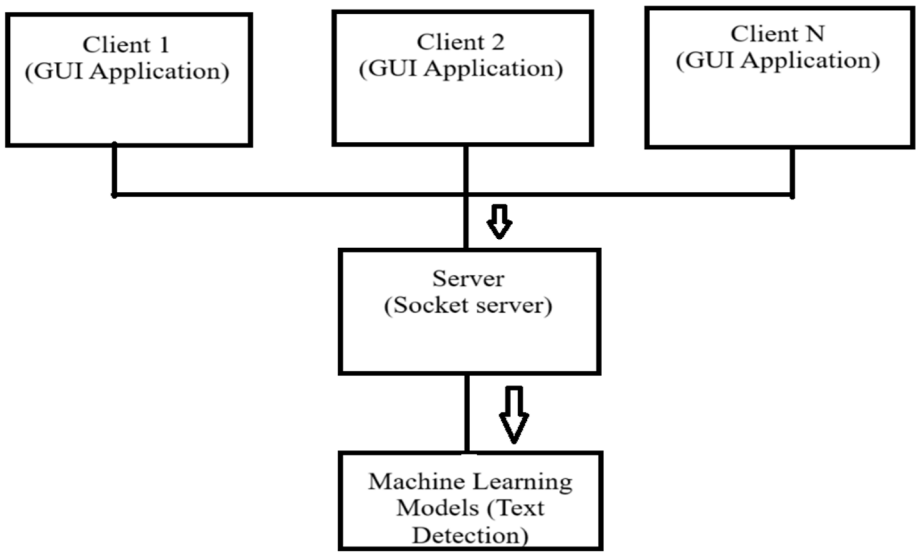


Fig 1: Methodology Diagram

A. Data Collection

Collecting a varied collection of labelled text data from forums, comment sections, and social media platforms with categories designating bullying and non-bullying content is the first stage. Training will be conducted using public datasets such as the Toxic Comment Classification Challenge and the Cyber bullying Detection dataset. Text cleaning includes lowercasing, eliminating links, special characters, and punctuation. Dividing text into words or tokens is known as tokenization.

We use publicly available datasets from social media platforms such as Twitter and Reddit. These datasets contain labelled instances of cyber bullying and non-bullying text.

B. Data Preprocessing

1) Text Cleaning: Remove special characters, stop words, and URLs.
2) Tokenization: Split text into words or phrases.
3) Stemming/Lemmatization: Convert words to their base forms.
4) Vectorization: Convert text into numerical representations using TF-IDF or word embedding.

C. Model Selection and Training

To create a successful cyberbullying detection model, a number of machine learning algorithms will be evaluated: Supervised Learning Models: We'll start with models like Support Vector Machines (SVM), Naive Bayes, and Logistic Regression. To identify the patterns that differentiate bullying content from non-bullying content, these algorithms use labeled data. Hyperparameter tuning: To tune hyper-parameters like learning rate, batch size, and number of layers in deep learning models, methods like Grid Search or Random Search will be used.

D. Model Evaluation

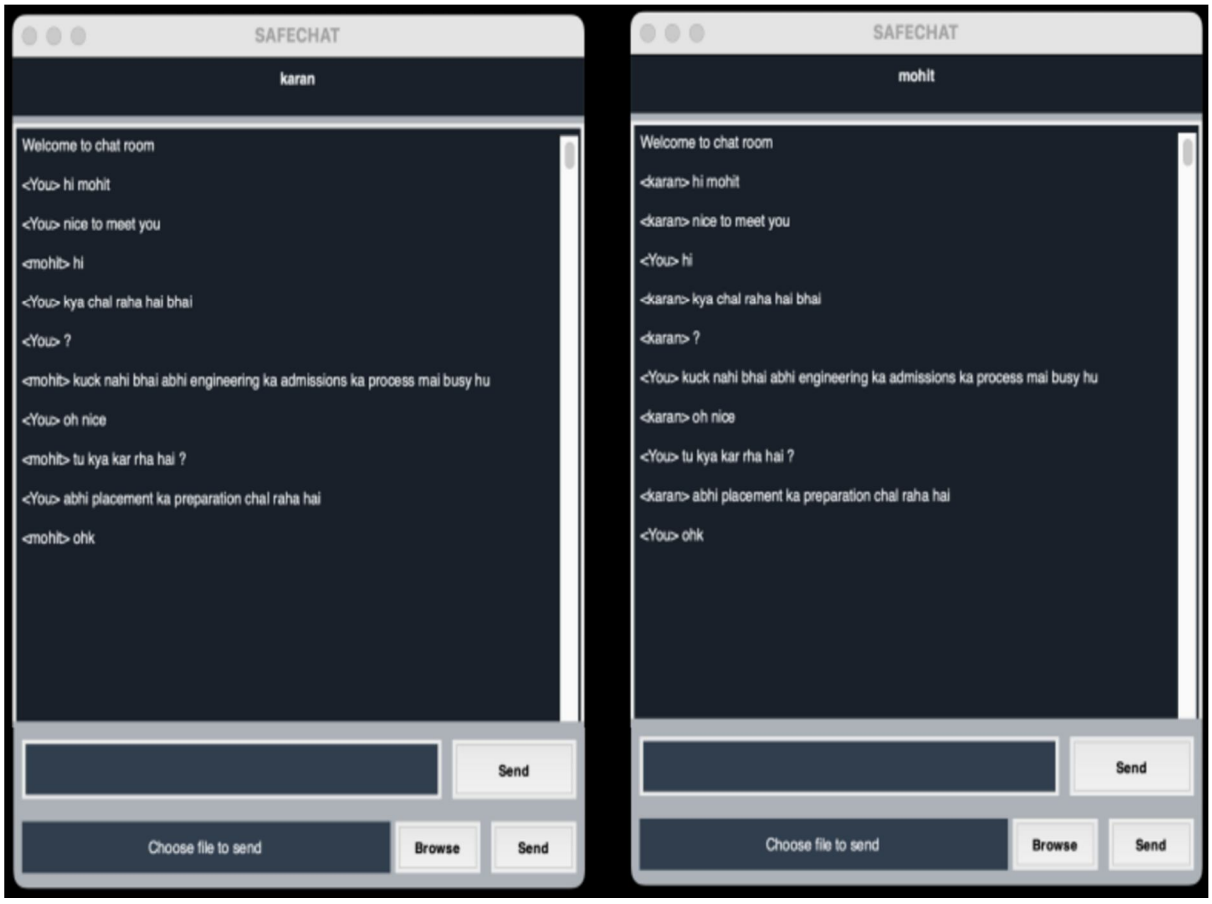
Model evaluation is a critical step in assessing the performance of a cyberbullying detection system. To evaluate the effectiveness of the trained models, several performance metrics will be used, including accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correct predictions, but for imbalanced datasets, precision and recall are more important as they focus on the model's ability to correctly identify bullying content (precision) and its capacity to detect all instances of bullying (recall). The F1-score, which is the harmonic mean of precision and recall, will be used to balance these two metrics, providing a more comprehensive view of the model's effectiveness. Additionally, the confusion matrix will be employed to visualize the distribution of true positives, true negatives, false positives, and false negatives, helping to identify areas where the model may be misclassifying content. This evaluation process ensures that the model not only performs well overall but also avoids critical misclassifications that could have harmful consequences in real-world applications.

IV. RESULT

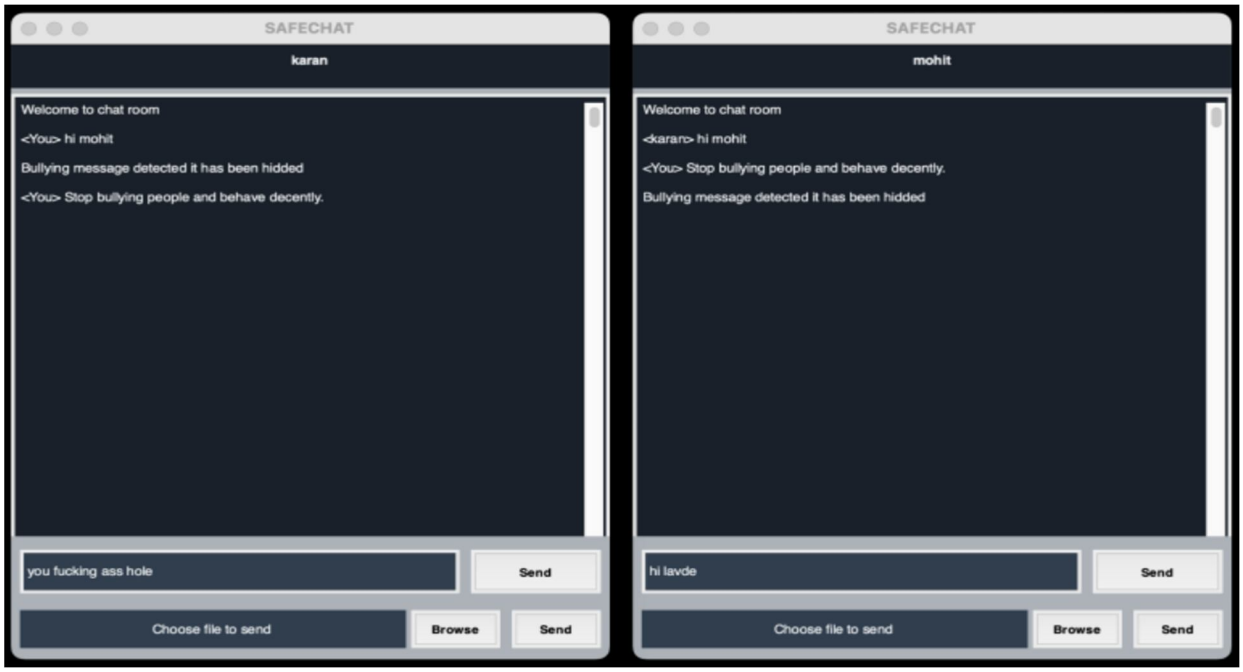
A. Bullying and Non-Bullying Flow

1) Non-Bullying Flow

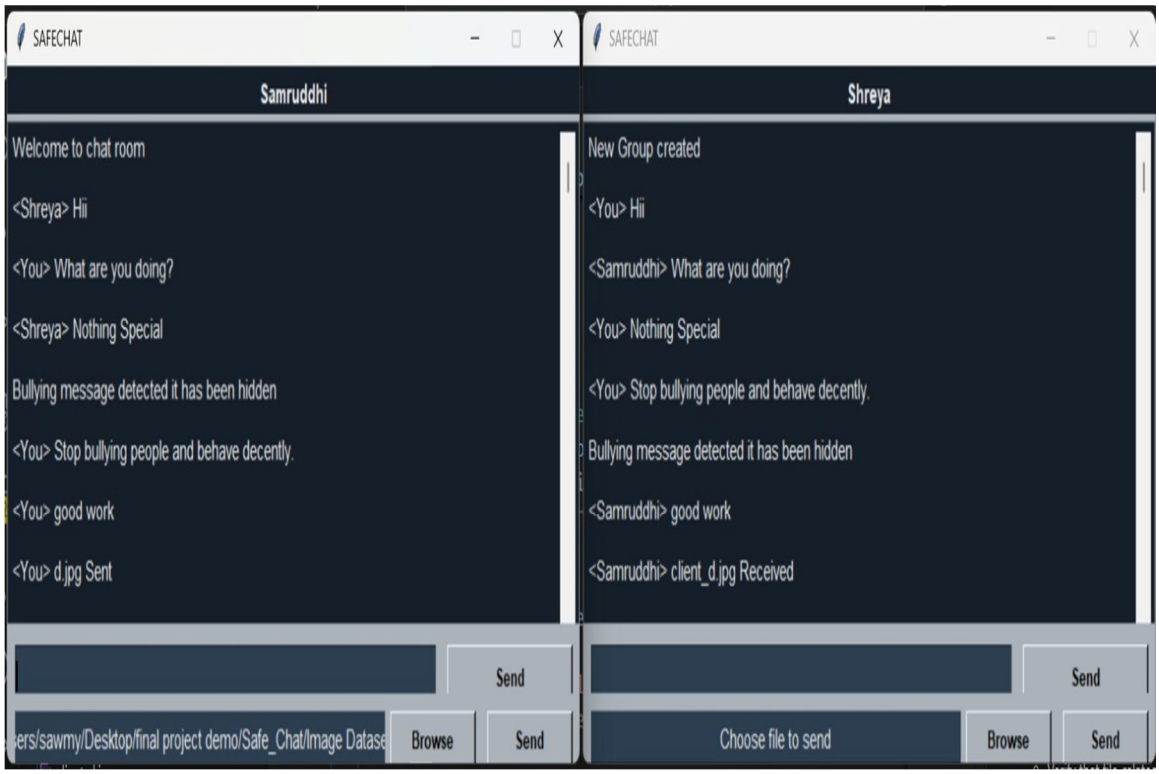
Whenever the user posts a message in the chat, our prediction service will the load the model and if the text enter is categorized as non-bullying then text or messages will be displayed on the chat screen as shown in the fig below.



2) *Bullying Flow*



Whenever the user posts a message in the chat, our prediction service will load the model and if the texts enter is categorized as bullying, then the message will be not displayed on the chat screen, the sender will get the warning as Stop bullying people and behave decently and the receiver will not receive the bullying message. Instead, they will be informed that a bullying message has been detected it and it is hidden as shown in the above figure.



V. CONCLUSION

By leveraging machine learning, it effectively blocks bullying content and ensures a safer communication environment. With further improvements in model accuracy and user experience, this application can be a valuable tool for promoting positive online interactions. Thus, we have successfully extracted, cleaned, and visualized the data using various Python libraries.

We also implemented several natural language processing (NLP) techniques such as tokenization, lemmatization, and vectorization (i.e., feature extraction). After reviewing multiple research papers in this field, we analysed different feature extraction techniques. We found that Count Vectorizer and TF-IDF provide better accuracy compared to Word2Vec and Bag of Words. To determine the best feature extraction method between Count Vectorizer and TF-IDF, we conducted a comparative analysis and observed that Count Vectorizer slightly outperforms TF-IDF in terms of accuracy.

Next, we explored various machine learning algorithms and applied several of them to our project. We trained our models using Count Vectorizer as the feature selection method and obtained good accuracy and efficiency. After training, we summarized all the algorithms in a single plot, comparing Accuracy and F1 score.

Upon analysing the results, we found that Linear SVC and Stochastic Gradient Descent (SGD) classifiers performed the best in classifying and predicting bullying messages in Hinglish. These models not only achieved better classification performance but also required less training and prediction time compared to other algorithms.

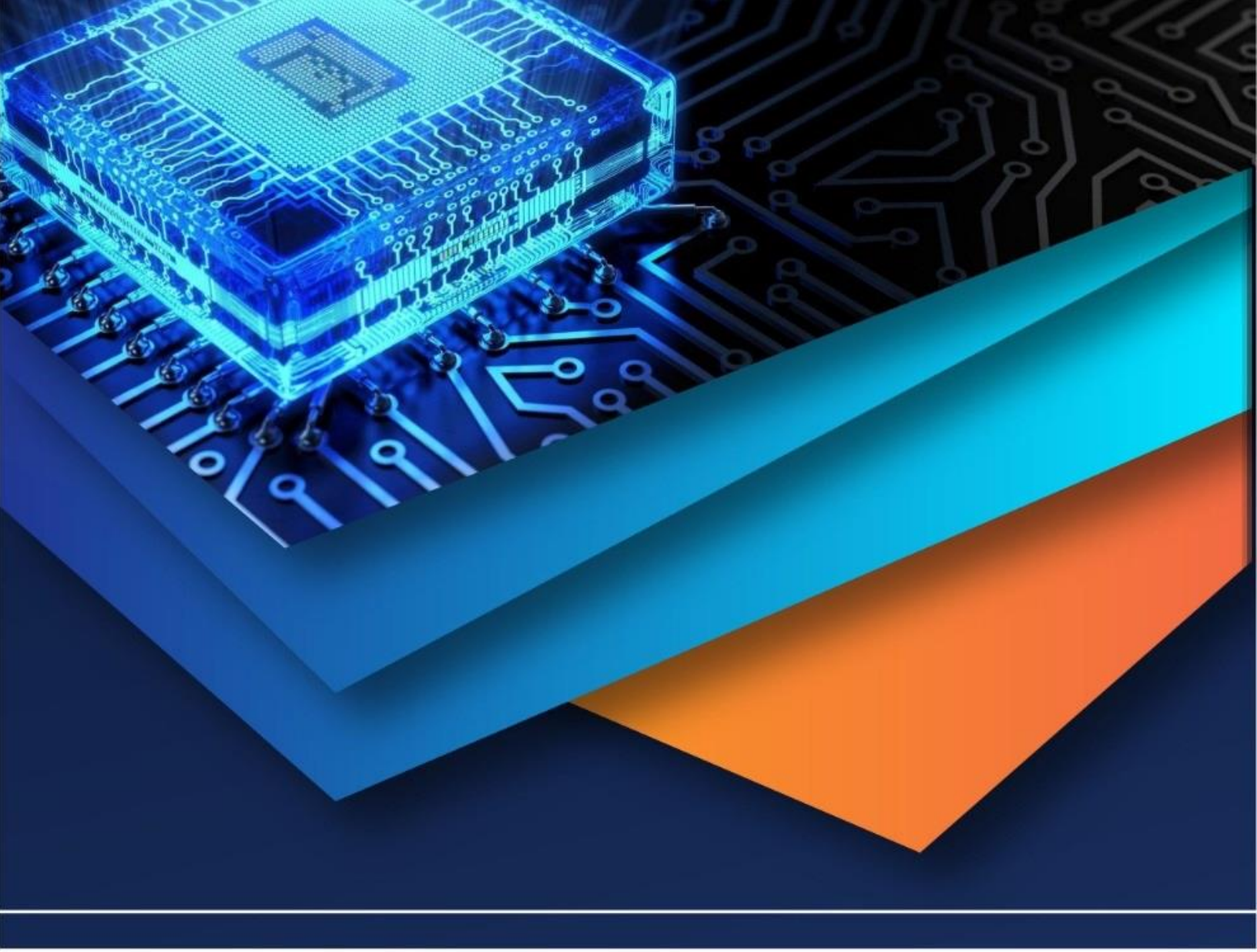
VI. FUTURE SCOPE

The future scope of cyber bullying detection systems holds significant potential for improving online safety across various platforms. As cyber bullying occurs globally, future systems can focus on expanding multi-language support, ensuring effective detection in diverse linguistic and cultural contexts. Additionally, real-time monitoring and feedback will become increasingly important, allowing for quicker interventions in detecting and addressing harmful content. Integrating cross-platform capabilities could further extend the reach of these systems to gaming, online education, and messaging platforms.

Advancements in context-aware models and emotional intelligence will enable the system to understand the emotional tone and intentions behind text, improving the detection of subtle forms of bullying. Furthermore, efforts to reduce bias and ensure fairness will be critical in making these systems effective across diverse communities. Lastly, integrating mental health resources and providing continuous model improvement will ensure the system remains adaptive to new trends and offers support for victims, ultimately creating a safer online environment.

REFERENCES

[1] Varun Jain, Vishant Kumar, Dinesh Kumar Vishwakarma and Vivek Pal, “Detection of Cyberbullying on Social-Media Using Machine Learning” IEEE 2021.
[2] Saloni Mahesh Kargutka and Prof. Vidya Chitre, “A Study of Cyberbullying Detection Using Machine Learning Techniques” IEEE 2020 .
[3] Vikas S Chavan and Shylaja S S, “Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network” IEEE 2015.
[4] Nanlir Sallau Mullah and Wan Mohd Nazeem Wan Zainon, ”Advance in Machine Learning Algorithms for Hate Speech Detection in Social Media” IEEE 2021.
[5] TEOH HWAI TENG AND KASTURI DEWI VARATHAN, “Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches” IEEE 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY