IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, May 2026
Impact Factor: 8.2

CyberBulling Detection System

Prof. Vikas Gaikwad[1], Prachi Chavan[2], Anklesha Dolas[3], Sakshi Janakwar[4], Minakshi Shirsat[5]
Professor, Department of Artificial Intelligence & Data Science[1]
Students, Department of Artificial Intelligence & Data Science[2-5]
Shree Ramchandra College of Engineering, Pune, India

Abstract: *This paper presents the new threat of cyberbullying is the social media, which is largely hidden under slang, abbreviations, and multiple languages. Keywords filters are not contextual hence rendering moderation ineffective. The NLP and transformer models form the basis of the project because they are sophisticated and can detect dangerous content besides providing real-time information on ways individuals can safely communicate on the Internet. In this paper, NLP and transformer models will help to recognize and describe the instances of cyber bullying in text on social media, and a transparent, accurate, and scalable content moderation tool will be offered to identify and categorize the content in real time.*

Keywords: Cyberbullying detection, NLP, Transformer models, Explainable AI, Content moderation, Online safety

I. INTRODUCTION

Cyberbullying has emerged as a menace on the social media, so disguised by use of slangs, abbreviations and different languages. The old-fashioned system of moderation based on the use of key-words cannot realize the real context of messages, thus, being not effective in detecting the harmful messages. This project involves determining the presence of the abusive or offensive text and giving real-time feedback with the help of the state-of-the-art Natural Language Processing (NLP) and transformer-based models with Explainable AI. The system does not only recognize harmful communication, but also provides recommendations that will ensure safer online interactions. Generally, the proposed project can be viewed as an NLP-based, transformer-based system aimed at automatic detection and analysis of cyberbullying examples in the social media. It will provide a clear, precise and scaleable content moderator that has the capability to work in real-time.

II. PROBLEM STATEMENT

Cyberbullying on social networking sites is growing at a very fast pace. Manually identifying cyberbullying content is tedious and not efficient. An automatic mechanism that can effectively identify such bullying content would be helpful.

III. OBJECTIVE

• Recognize objectionable or threatening text in the online data with the help of the high-level machine learning algorithms to determine the case of cyberbullying precisely.
• Use real time identification to automatically mark abusive or harmful messages, which can be used to make the Internet a safer place.
• The model is to be trained on well-labeled datasets containing samples of bullying and non-bullying text in order to enhance its performance in regard to classifying various types of messages correctly.
• Continuous model performance is achieved through testing and refinement so that a given model can be more accurate, precise, and reliable on different datasets.
• Establish an easy-to-use interface or dashboard that will allow real-time analysis and visualization of the detected content so as to monitor and control it better.
• Make the system scalable and flexible so as to suit in various languages, platforms real-life applications.

DOI: 10.48175/IJARSCT-35374

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 6, May 2026 Impact Factor: 8.2

• To find cyberbullying in online text, this project employs natural language processing (NLP) and machine learning. Its primary objective is to automatically find and flag abusive or offensive communications published on social media or other internet channels, hence assisting in the establishment of a Safer and more courteous virtual area.
• Trained on datasets including bullying as well as non-bullying text, the system will pick up patterns and language characteristics that point to aggressive conduct. Regular reviews of its accuracy and performance will result from precise tuning with several data sources.
• The project goes beyond technical correctness to concentrate on ethical AI—that is, ensuring privacy, openness, and justice. Explainable Artificial Intelligence helps consumers to thoroughly grasp the reasons behind some communications being regarded as detrimental.
• Ultimately, this project seeks to create a scalable, clear, and responsible cyberbullying detection system that advances safer online communication and benefits everyone's digital well-being.

V. LITERATURE SURVEY

TABLE I: LITERATURE SURVEY

Sr No.	Title	Author	Year	Methodology Used	Conclusion
1.	"A Hybrid Machine Learning Model for Cyberbullyi ng Detection"	Hossein Hosseinmardi, Saberi M	2015	The system combined machine learning classifiers, feature extraction, and behavioral analysis to detect cyberbullying activities across social media platforms.	The research concluded that hybrid models improve detection accuracy and reduce false positive results in cyberbullying identification systems.
2.	"Cyberbully ing Detection on Social Media Using Machine Learning"	Zeerak Waseem, Dirk Hovy	2016	The study used machine learning algorithms such as Support Vector Machine (SVM) and Natural Language Processing (NLP) techniques to classify harmful social media comments.	The research concluded that machine learning models can effectively identify offensive and bullying content on social media platforms.
3.	"Automatio n Detection of Cyberbullyi ng in Social Media Text"	Vinita Nehar, Xue Li, Chaoyi Pang	2013	The system applied text mining, sentiment analysis, and supervised learning techniques to detect abusive and threatening messages from online communications	The study showed that sentiment-based analysis helps in identifying cyberbullying behavior and improves online safety monitoring.
4.	"Deep Learning for Cyberbullying Detection"	Pete Burnap, Matthew Williams	2015	The researchers implemented deep learning models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks for detecting bullying patterns in text data.	research concluded that deep learning approaches provide better contextual understanding and higher detection performance compared to traditional methods.
5.	"Detection of Cyberbullying on Twitter Using NLP"	Karthik Dinakar, Randal Picard	2012	The methodology focused on Natural Language Processing, keyword extraction, and topic classification to identify bullying related tweets and comments.	The study concluded that NLP-based systems can efficiently recognize harassment and abusive behavior in online social networks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, May 2026
Impact Factor: 8.2

VI. METHODOLOGY

The approach used in the cyberbullying detection model has a number of systematic steps meant to guarantee both exact and quick identification of dangerous material in online text. Data gathering, preprocessing, feature extraction, model building, assessment, and results interpretation are all part of the process:

A. Gathering Information

Collected from publicly accessible datasets or social media text data from sites like Reddit, YouTube, and Twitter. To guarantee variety and balanced training, the data comprises both bullying and non-bullying examples.

B. Preprocessing of Data

URLs, emojis, hashtags, and stopwords are eliminated to help clean collected text. To get the data ready for model training, it is then tokenized, lemmatized, and class labels are encoded.

C. Feature Extraction

TF-IDF, Word2Vec, or BERT embeddings let one turn the processed text into numerical form. This enables the model to accurately grab the significance and background of terms.

D. Model Designing

Training several Machine Learning models (SVM, Random Forest, Logistic Regression) and Deep Learning models (LSTM, CNN, BERT) helps to categorize text as either bullying or non-bullying. Every model is adjusted for best performance and precision..

E. Model Appraisal

Build a visual dashboard using PyQt6. To maintain a true voice-assistant experience, traditional text outputs are bypassed, and all system responses and confirmations are delivered vocally using Edge TTS.

F. Outcome Interpretation

Different models are evaluated to find the best way to spot cyberbullying. Understanding from this study helps to improve system dependability and detection precision.

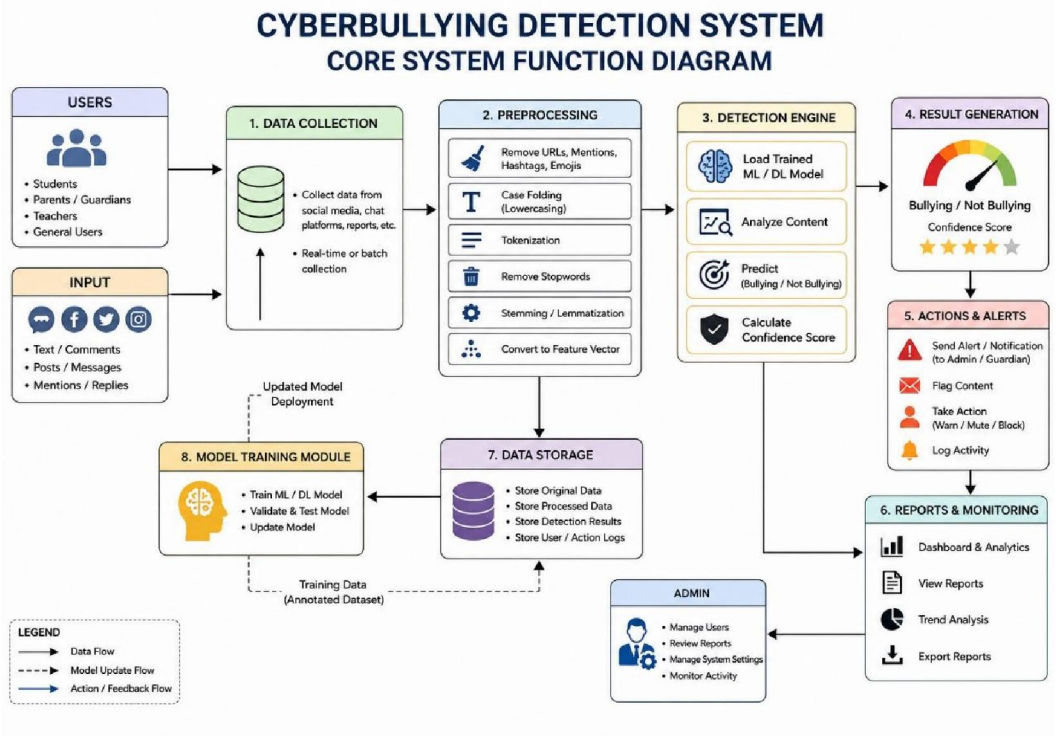

Figure 1. System Functions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, May 2026
Impact Factor: 8.2

VII. MODELING & ANALYSIS

The Cyberbullying Detection System is designed to detect abusive and harmful text messages using Machine Learning and NLP techniques. The system collects user text data, preprocesses it by removing unwanted words and symbols, and converts the text into numerical features using TF-IDF or Bag of Words methods. Machine learning algorithms such as Naive Bayes, SVM, or Logistic Regression are used to classify the text as bullying or non-bullying.

1. The system consists of the following modules:
2. Data Collection
3. Text Preprocessing
4. Feature Extraction
5. Classification
6. Result Generation

The performance of the system is evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy Formula:

$$\text{Accuracy} = \{TP + TN\}/\{TP + TN + FP + FN\}$$

VIII. SYSTEM DESIGN

The system is built on a modular architecture with the following components: The system is built on a modular architecture with the following components:

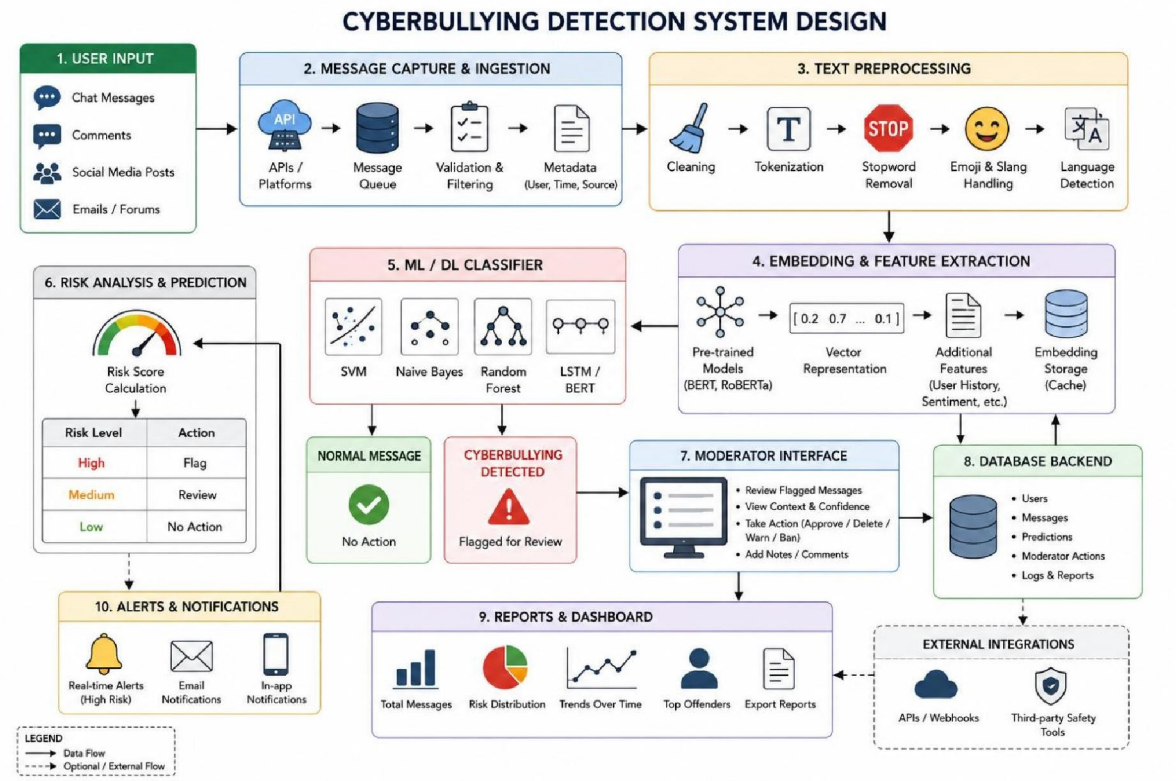

Figure 2. System Design

A. User Interface Layer: Provides voice and text communication between the user and AI assistant.
B. AI Processing Layer: Processes speech recognition, Natural Language Processing (NLP), and response generation.
C. Database Layer: Stores user data, reminders, health records, and chat logs securely.
D. Notification Layer: Sends alerts, reminders, and emergency notifications to family members and doctors.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, May 2026
Impact Factor: 8.2

E. Design Flow: User → Input → AI Processing → Response Generation → Notification/Action.

IX. RESULT & DISCUSSION

This study aimed to find out if cyberbullying may be automatically detected via the characteristics and identity of it depend on these parameters. Therefore, we applied a methodical Method of literature review gives a thorough survey of research on automated cyberbullying. detection. According on the findings of the provided review, we provide suggestions for next Study and propose enhancements to current machine learning models and classifiers in automated Detection of cyberbullying in this segment.

User Authentication Module – Login and Registration Page:

This module allows users to create accounts, authenticate securely, and access the cyberbullying detection features of the system.

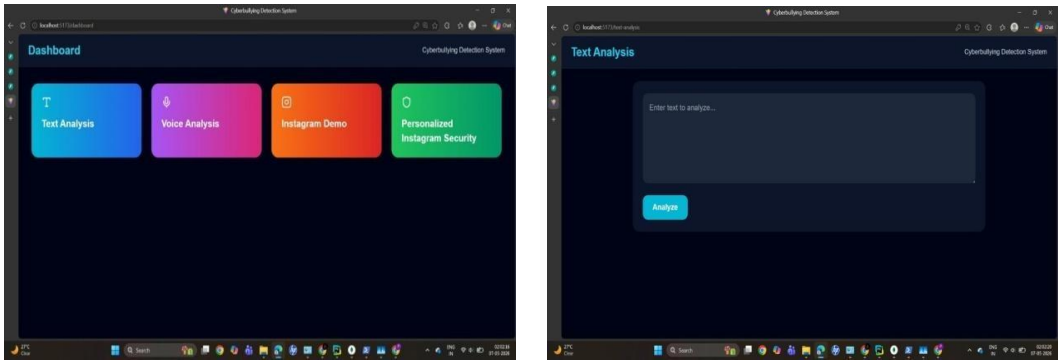

Figure 3: Cyberbullying Detection System Dashboard

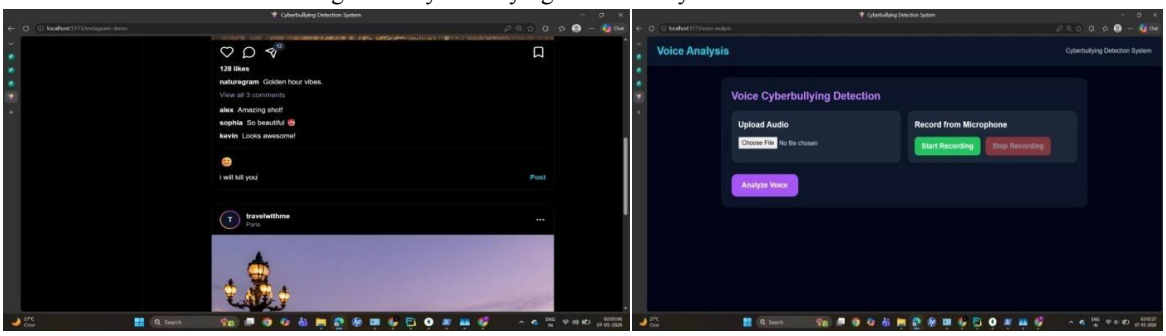

Figure 4:Cyberbulling msg/voice analysis

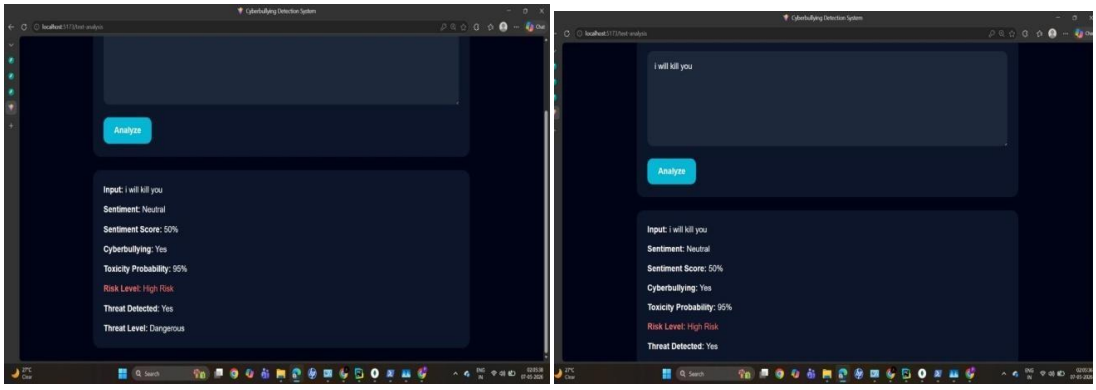

Figure 5: Result – Warning

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 6, May 2026
Impact Factor: 8.2

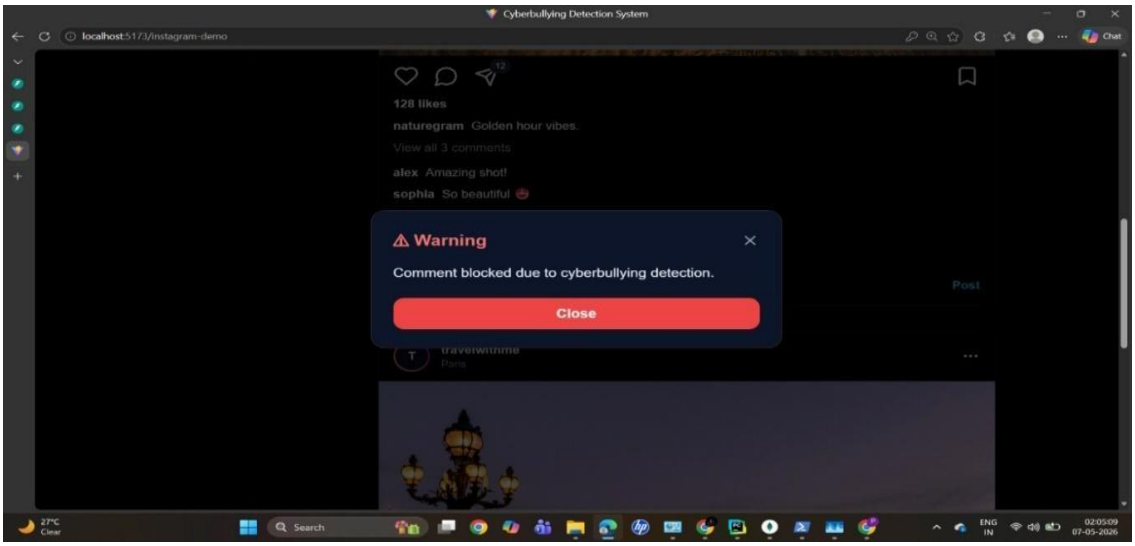

Figure 6: Result- Analysis

X. CONCLUSION

This work shows how machine learning and natural language processing (NLP) can be successfully employed to identify cyberbullying in online text. The system uses Explainable Artificial Intelligence and training of models on actual data to not only detect abusive or destructive communications but also explain why a given message is flagged. This increases the transparency and reliability of the procedure. Supporting real-time detecting and moderation of online content, the evolved model helps to build a safer digital environment. This system helps to decrease online harassment and foster good digital communication by means of constant improvement, scalability, and an emphasis on ethical artificial intelligence policies.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to everyone who supported us throughout the development of this paper. First and foremost, we are deeply thankful to our guide, whose invaluable expertise, continuous encouragement, and constructive feedback were instrumental in shaping this paper. Their guidance helped us navigate technical challenges and stay on track with our objectives. We extend our appreciation to the HOD and Faculty (Department of AI & DS Engineering) at Shree Ramchandra College of Engineering, Pune, for providing us with the necessary resources and a supportive academic environment. The curriculum and facilities have been essential to our learning and execution. We are also grateful to our family and friends for their unwavering support, patience, and motivation during this process. This paper would not have been possible without the collective effort and support of all these individuals and institutions. VIII.

REFERENCES

[1]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. https://arxiv.org/abs/1703.04009
[2]. Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. https://arxiv.org/abs/1801.04509
[3]. Vaswani, A., et al. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762
[4]. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805