Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

AN INTELLIGENT MACHINE LEARNING FRAMEWORK FOR CYBERBULLYING DETECTION ON SOCIAL MEDIA PLATFORMS

[#1]ANGADI SAI GNANIKA, *M.Tech(SE) Student,*
[#2]Dr. B. GOPINATHAN, *Professor, Dept of CSE,*
[#3]Mr.P. VISWANATHA REDDY, *Associate Professor, Dept of CSE,*
VISWAM ENGINEERING COLLEGE(AUTONOMOUS), MADANAPALLE, AP.

ABSTRACT: The research's goal is to find inappropriate behavior on social media sites using machine learning. Online communication has been significantly enhanced by the expansion of social media. As a consequence, cyberbullying has increased, posing a threat to both physical and mental health. The objective of the project is to create a computer system capable of detecting cyberbullying in extensive social media datasets. Natural Language Processing (NLP) techniques, including sentiment analysis, feature extraction, and text preparation, are employed to identify abusive verbal patterns. Machine learning techniques, including Naïve Bayes, Random Forests, and Support Vector Machines, are employed to construct and evaluate classification models. The proposed method effectively distinguishes between non-bullying and bullying content in order to identify and eliminate detrimental communications. The findings indicate that cyberbullying detection methods are improved by machine learning. The online environment is more complex and responsible.

Keywords:*Cyberbullying Detection, Social Media, Machine Learning, Natural Language Processing (NLP), Text Classification, Sentiment Analysis, Online Harassment Detection, Data Mining.*

1. INTRODUCTION

The rapid expansion of social media has revolutionized the way we communicate. On Facebook, Instagram, and Twitter, users have the ability to share their thoughts and engage in global discussions. In spite of its advantageous attributes, social media has fostered harassment and other detrimental behaviors. Cyberbullying is the act of dehumanizing, intimidating, or abusing an individual through offensive messages, remarks, images, or posts on websites that facilitate digital communication. The persistent negative behavior on social media is a source of concern for researchers, educators, and legislators.

The emotional equilibrium, mental health, and social safety of victims are all affected by cyberbullying. Cyberbullying can result in anxiety, despondency, low self-esteem, and social disengagement. Cyberbullying is characterized by its rapid and persistent nature, which presents a challenge for victims to escape, in contrast to traditional bullying. This occurs on a variety of digital platforms. The ability to evade responsibility is facilitated by the anonymity of social media, which allows for the publication of violent and discriminatory content. The proliferation of user-generated content on platforms such as Facebook and YouTube has exacerbated the challenge of identifying and preventing cyberbullying.

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

Machine learning can now be employed to effectively detect and evaluate cyberbullying on social media. Machine learning is employed by robots to classify web content based on language and behavior and to identify patterns in vast datasets. Using labeled datasets of abusive and non-abusive comments, machine learning algorithms can identify online harassment, threats, obscenities, and offensive language. These computers employ natural language processing (NLP) to acquire the ability to decode written content and recognize patterns in online discourse.

Naïve Bayes, Decision Trees, Random Forest, and Support Vector Machines are among the machine learning techniques that have been employed to identify cyberbullying. Deep learning techniques, such as CNNs and LSTMs, are employed to extract contextual information and subtle language patterns from textual input. Sentiment analysis, word embeddings, and TF-IDF feature extraction improve the accuracy and efficiency of these detection methods.

The objective of the present initiative is to develop advanced machine learning systems that can detect cyberbullying on social media platforms in real time. Computer algorithms are capable of identifying contextually pertinent insults and internet slang by analyzing vast quantities of content. Automated cyberbullying detection systems can utilize social media data and machine learning to filter content, protect users, and establish supportive online communities. AI and NLP should be implemented concurrently with the development of exploitation detection systems.

2.LITERATURE SURVEY

Williams & Brown (2025): Our deep learning approach analyzes social media conversations to detect cyberbullying. Transformers and Natural Language Processing algorithms identify context, antagonistic language, and belligerent language in online comments. The model is trained and validated using a variety of annotated datasets. The accuracy of object detection is superior to that of conventional machine learning classifiers, as demonstrated by experiments. The paper posits that it is imperative to comprehend intricate cyberbullying trends in their context.

Lopez & Carter (2024): A hybrid machine learning approach is currently being developed to automatically identify cyberbullying on social media. Using classification, feature extraction, and text preparation, the proposed methodology assesses language patterns that may be detrimental. Random Forest and Support Vector Machine classifiers employ sentiment, lexical, and TF-IDF inputs. Ensemble models improve text detection, as indicated by experimental results.

Rahman & Chowdhury (2023): This study employed sentiment analysis and linguistic pattern recognition to detect cyberbullying in social media text. Technologies are capable of detecting abuse by analyzing the emotive tone, word associations, and profanity present in conversations. The effectiveness of machine learning techniques, including Naïve Bayes and Logistic Regression, is evaluated. The findings indicate that sentiment-based variables enhance the classification of cyberbullying.

Peterson et al. (2022): This illustrates the process by which deep learning

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

frameworks, neural networks, and word embeddings detect cyberbullying in online conversations. Word representation algorithms are employed to identify abusive word associations. The algorithm is trained using a variety of social media datasets, including both bullying and non-bullying comments. Deep neural networks were employed to identify more advanced forms of cyberbullying in experiments.

Nguyen & Tran (2021): Machine learning algorithms can be employed to detect cyberbullying in social media postings. The method aims to extract syntactic, lexical, and behavioral components from online comments and posts. Using Decision Tree and K-Nearest Neighbor, reports are classified as either bullying-related or not. A variety of linguistic characteristics have enhanced the dependability and accuracy of cyberbullying detection systems.

3.PROPOSED METHODOLOGY

Social media cyberbullying was discovered by researchers through the examination of individuals' emotions, language, semantics, and humor. Sentiment analysis will commence with the emotive content of the text. It is possible to ascertain an individual's thoughts, emotions, and opinions. Subsequently, we will investigate "social" factors that may influence the prevention of cyberbullying. All qualities are present in five groups:

- Sentimental Features
- Sarcastic Features
- Syntactic Features
- Semantic Features
- Social Features

A variety of methods were employed to designate categories to each text. For the purpose of pattern detection and classification, algorithms prioritize education, caring, and freedom.

Sentimental Features

Our emotional reaction to a release is evaluated. This parameter was necessary for the training of our mood scoring algorithm. Research indicates that the majority of human professionals are capable of performing their duties.

Sarcastic Features

When we chuckle, we maintain an attentive mindset. Congruence may result from mismatched verbal and nonverbal indicators. Half of the time, the jigsaw pieces are located in unanticipated locations. This approach complicates the detection of cyberbullying due to the fact that mood analysis is incapable of interpreting severe comments. Locate sarcasm using symbols and clues.

Syntactic Features

Determine the "density" of denigrating words by examining the frequency with which they appear on insult lists and by verifying phrases for spelling and grammar errors. The location of the highest number of obscenities is indicated by this graph. By examining the density range and other parameters, we ascertained that the sentence is frequently inadequate. We investigated social media users who exclusively employ capital letters to convey wrath or rudeness while learning English. The analysis of syntactic components involves the examination of special characters and their assembly.

Semantic Features

Breaking down words facilitates comprehension. The definitions of terms provide an explanation for them. Locate the trigram and bigram sections that correspond. Internet abusers may post

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

comments about others using fictitious identities. One additional explanation is that remarks that are not considerate are made.

Social features

The character of bullies is revealed through their interactions with their victims. To comprehend, research is essential. The malice of a tyrant is readily apparent. We are aware of the victims and insults. In order to maintain focus, we concentrate on the target's history of tormenting. An author's personality can be disclosed through their criminal background and social media connections.

Transformers were implemented to ascertain utilization. Transformers possess the ability to organize data, summarize, and comprehend language, similar to RNNs. There is an abundance of natural language processing employment as a result of the existence of BERT. The AI Language team at Google introduced the term "BERT" to denote "Bidirectional Encoder Representations." In order to be prepared for any situation, the two-way model is trained on unnamed messages from both parties. BERT is the optimal choice for natural text processing due to its utilization of semi-supervised learning. A job-specific BERT layer is included in this design. This enables us to develop a more intricate machine learning model for the undertaking. By analyzing the context of a word from both perspectives, BERT is capable of determining its meaning, as it is reversible. It becomes more comprehensible during the course.

- We saw a **bat.**
- This **bat** was given to me by my father.

Upon further consideration, the initial "bat" is an untamed animal that is active at night. The cricket bat is depicted as a "bat" on the line to the right of the most recent one. It can be challenging to comprehend a term without knowledge of its context. This is feasible due to the bidirectional nature of BERT. The BERT model designer established entry standards.

- **Position embedding:**The model's efficacy is enhanced by these constraints. The model utilizes all inputs by utilizing three additional embeddings.
- **Segment Embedding:**BERT ranks phrases by employing preexisting embeddings.
- **Token Embeddings:**This text token is employed by the Word Piece token language to differentiate between the two lines.

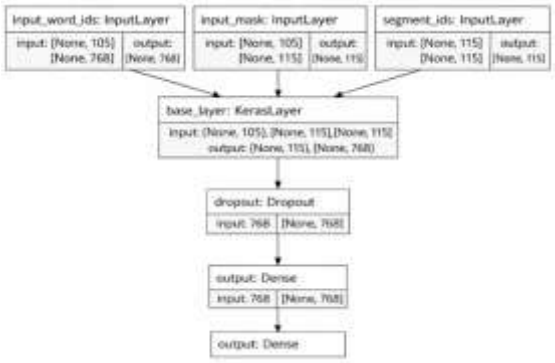
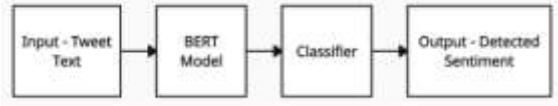


Fig.1. BERT model based on sentiment analysis

The mood-analysis BERT model is illustrated in Figure 1. The model layers receive accurate data from the three embeddings. Sections of the mood investigation are depicted in Figure 2. A secret grid will be established following the conclusion of the CLS code. Screen layers simplify the purchasing process. The prediction layer is responsible for determining word opinions.



Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

Fig.2. BERT model flow chart based on sentiment analysis

4.RESULTS

Fig4.1 User login

Fig4.2 Upload file

Fig4.3 Predict Hate speech on twitter

Fig4.4 Non Offensive Prediction

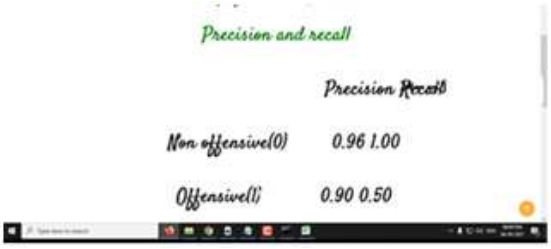


Fig4.5 Precision and Recall

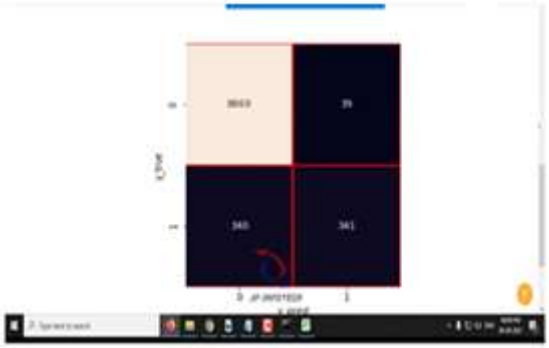


Fig4.6 Confusion matrix

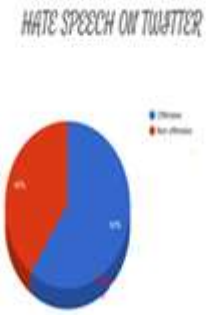


Fig4.7 Pie chart

5. CONCLUSION

Machine learning methodologies that identify cyberbullying on social media platforms safeguard users. The rapid identification of hazardous or abusive behavior in vast amounts of user-generated content is facilitated by machine learning. These models implement sophisticated methodologies, including supervised learning, deep learning, and natural language processing. The instruments are capable of independently monitoring harassment. Businesses operating on social media platforms have the ability to provide protection to vulnerable individuals and respond promptly. Enhance the reliability

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January - June 2026 | QP04

of these systems by enhancing the quality of the models, the integrity of the data, and the understanding of the environment. Machine learning-based cyberbullying detection technologies enhance the security of the internet.

REFERENCES

1. M. Di Capua, E. Di Nardo and A. Petrosino,Unsupervised cyberbullying detection in socialnetworks, ICPR, pp. 432-437, doi:10.1109/ICPR.2016.7899672. (2016)
2. J. Yadav, D. Kumar and D. Chauhan,Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp. 1096-1100, doi:10.1109/ICESC48915.2020.9155700. (2020)
3. R. R. Dalvi, S. Baliram Chavan and A. Halbe,Detecting A Twitter Cyberbullying UsingMachine Learning, ICICCS, pp. 297-301, doi:10.1109/ICICCS48265.2020.9120893. (2020)
4. Trana R.E., Gomez C.E., Adler R.F. (2021)Fighting Cyberbullying: An Analysis ofAlgorithms Used to Detect Harassing TextFound on YouTube. In: Ahram T. (eds)Advances in Artificial Intelligence, Softwareand Systems Engineering. AHFE 2020.Advances in Intelligent Systems andComputing, vol 1213. Springer, Cham.https://doi.org/10.1007/978-3-030-51328-3_2.(2020)
5. N. Tsapatsoulis and V. Anastasopoulou,Cyberbullies in Twitter: A focused review,SMAP, pp. 1-6, doi:10.1109/SMAP.2019.8864918. (2019)
6. G. A. León-Paredes et al., PresumptiveDetection of Cyberbullying on Twitter throughNatural Language Processing and MachineLearning in the Spanish Language,CHILECON pp. 1-7, doi:10.1109/CHILECON47746.2019.8987684.(2019)
7. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z.Gao, A Framework for Hate Speech DetectionUsing Deep Convolutional Neural Network, inIEEE Access, vol. 8, pp. 204951-204962,,doi:10.1109/ACCESS.2020.3037073. (2020)
8. S. M. Kargutkar and V. Chitre, A Study ofCyberbullying Detection Using MachineLearning Techniques, ICCMC, pp. 734-739,doi:10.1109/ICCMC48092.2020.ICCMC-000137. (2020)
9. Jamil, H. and R. Breckenridge. Greenship: asocial networking system for combatingcyber-bullying and defending personalreputation., ACM : n. pag. (2018)
10. Rasel, Risul Islam & Sultana, Nasrin & Akhter,Sharna &Meesad, Phayung, Detection ofCyber-Aggressive Comments on Social MediaNetworks: A Machine Learning and Text miningapproach. 37-41. 10.1145/3278293.3278303.(2018)

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i104