isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Mar,2026 ,ISSN_NO: 2321-3337

CYBERBULLYING DETECTION IN SOCIAL MEDIA USING HYBRID LEARNING APPROACHES

Mohan G[1], N Vishnu[2], C J Santhosh Kumar[3], Akash P[4]

[1234] UG Scholar-Dept CSE, Grt Institute of Engineering and Technology Tiruttani, India.

[5*]Assistant Professor-Dept CSE, Grt Institute of Engineering and Technology Tiruttani, India.

mohancse2022@grt.edu.in , vishnucse2022@grt.edu.in,

santhoshkumarcse2022@grt.edu.in , akashcse2022@grt.edu.in

[5*]Corresponding Author:

Abstract

Cyberbullying has become a major concern in online communication platforms due to the increasing usage of social media networks. Harmful comments, abusive language, and threatening messages can negatively affect individuals and communities. Manual monitoring of such content is difficult because of the large volume of data generated every day. Therefore, an automated system is required to identify cyberbullying activities effectively.This paper proposes a cyberbullying detection system that uses machine learning techniques to analyze textual data and identify harmful messages. The dataset used in this research contains labeled bullying and non-bullying text collected from online sources. The data is preprocessed and transformed into suitable features for classification. The classification process is performed using the Support Vector Machine algorithm, which is widely used for text classification problems. The system automatically detects abusive messages and allows administrators to take appropriate actions such as flagging or deleting harmful content. The proposed system improves the efficiency of cyberbullying detection and helps in maintaining a safer online environment.

Keywords: *Cyberbullying Detection, Machine Learning, Text Classification, Natural Language Processing, Support Vector Machine.*

1. Introduction

With the rapid development of internet technologies and social media platforms, online communication has become an essential part of daily life. People use social networking sites to interact, share opinions, and exchange information. However, the increased use of these platforms has also resulted in the growth of cyberbullying activities. Cyberbullying refers to the act of harassing, threatening, or insulting individuals through digital platforms such as social media, messaging applications, and online forums. Victims of cyberbullying may experience psychological stress, anxiety, and emotional distress. Due to the massive amount of content generated on social media platforms, it becomes difficult for administrators to monitor and control abusive messages manually.Machine learning techniques provide an efficient solution for automatically identifying harmful content in online communication. By analyzing textual patterns and linguistic features, machine learning models can classify messages into bullying and non-bullying categories. In this research, a cyberbullying detection system is developed using a supervised machine learning approach. The system analyzes user-generated text and detects harmful messages using the Support Vector Machine classification algorithm.

2. Related Work

With the rapid growth of the internet and social media platforms, cyberbullying has emerged as a major issue in modern digital communication. Millions of users interact daily through social networking websites, online forums, gaming platforms, and messaging applications. While these technologies provide various benefits such as fast communication, knowledge sharing, and social connectivity, they also create opportunities for harmful behaviors such as harassment, abusive messaging, and online bullying. Cyberbullying can cause serious psychological and emotional effects on individuals, especially among teenagers and young users who are highly active on digital platforms. Due to the increasing number of cyberbullying incidents reported on social media platforms, researchers have focused on developing automated systems capable of detecting and

controlling such harmful content. Early research in this area primarily relied on rule-based methods and keyword filtering techniques to detect offensive language in online messages. These approaches identified predefined abusive words or phrases within the text. Although these systems were able to detect explicit offensive terms, they were limited in their ability to understand the context and meaning of sentences, making them ineffective for identifying complex forms of harassment or indirect bullying behavior. As a result, researchers began exploring machine learning techniques to improve the performance and reliability of cyberbullying detection systems. [1]

One important research direction involves the use of supervised machine learning algorithms for classifying online messages into bullying and non-bullying categories. In these approaches, a labeled dataset containing examples of abusive and non-abusive text is used to train classification models. During the training process, the models learn patterns and linguistic characteristics associated with cyberbullying messages. Researchers have applied various classification algorithms such as Naïve Bayes, Decision Trees, Logistic Regression, and Support Vector Machines for this task. These algorithms analyze textual features such as word frequency, sentence patterns, grammar structure, and semantic relationships between words. Once trained, the models can automatically classify new incoming messages and determine whether they contain cyberbullying content. Experimental studies have shown that machine learning- based detection systems perform significantly better than traditional keyword filtering approaches because they can recognize patterns in language rather than relying only on predefined abusive words. [2]

Another important research area focuses on the application of Natural Language Processing (NLP) techniques for analyzing textual data in cyberbullying detection systems. NLP enables computers to understand and process human language by extracting meaningful information from textual data. In cyberbullying detection tasks, NLP techniques are commonly used during the preprocessing stage to prepare the text data for machine learning algorithms. These techniques include tokenization, which breaks sentences into individual words or tokens; stop-word removal, which eliminates common words that do not contribute significant meaning; stemming and lemmatization, which reduce words to their base forms; and normalization processes that standardize textual data. By applying these preprocessing steps, raw text data can be transformed into a structured format that is easier for machine learning models to analyze. Several studies have demonstrated that combining NLP techniques with machine learning algorithms significantly improves the accuracy and efficiency of cyberbullying detection systems. [3]

In addition to text preprocessing, researchers have also investigated the role of sentiment analysis in detecting cyberbullying behavior. Sentiment analysis is a technique used to determine the emotional tone or polarity of a text message by analyzing the words and expressions used within it. Cyberbullying messages often contain strong negative sentiments such as anger, hatred, threats, or aggressive language. By analyzing sentiment scores and emotional indicators in text data, detection systems can identify messages that potentially contain harmful intent. Sentiment analysis can also help distinguish between neutral conversations and abusive interactions in social media discussions. Integrating sentiment analysis with machine learning classifiers provides additional contextual information that improves the detection of offensive content. This approach has been widely used to analyze user-generated content such as social media posts, online comments, and chat conversations across different digital platforms. [4]

Another key aspect of cyberbullying detection research involves the use of large-scale datasets collected from social media platforms. The availability of large textual datasets allows machine learning algorithms to learn more complex patterns associated with abusive language. However, working with large datasets requires effective preprocessing and feature extraction methods to ensure that relevant information is captured accurately. Researchers have emphasized the importance of feature extraction techniques such as Term Frequency (TF), Term Frequency–Inverse Document Frequency (TF-IDF), and word embedding models. These techniques convert textual information into numerical feature vectors that represent the importance of words within a

document or message. Machine learning algorithms then use these feature vectors to perform classification tasks. Proper feature representation plays a crucial role in improving the performance and accuracy of cyberbullying detection systems. Studies have shown that well-prepared datasets combined with effective feature extraction techniques can significantly enhance the performance of automated detection models. [5]

Several studies have also focused on comparing the performance of different machine learning algorithms in cyberbullying detection tasks. Among the various classification techniques, the Support Vector Machine (SVM) algorithm has been widely used due to its strong performance in text classification problems. SVM is particularly effective when dealing with high-dimensional datasets where the number of features is large. The algorithm works by mapping input data into a higher- dimensional feature space and identifying an optimal hyperplane that separates different classes with maximum margin. This ability allows SVM to create clear boundaries between bullying and non-bullying messages, making it highly suitable for cyberbullying detection systems. Experimental evaluations conducted in various studies have demonstrated that SVM- based models achieve high accuracy, precision, and recall in detecting abusive content in online communication platforms. [6]

Despite the significant progress achieved in cyberbullying detection research, several challenges still remain. Many existing systems struggle to identify subtle forms of harassment such as sarcasm, indirect bullying, and context-dependent abusive language. Online communication also evolves rapidly, introducing new slang terms, abbreviations, and expressions that may not be included in existing training datasets. These factors can reduce the effectiveness of automated detection systems over time. Therefore, researchers continue to explore advanced machine learning models, improved preprocessing techniques, and larger training datasets to enhance the adaptability and performance of cyberbullying detection systems. The proposed system in this study contributes to this research area by implementing an automated cyberbullying detection framework that uses machine learning techniques to classify online messages and assist administrators in monitoring and controlling harmful content on digital platforms.

3. Objective

The primary objective of this project is to develop an automated system that can detect cyberbullying content in online communication platforms. The system analyzes textual data from user messages and classifies them as bullying or non-bullying using machine learning techniques. By applying text preprocessing, feature extraction, and classification methods, the system aims to accurately identify abusive or harmful language in digital conversations. Another objective is to assist administrators in monitoring and controlling inappropriate content on online platforms by automatically flagging harmful messages for review and removal. This reduces the need for continuous manual monitoring and improves the efficiency of content moderation. Overall, the project aims to create a safer online communication environment by reducing the spread of abusive language and minimizing the negative impact of cyberbullying on users.

4. Proposed System

The proposed system introduces an automated cyberbullying detection framework that analyzes user-generated messages and identifies harmful content using machine learning techniques. The system uses a dataset containing labeled bullying and non-bullying messages for training the classification model.The collected data is first processed using several preprocessing techniques such as removing unnecessary characters, converting text to lowercase, and eliminating stop words. These preprocessing steps improve the quality of the dataset and enhance the performance of the classification model.After preprocessing, relevant textual features are extracted from the dataset and used to train the machine learning classifier. The system uses the Support Vector Machine algorithm to classify messages based on their textual patterns. When a user submits a message, the trained model analyzes the text and predicts whether the message contains cyberbullying content. If bullying content is detected, the system alerts the administrator so that appropriate action can be taken.

5. Architecture Diagram

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Mar,2026 ,ISSN_NO: 2321-3337

The architecture of the proposed system is designed as a streamlined pipeline that begins at the user interface layer, where raw text and image inputs are captured via Flask and routed to the preprocessing module for noise reduction, stop-word removal, and lemmatization using NLTK. Following this, the normalized text enters the feature extraction stage, where it is converted into high-dimensional numerical vectors through TF-IDF vectorization, allowing the Support Vector Machine classification engine to analyze linguistic patterns and categorize the content as safe or bullying. The final decision and enforcement stage executes automated protocols based on this classification: safe content is committed to the public posts database, while flagged messages are intercepted, logged in the warning database to update the user's strike count, and simultaneously displayed on the global surveillance feed for real-time administrative oversight and manual verification.

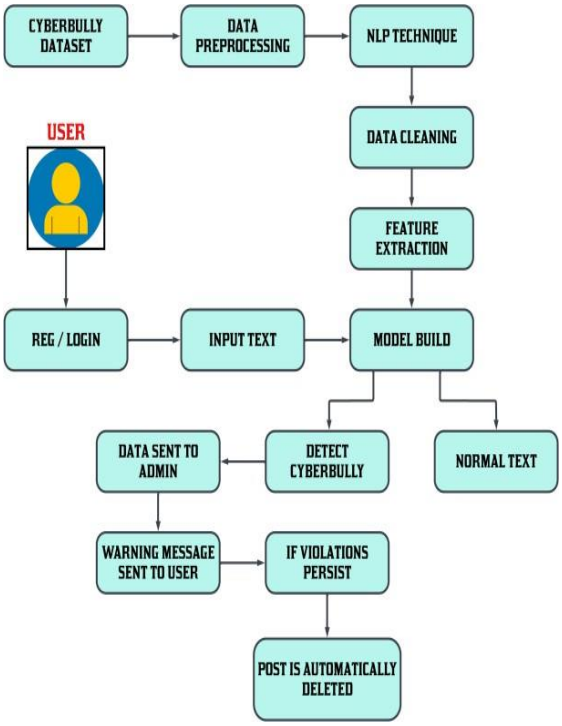
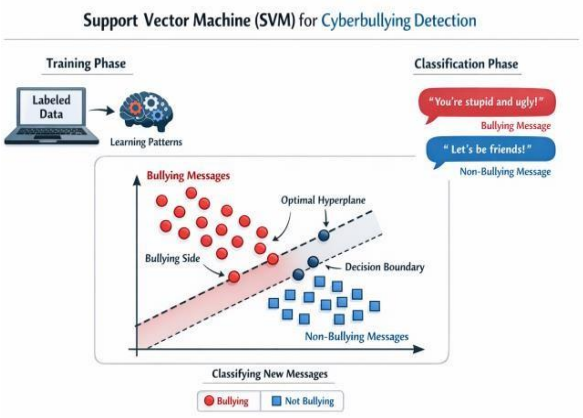


Fig.5.1: Architecture Diagram

6. Algorithm

The classification process in the proposed system is performed using the Support Vector Machine **(SVM)** algorithm. SVM is a supervised machine learning algorithm widely used for classification tasks.The main objective of the SVM algorithm is to find an optimal hyperplane that separates data points belonging to different classes. In the case of cyberbullying detection, the algorithm separates bullying messages from non-bullying messages based on textual features.During the training phase, the SVM model learns patterns from labeled datasets. After training, the model can classify new messages by determining which side of the decision boundary the message belongs to. Due to its ability to handle high-dimensional data, SVM is well suited for text classification problems such as cyberbullying detection.



7. Implementation

7.1 Data Collection Module

The data collection module is responsible for obtaining the dataset required for training and testing the cyberbullying detection system. The dataset contains two categories of text data: bullying text and non-bullying text. This dataset is collected from publicly available online sources and repositories such as social media comment datasets and online text corpora. The collected data provides examples of abusive and normal messages, which helps the system learn patterns associated with cyberbullying. The dataset is then stored in a structured format so that it can be used for training the machine learning model

7.2 Data Preprocessing Module

The data preprocessing module prepares the raw text data for further analysis. Text data collected from online sources usually contains noise such as punctuation marks, special characters, numbers, and unnecessary symbols that may affect the performance of the classification model. In this stage, the text is cleaned by removing unwanted characters and converting all words into a consistent format such as lowercase letters. Stop words that do not contribute meaningful information, such as “is”, “the”, and “and”, are also removed. This process improves the quality of the dataset and ensures that the machine learning model focuses only on relevant information.

7.3 Feature Extraction Module

The feature extraction module converts the cleaned text data into numerical representations that can be understood by machine learning algorithms. Since machine learning models cannot directly process textual data, the text must be transformed into a vector format. Techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) are used to measure the importance of words within the dataset. This technique assigns weights to words based on their frequency and relevance in the text. The resulting feature vectors represent the textual information in a numerical format, which can then be used as input for the classification model.

7.4 Classification Module

The classification module is responsible for detecting whether the given input text contains cyberbullying content. In this system, a machine learning algorithm called Support Vector Machine is used for text classification. The Support Vector Machine model is trained using the extracted features from the dataset. Once the training process is completed, the model can analyze new input messages and classify them into two categories: cyberbullying text or normal text. The classification results help in identifying harmful content posted by users on the platform.

7.5 Administration Module

The administration module provides monitoring and control functionalities for managing detected cyberbullying content. When the system identifies a message as cyberbullying, the information is sent to the administrator for review. The administrator can monitor flagged messages and take appropriate actions to maintain a safe environment on the platform. These actions may include sending warning notifications to the user, removing harmful content, or blocking repeated offenders. This module ensures that cyberbullying activities are controlled effectively and that users follow appropriate online behavior.

8. Experimental Results

This section addresses the implementation of the proposed cyberbullying detection system using machine learning-based text analysis. Various scenarios have been considered, and the figures below—Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5, and Fig. 6—illustrate the workflow of the system from user interaction to administrative actions and final results. Specifically, Fig. 1 shows the dashboard or application frontend, providing an overview of system functionality. Fig. 2 presents the user login interface, while Fig. 3 depicts the process of users submitting text messages into the system. Fig. 4 illustrates the administrator login interface, Fig. 5 demonstrates the administrative actions for monitoring and handling flagged messages and Fig. 6 shows the account blocking after repeated abuses

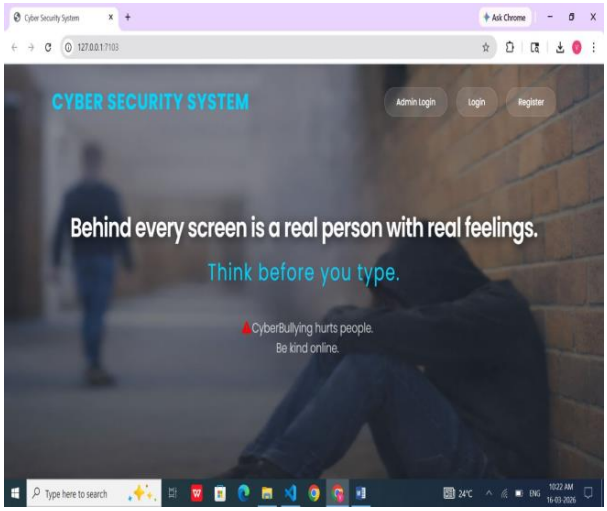


Fig.8.1. Application Frontend

isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6, Issue: 3,Special Issue: 3 ,Mar,2026 ,ISSN_NO: 2321-3337

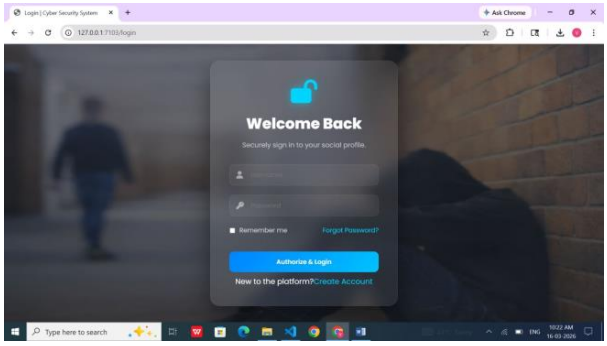


Fig.8.2. User Login

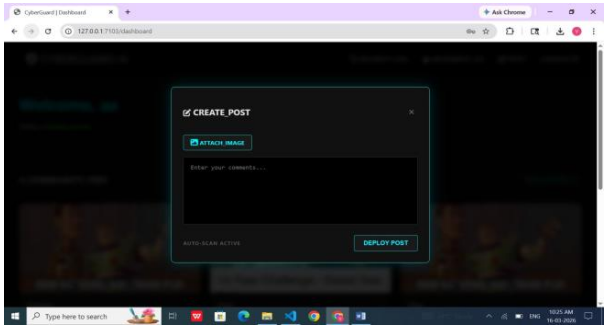


Fig.8.3. User Input

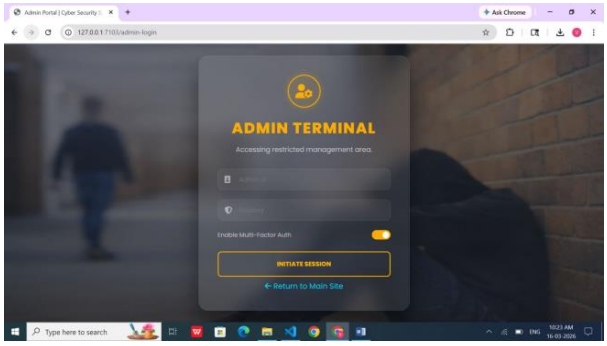


Fig.8.4. Admin Login

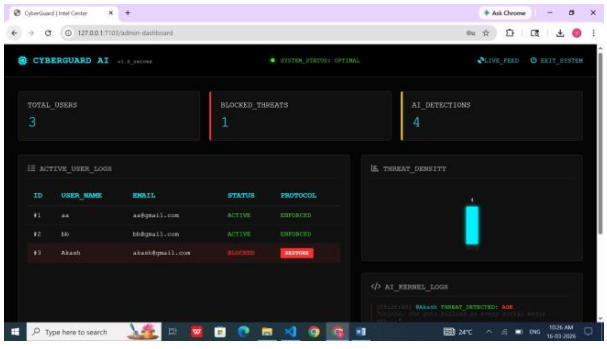


Fig. 8.5. Admin panel

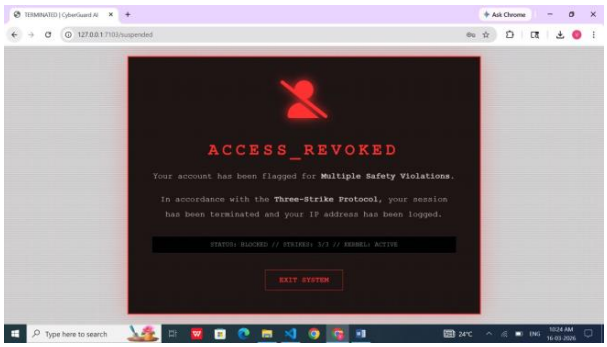


Fig.8.6. Result (Account Blocked)

9. Conclusion & Future work

The proposed cyberbullying detection system analyzes user-generated text data from online platforms to identify harmful and abusive content. By processing and comparing textual patterns from a prepared dataset with user input messages, the system predicts whether the content contains cyberbullying behavior. The application utilizes machine learning techniques to automatically detect offensive language and assist administrators in maintaining a safer online communication environment.

This system helps reduce the spread of harmful messages by enabling early detection and administrative monitoring of cyberbullying activities. The platform can be extended to monitor multiple social media environments and online communities, making it useful for protecting users across different digital spaces. In addition, the collected data can be securely stored for future analysis and system improvement, helping researchers and administrators develop more effective cyberbullying prevention strategies.

Reference

[1] T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches," in IEEE Access, vol. 11, pp. 55533-55560, 2023, doi:10.1109/ACCESS.2023.3275130.

[2] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying

Detection in Twitter Social Media Platform," in IEEEAccess,vol.10,pp.2585725871,2022,doi:10.1109/ACCESS.2022.3153675.

[3] M. H. Obaid, S. K. Guirguis and S. M. Elkaffas, "Cyberbullying Detection and Severity Determination Model," in IEEE Access, vol. 11, pp. 97391-97399, 2023, doi: 10.1109/ ACCESS. 2023.3313113

[4] M. Al-Hashedi, L. -K. Soon, H. -N. Goh, A. H. L. Lim and E. -G. Siew, "Cyberbullying Detection Based on Emotion," in IEEE Access, vol. 11, pp. 53907-53918, 2023, doi: 10.1109/ ACCESS .2023.3280556.

[5] J. Bacha, F. Ullah, J. Khan, A. W. Sardar and S. Lee, "A Deep Learning-Based Framework for Offensive Text Detection in Unstructured Data for Heterogeneous Social Media," in IEEE Access, vol. 11,pp.124484-124498, 2023, doi: 10.1109/ ACCESS.2023.3330081.

[6] P. K. Roy and F. U. Mali, ‘‘Cyberbullying detection using deep transfer learning,’’ Complex Intell. Syst., vol. 8, pp. 5449–5467, May 2022

[7] N. Vishwamitra, H. Hu, F. Luo, and L. Cheng, ‘‘Towards understanding and detecting cyberbullying in real-world images,’’ in Proc. Netw. Distrib. Syst. Secur. Symp., Jan. 2021, pp. 1–18

[8] G. Jacobs, C. Van Hee, and V. Hoste, ‘‘Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?’’ Natural Lang. Eng., vol. 28, no. 2, pp. 141–166, Mar. 2022

[9] M. Gada, K. Damania, and S. Sankhe, ‘‘Cyberbullying detection using LSTM-CNN architecture and its applications,’’ in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2021, pp. 1–6

[10] H. H.-P. Vo, H. Trung Tran, and S. T. Luu, ‘‘Automatically detecting cyberbullying comments on online game forums,’’ in Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF), Aug. 2021, pp. 1–5

[11] R. Zhao, A. Zhou, and K. Mao, “Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features,” in *Proceedings of the International Conference on Distributed Computing Systems Workshops*, Atlanta, GA, USA, 2016, pp. 43–48.

[12] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” in *Proceedings of the International AAAI Conference on Web and Social Media*, Barcelona, Spain, 2011, pp. 11–17.

[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” in *Proceedings of the International AAAI Conference on Web and Social Media*, Montréal, Canada, 2017, pp. 512–515.

[14] V. Nahar, X. Li, and C. Pang, “An Effective Approach for Cyberbullying Detection,” *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238–247, 2013.

[15] S. Agrawal and A. Awekar, “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,” in *Proceedings of the European Conference on Information Retrieval*, Grenoble, France, 2018, pp. 141–153.

[16] Y. M. Ibrahim, R. Essameldin, and S. M. Saad, “Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty,” in *IEEE Access*, vol. 12, pp. 59474–59484, 2024, doi: 10.1109/access.2024.3393295.