International Journal of Innovative Research in Science and Engineering
Volume No. 12, Issue No. 03, Year 2026
ISSN : 2454-9665

Recognition and Classification of Cyberbullying on Social Media with Machine Learning Techniques

K. Naga Sindhura[1], K. Kiran[2], MD. Yaseen[3], K. Narashima Naik[4]
Department of Computer Science and Engineering
Tirumala Engineering College

Abstract— The rise of social media platforms has brought about new challenges, one of the most concerning being cyberbullying. With millions of users posting content daily, manually identifying harmful interactions is both time-consuming and inefficient. This paper explores the use of machine learning techniques to automatically detect and classify instances of cyberbullying within social media content. By employing natural language processing (NLP) and sentiment analysis, the study analyzes textual data to identify toxic language, aggressive behaviors, and harmful interactions. Cyberbullying is a growing problem on social media, affecting many people through online harassment, hate speech, and personal attacks. To help detect and prevent such harmful content, this project uses Machine Learning and Natural Language Processing (NLP) techniques.

Keywords— *Cyberbullying Detection, Machine Learning, Natural Language Processing, Social Media, Text Classification, Hate Speech Detection.*

I. INTRODUCTION

Cyberbullying on social media has become a major concern, impacting users' mental health and safety. Machine learning techniques, including NLP help in detecting abusive and harmful content automatically. The challenge lies in handling informal language, sarcasm, and evolving slang used in online interactions. Automated detection systems can aid social media platforms in mitigating cyberbullying and ensuring a safer digital environment.

With the rapid growth of the internet and social media, cyberbullying has emerged as a significant concern, affecting both teenagers and adults. Online harassment, hate speech, and personal attacks have led to serious consequences, including depression and suicide. Addressing this issue requires efficient detection mechanisms to regulate harmful content on digital platforms

This project focuses on detecting cyberbullying using Natural Language Processing (NLP) and Machine Learning (ML) techniques. By analyzing text data from Twitter (hate speech) and Wikipedia forums (personal attacks), the study aims to classify whether a given text contains cyberbullying content. Three feature extraction methods— Bag of Words, TFIDF, and Word2Vec—are employed, along with four classificationalgorithms:

Support Vector Machines (SVM), Logistic Regression, Random Forest, and Multi-Layered Perceptrons (MLP)

II. LITERATURE SURVEY

The automatic detection of cyberbullying on social networks involves the use of advanced algorithms and machine learning techniques to identify harmful, abusive, or bullying behavior in online interactions. This process is vital due to the growing prevalence of cyberbullying, which can negatively affect individuals' mental health, particularly among adolescents and young adults. Here's a breakdown of what this type of system entails Automatic detection of cyberbullying on social networks based on bullying features leverages modern AI techniques to identify harmful behaviors in real- time, offering the potential to reduce cyberbullying and protect users. However, the system must be continuously refined and balanced with ethical considerations to avoid misidentifications and respect user privacy.

Cyberbullying detection has become an increasingly important task due to the rise of online harassment on social media platforms. Traditional methods of detecting cyberbullying, which often rely on keyword filtering or basic rule-based systems, are becoming insufficient due to the complexity, evolving language, and nuances of online communication. Social network mining techniques present a more sophisticated approach to tackling the issue, by leveraging data from user interactions, relationships, and behaviors within social networks to detect bullying behavior. Social network mining involves analyzing the structure, dynamics, and interactions of individuals within a social network to uncover patterns and insights. In the context of cyberbullying detection, social network mining techniques can extract useful features from both social interactions and textual content. These features can then be used to identify harmful behaviors such as harassment, exclusion, trolling, and other forms of cyberbullying
.
Cyberbullying is a growing concern on social networks, with individuals, especially adolescents, suffering from its harmful effects. The ability to automatically detect cyberbullying in online environments is crucial for mitigating its impact. Traditional methods for cyberbullying detection often rely on simple keyword filtering or rule-based systems, which are insufficient for dealing with the complexities and nuances of onlinecommunication.

Cyberbullying is a serious problem on social media platforms, and detecting harmful content automatically is crucial to ensure the safety of online communities. Traditional methods of detecting cyberbullying often rely on predefined rules or keywordbased systems, but these approaches tend to be limited in their accuracy and ability to understand the context of language. BERT (Bidirectional Encoder Representations from Transformers), a powerful pre- trained deep learning model, has revolutionized natural language processing (NLP) tasks, including text classification, by capturing complex contextual relationships in language. Using BERT for cyberbullying detection leverages its ability to understand the nuanced meanings of words in context, making it an effective tool for identifying abusive behavior in online communications.

III. Comparison With Previous Methodology

Keyword-based filtering is another approach used in the existing system. This method identifies cyberbullying by matching words in the text against a predefined list of offensive or abusive terms. While this approach is simple and easy to implement, it has several limitations. It cannot understand the context in which a word is used. For example, certain words may be offensive in one context but harmless in another. Additionally, users often use slang, abbreviations, or misspellings to bypass these filters. Another major limitation of the existing system is its inability to detect implicit cyberbullying.

Cyberbullying is not always direct; it can involve sarcasm, irony, or indirect insults. Keyword-based systems fail to recognize such complex patterns in language. As a result, many harmful messages go undetected. The existing system also suffers from scalability issues. As the number of users on social media platforms continues to grow, the amount of data increases exponentially. Manual and rule-based systems are not capable of handling such large-scale data efficiently. Furthermore, the response time of the existing system is relatively slow.

Since it relies on user reports and manual review, there is a delay in identifying and removing harmful content. During this time, the content may continue to spread and cause harm. The existing system for detecting cyberbullying primarily relies on manual monitoring and basic filtering techniques.

Social media platforms depend on users to report abusive content, which is then reviewed by human moderators. In some cases, simple keywordbased filtering systems are used to detect offensive words. Manual moderation involves reviewing reported content and taking appropriate actions such as removing posts or banning users. Although this method ensures human judgment, it is highly inefficient due to the massive volume of data generated on social media platforms every second. Millions of posts, comments, and messages are shared daily, making it impossible for human moderators to analyze all content effectively

Table 1. Comparison Table

Aspect	Previous Methodology (Single-Task Models)	Proposed Methodology (Multi-Task Model)
Cyberbullying detection	Performs only keyword matching or simple abuse detection as a standalone task	Simultaneously performs cyberbullying classification, toxicity detection, and contextual feature learning
Model Architecture	Uses traditional machine learning	Uses advanced machine learning models such as SVM, Logistic Regression, Random Forest with optimized training
Context Understanding	Limited understanding of sarcasm, slang, and sentence context	Learns complex linguistic patterns and contextual relationships
Input Handling	Primarily text-based static input	Supports dynamic social media text from comments, posts, and chats
Adaptability	Static behavior with limited real-time learning	Adapts to different bullying patterns across platforms
Recommendation Logic	Separate or rule-based recommendation process	Integrated mood-aware music recommendation pipeline
Accuracy	Lower accuracy (around 70%–78%)	Higheraccuracy (around 88%–95%)
User Experience	Less personalized	Highly personalized

IV. Proposed Framework

Algorithm Involved:

To overcome the limitations of the existing system, the proposed system introduces a machine learning-based approach for detecting and classifying cyberbullying on social media. This system is designed to automatically analyze text data and identify whether it contains bullying content. The proposed system uses Natural Language Processing (NLP) techniques to process textual data. NLP enables the system to understand human language and extract meaningful information from it. The system begins by collecting data from publicly available datasets containing examples of both bullying and non-bullying.

The neural network architecture consists of multiple hidden layers that enable the model to capture both low-level linguistic features and high-level emotional cues. During training, the model minimizes classification loss across predefined mood categories, allowing it to distinguish subtle differences between emotional states. Once trained, the algorithm performs real-time mood classification by forwarding user input through the network and selecting the most probable emotion class. The detected mood is then passed to the music recommendation module, which maps the emotion to appropriate music categories and tracks, ensuring accurate and personalized recommendations.

Once the model is trained, it can be used to predict whether a given piece of text is cyberbullying or not. The system can be integrated into social media platforms to automatically monitor user-generated content and flag harmful messages. The proposed system offers several advantages over the existing system. It is capable of handling large volumes of data efficiently, making it suitable for real-time applications. It provides higher accuracy by considering the context of the text. It also reduces the need for manual intervention, saving time and resources. In addition, the system can be continuously improved by retraining the model with new data. This allows it to adapt to changing language patterns and emerging trends in cyberbullying.

System architecture defines the overall structure and workflow of the proposed system. It illustrates how different components interact with each other to perform the task of cyberbullying detection. The system architecture consists of several key components, including input, preprocessing, feature extraction, machine learning model, and output. The first component is the input layer, where the system receives text data from social media platforms. This data may include posts, comments, or messages entered by users. The input data is usually unstructured and may contain noise such as special characters, emojis , and irrelevant words. The processed data is then fed into classification algorithms like Logistic Regression, Naive Bayes, or Support Vector Machine to determine whether the content is cyberbullying or not. Additionally, the system can analyze the severity level of the detected content, categorizing it into different levels such as low, medium, or high risk. Based on the classification results, appropriate actions are taken, such as warning the user, flagging the content, or notifying the administrator for further review.

The proposed system aims to enhance online safety by reducing harmful interactions and providing an automated solution for content moderation. It minimizes human effort, improves detection accuracy, and ensures faster response to abusive behavior. Future improvements of the system may include support for multiple languages, detection of cyberbullying in images and videos, and the use of advanced deep learning models for better performance.

Proposed Model Information:

The proposed system for cyberbullying detection is designed to automatically identify and manage harmful content on social media platforms using Natural Language Processing (NLP) and Machine Learning (ML) techniques. In this system, user-generated content such as comments, posts, or messages is taken as input and collected for analysis. The text data is first preprocessed by removing unnecessary elements like stopwords, punctuation, and special characters, and then converted into a suitable format through techniques such as tokenization and stemming. After preprocessing, feature extraction methods like TF-IDF or Bag of Words are used to transform the text into numerical representations. These features are then fed into machine learning models such as Logistic Regression, Naive Bayes, or Support Vector Machine to classify whether the content is bullying or non-bullying. The system further analyzes the severity level of the detected content, categorizing it into low, medium, or high risk. Based on the results, appropriate actions are taken, such as warning the user, flagging the content, or notifying administrators. Finally, the output is displayed to the user or system moderator. This proposed system helps in reducing online harassment, improves user safety, and supports efficient content moderation, with potential future enhancements including voice and image-based cyberbullying detection and multilingual support.

Step 1: User input acquisition

User Input Acquisition is the first stage of the cyberbullying detection system, where the system collects data provided by users for analysis. In this phase, input is gathered from various sources such as social media posts, comments, chat messages, or online forums. The system can accept different types of input, primarily text, and in advanced cases, voice input which is converted into text using speech-to-text technology.

This step ensures that all relevant user-generated content is captured accurately and efficiently. The collected data may also include additional information such as timestamps, user IDs, or platform details to provide better context for analysis. Proper input acquisition is important because the quality and completeness of the collected data directly affect the accuracy of cyberbullying detection in later stages. Data validation is also performed at this stage to remove incomplete, duplicate, or irrelevant inputs, ensuring that only meaningful data is passed to the next step. Security and privacy are important considerations, so the system may anonymize user information or follow data protection guidelines while collecting and storing inputs.

Overall, User Input Acquisition acts as the foundation of the entire system. If the input data is accurate, relevant, and well-collected, the performance of later stages like preprocessing, feature extraction, and classification will be significantly improved, leading to more reliable detection of cyberbullying content.

Step 2: Text Preprocessing

Text preprocessing is the step where raw user input is cleaned and transformed into a structured format so that machine learning models can understand it effectively. The process starts by converting all text into lowercase to maintain uniformity (e.g., "Hate" → "hate"). Next, unwanted elements such as punctuation marks, numbers, special characters, and extra spaces are removed because they do not add meaningful information for detecting cyberbullying.

After cleaning, the text is broken down into smaller units called tokens using tokenization (for example, splitting a sentence into individual words). Then, common words that do not carry significant meaning, known as stopwords (like "is", "the", "and"), are removed using libraries such as NLTK. Following this, stemming or lemmatization is applied to reduce words to their root form (e.g., "bullying", "bullied" → "bully"), which helps in treating similar words as the same.
The next step is handling repeated characters and slang words (e.g., "looooser" → "loser"), which are common in social media text. The system may also expand abbreviations (e.g., "u" → "you", "idk" → "I don't know") to improve understanding. Additionally, URLs, hashtags, and mentions (@user) are either removed or processed separately depending on their importance.

Finally, the cleaned and processed text is ready for feature extraction techniques like TF-IDF or Bag of Words. Proper preprocessing improves the accuracy of the model by reducing noise and ensuring that only meaningful and relevant information is used for cyberbullying detection.

Step 3: Feature Representation

Feature representation is the process of converting cleaned text data into numerical form so that machine learning models can understand and process it. Since algorithms cannot directly work with raw text, this step transforms words and sentences into vectors (numbers) that capture their meaning and importance.

Step 4: Mood Classification

Model classification is the stage where the processed and feature-represented data is analyzed using machine learning algorithms to determine whether the given content is cyberbullying or not. After converting text into numerical form using techniques like TF-IDF or Bag of Words, the data is fed into classification models that are trained to recognize patterns of abusive or harmful language.

During training, the model learns from labeled datasets where examples of bullying and non-bullying content are already identified. After training, the model can classify new, unseen data. The output is typically a label such as "bullying" or "non-bullying," and in some cases, it may also provide a probability score indicating the confidence of the prediction.

User presentation is the final stage of the cyberbullying detection system, where the results of the analysis are displayed in a clear and understandable format to the user or system administrator. After the model classifies the input text, the system presents the output in the form of labels such as "Bullying" or "Non-Bullying," along with additional details like confidence score or probability level. In some cases, the system may also indicate the severity level of the content (low, medium, or high), helping users better understand the seriousness of the situation.

The presentation layer is designed to be simple and user-friendly, often implemented through a web interface, dashboard, or mobile application. It may highlight offensive words or phrases in the text, making it easier for users to identify problematic content. For administrators, the system can display reports, charts, or summaries showing the number of detected cases, trends over time, and user activity.

Additionally, the system provides actionable options such as warning the user, flagging or removing the content, blocking the user, or reporting to higher authorities. Notifications or alerts may also be generated in real time to ensure quick response. Overall, the user presentation layer ensures that the results are communicated effectively, enabling proper decision-making and promoting a safer online environment.

V. RESULTS AND DISCUSSION

The performance of the cyberbullying detection system was analyzed using multiple evaluation techniques to ensure reliability and accuracy. The model was tested on unseen data to evaluate its generalization capability. The dataset was divided into training and testing sets in an 80:20 ratio, which is a standard approach in machine learning. During testing, the model successfully classified most of the text data into the correct categories. The evaluation metrics indicated that the model performs consistently across different types of input data. The results demonstrate that the system is capable of identifying cyberbullying content with a high degree of accuracy.

A. The evaluation process also involved comparing different machine learning algorithms to determine which performs best for this task. It was observed that simpler models like Naive Bayes perform well for basic classification, while more advanced models like Support Vector Machine (SVM) provide better accuracy for complex datasets

Analysis of Results Graphical representation plays an important role in understanding the performance of the Graphical model. Various graphs such as bar charts and confusion matrix heatmaps can be used to visualize the results. Accuracy comparison graph shows how different algorithms perform Precision and recall graphs highlight

the balance between false positives and false negatives Confusion matrix heatmap shows classification performance visually These visualizations make it easier to interpret the results and identify areas for improvement. Error analysis is an important part of result evaluation. It helps in identifying the types of mistakes made by the model

The system mainly makes two types of errors: False Positives – Non-bullying content classified as bullying False Negatives – Bullying content classified as non-bullying False positives may lead to unnecessary filtering of harmless content, while false negatives are more serious as they allow harmful content to pass undetected. The analysis shows that most errors occur in cases where: 1. The text contains sarcasm or irony 2. The language is ambiguous 3. Slang or informal words are used.

B. Model Optimization

To improve performance, several optimization techniques were applied: 1. Hyperparameter tuning 2. Cross-validation 3. Feature selection These techniques help in improving model accuracy and reducing overfitting. Crossvalidation ensures that the model performs well on different subsets of data. Mood History Tracking

Scalability and Real-Time Performance The system is designed to handle large datasets efficiently. It can process multiple inputs simultaneously and provide results in real time. This makes it suitable for integration with social media platforms where large amounts of data are generated continuously. The system can also be scaled by:
1. Using cloud computing
2. Implementing distributed processing.

Challenges in Result Analysis Despite achieving good performance, several challenges were observed:
1. Difficulty in understanding sarcasm
2. Handling multilingual data
3.Dataimbalance
4. Limited context understanding

These challenges highlight the need for more advanced models andlargerdatasets

The result analysis of the project shows that the model performs with high efficiency and reliability. The system successfully detects a significant number of hate speech and cyberbullying incidents from the given data. It achieves an overall classification accuracy of around 94%, which indicates that the predictions made by the model are highly dependable. This level of accuracy demonstrates the effectiveness of machine learning algorithms in handling real-time social media data and identifying inappropriate behavior. To further evaluate the performance of the system, important metrics such as precision, recall, and F1-score are considered. The precision value of 93% indicates that most of the content identified as cyberbullying is actually correct, minimizing false positives. The recall value of 91% shows that the system is able to capture the majority of actual bullying 43 instances, reducing the chances of missing harmful content. The F1-score of 92% provides a balanced evaluation of both precision and recall, confirming that the model maintains a strong overall performance.

In addition to detection, the system also classifies cyberbullying into different categories for better understanding and analysis. The results indicate that hate speech constitutes about 40% of the detected cases, harassment accounts for 35%, and insults make up around 25%. This categorization helps in identifying the most common types of online abuse and enables platforms to take targeted actions against specific types of harmful behavior. Moreover, the implementation of this system significantly improves content moderation on social media platforms. By automating the detection process, it reduces the number of cyberbullying cases and ensures quicker response times. The system helps moderators take immediate action against offensive content, thereby preventing its spread.

It also enhances user safety by creating a healthier and more secure online environment. In conclusion, the project proves that machine learning is a powerful tool for recognizing and classifying cyberbullying on social media. With high accuracy and efficient performance metrics, the system not only detects harmful content but also categorizes it effectively.

This contributes to better moderation, improved user experience, and safer digital communication platforms. Once the input is provided, the system performs preprocessing steps before passing the data to the machine learning model. The preprocessing stage includes removing unnecessary characters, converting text to lowercase, eliminating stop words, and performing tokenization. These steps are essential to improve the accuracy of the model.

The preprocessing output may also be shown in a screenshot where the cleaned and processed text is displayed. This ensures that the input data is in a suitable format for further analysis by the ML model. After preprocessing, the processed text is fed into the trained machine learning model. The model has been trained using labeled datasets containing examples of both normal and cyberbullying content.

The model applies algorithms such as Naive Bayes, Support Vector Machine, or Logistic Regression to classify the input text. In the processing screenshot, the system may display intermediate steps such as feature extraction using techniques like TFIDF or Bag of Words. This stage highlights how the system converts text into numerical representations that can be understood by the model. The output generated by the system is displayed in a clear and understandable format. The output screenshot shows whether the given input text is classified as “Cyberbullying” or 46 “Non-Cyberbullying.” If the text is identified as cyberbullying, the system further categorizes it into types such as hate speech, harassment, or insult. For example, the output may display: “Result: Cyberbullying Detected – Category: Hate Speech.” This classification helps users understand the nature of the harmful content. Additionally, the system may also show a confidence score or probability value indicating how certain the model is about its prediction

CONCLUSION

The project “Recognition and Classification of Cyberbullying on Social Media with Machine Learning” focuses on detecting and analyzing harmful content using machine learning techniques. It uses labeled social media data (such as hate speech, insults, threats, and normal comments) along with preprocessing methods like text cleaning and tokenization to prepare the data. Feature extraction techniques like TF-IDF convert text into numerical form, enabling the model to classify content accurately. The system demonstrates strong performance using evaluation metrics like accuracy, precision, recall, and F1-score, making it effective for identifying cyberbullying. Overall, the system provides an efficient and scalable solution for real-time content moderation, reducing the need for manual monitoring and helping create a safer online environment. Although challenges like understanding context and sarcasm remain, future improvements using advanced models like LSTM or BERT and expanding to multilingual and multimedia data can further enhance its capabilities

REFERENCES

1. Ting, H., Liou, W. S., Liberona, D., Wang, S. L., & Bermudez, G. M. T. (2016). Towards the detection of cyberbullying based on social networking mining techniques. IEEE
2. Dadvar, M., & Eckert, K. (2018). Cyberbullying detection in social networks using deep learning models: A reproducibility study. IEEE.
3. .Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. (2019). Detection of cyberbullying using deep neural networks. IEEE.
4. Yadav, J., Kumar, D., & Chauhan, D. (2020). Cyberbullying detection using pre-trained BERT model. IEEE.
5. Hosseinmardi, H., Ardon, S., Ke, Q., Tran, H., Mishra, S., & Bailey, K. P. (2015). Detection of cyberbullying incidents on the Instagram social network. IEEE.
6. Zhang, R., Jin, K., Duan, S., Li, X., & Huang, M. (2016). Detection of cyberbullying in social media using a deep learning approach. IEEE
7. Ukey, A. K., & Ingle, V. S. (2017). Automatic detection of cyberbullying in Twitter using machine learning. IEEE.
8. Maithri, R., Raj, A. P., & Jana, D. (2019). Cyberbullying detection on social media using convolutional neural networks. IEEE.
9. Rezaei, M., Shahid, F., & Ahmed, S. (2021). Cyberbullying detection on social media using transformer models. IEEE.
10. Freitas, L., Freiz, J., & Siang, H. K. (2020). Combining text and image data for social media cyberbullying detection. IEEE. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. AAAI Conference.
11. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. ICWSM.
12. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Language Resources and Evaluation.
13. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. IEEE.
14. Xu, J., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. NAACL. . Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Miller, D., & Caragea, C. (2016). Content-driven detection of cyberbullying on the Instagram social network. IJCAI.
15. Rosa, H., Pereira, N., Ribeiro, R., et al. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior.
16. Al-Garadi, M. A., et al. (2016). Text classification models for cyberbullying detection: A review . IEEE Access.