

# Cyberbullying Detection with Machine Learning Techniques

Y Jyothika<sup>1</sup>, Likitha R<sup>2</sup>, Soundarya H J<sup>3</sup>, Dr. J Amutharaj<sup>4</sup>

Department of Information Science and Engineering <sup>1-4</sup>, Rajarajeswari College Of Engineering India<sup>1-4</sup>

\*\*\*

**Abstract** – The rise of social media platforms has inadvertently facilitated the spread of cyberbullying, affecting numerous young individuals. As these platforms proliferate, so does the prevalence of online harassment. This study introduces a machine learning-based approach to detect cyberbullying on social media, particularly Twitter. By leveraging Natural Language Processing (NLP) to analyze textual data and Long Short Term Memory (LSTM) networks for image recognition, the proposed model aims to identify bullying content effectively. The system utilizes Twitter's API to retrieve tweets, which are then processed using classifiers like XGBoost and Decision Trees to determine the presence of bullying behavior. However, many social media bullying detection techniques have been implemented, but many of them were textual based. The objective of our project work is to show the implementation of NLP and LSTM which will identify and classify tweets, posts, and other content associated with bullying. Accordingly, machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. NLP(Natural Language Processing) are used for identifying the complete sentence in the comments and LSTM(Long Short Term Memory) for image identification. Both NLP and LSTM were able to detect the true positives with more accuracy. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not along with we use XGBoost and Decision Tree algorithms.

**Key Words:** Cyberbullying, NLP(Natural Language Processing), LSTM(Long Short Term Memory), Twitter API, XGBoost, Decision Tree.

## 1.INTRODUCTION

Cyberbullying, a modern manifestation of harassment, has become increasingly common with the advent of digital communication platforms. Unlike traditional bullying, cyberbullying allows perpetrators to remain anonymous, making it more challenging to detect and address.

For instance, suspects in several recent hate-related terror attacks had an extensive social media history of hate related posts, suggesting that social media contributes to their radicalization.

Around 87 percent of the today's youth have witnessed some form of cyber bullying. Cyber Bullying can take different structures like Sexual Harassment, Hostile

Environment, Revenge, and Retaliation. Since offender is hidden to the victim, the problem statement gets complex. With the multiplication of online life and internet access, the act of cyber bullying too has increased, and it's difficult to detect. Thus, it is necessary to detect cyber bullying in order to protect adolescents. In this research, this vital data is utilized and information in the form of texts to improve the existing cyber bullying detection performance.

To address this, we propose a system which employs natural language processing techniques and the classification is done using machine learning approach that incorporates various classification techniques.

### 1.1 Purpose and Scope

The purpose of this project is to provide a desktop UI applications for classification of hate speech. To identify the maximum number of hate speech related tweets from twitter as soon as it is posted by users. The problem can be framed as a multi-class classification task, where tweets are categorized into three distinct classes: hate speech (HS), nonhate speech (NHS), or offensive content. This application can be used to classify the hate speech (HS), not hate speech (NHS) or offensive on social media network.

### 1.2 Problem Statement

In today's digital era, cyberbullying poses a significant threat, leading to emotional and psychological distress among users, especially adolescents and young adults. The challenge lies in the subtlety and anonymity of such acts, making manual detection arduous.

. To address this problem there is a critical need for an automated system that can effectively detect and mitigate instances of cyberbullying on social media platforms, chat application, and other online communication channels. The goal of this project is to develop a robust and accurate cyberbullying detection system using deep learning techniques, specifically long short-term memory(LST) and logistic regression. This system should be capable of analysing text and multi media content( such as images and videos) to identify and classify instances of cyberbullying, hate speech, or offensive content in real time.

## 2. Literature review

### 1. MelNet: A Generative model for audio in the frequency domain

In this paper a MelNet model, a generative model for spectral representation of audio is introduced MelNet combines a highly expensive autoregressive model with a multi scale modelling scheme to generate high resolution spectrograms generated by MelNet exhibit a realistic structure at both local and global levels. Unlike previous approaches that model time-domain signals directly, MelNet is especially effective at capturing long-range temporal dependencies. Experimental results indicate promising outcomes across various tasks, such as unconditional speech generation, music synthesis, and text-to-speech conversion..

### 2. MMM: Exploring conditional multi-track music generation with the transformer

In this paper a novel generative model for multitrack sequence generation under the framework of GANs. We have also implemented such a model with deep CNNs for generating multi-track piano-roles. We designed several objective metrics and showed that we can gain insights into the learning processes via this objective metrics. The objective metrics and the subjective user study shows that the proposed models can start to learn something about music. All though musically and aesthetically it may still fall behind the level of human musicians, the proposed model has a few desirable properties.

### 3. Counterpoint by convolution

The paper presents the process of placing notes against notes to construct a polyphonic musical piece. This is a challenging task, as each note has strong musical influences on its neighbors and notes beyond. Human composers have developed systems of rules to guide their compositional decisions. However, these rules sometimes contradict each other, and can fail to prevent their users from going down musical dead ends. Our current focus on statistical models of music represents one of several computational methods that enable composers to experiment with ideas more efficiently, thereby lowering the costs associated with creative exploration. Whereas previous work in statistical music modelling has relied mainly on sequence models such as Hidden Markov Models and Recurrent Neural Network (RNNs), we instead employ convolution neural

networks due to their invariance properties and emphasis on capturing local structure

### 4. Cyberbullying detection in social network: a comparison between machine learning and transfer learning

It explores the effectiveness of traditional machine learning versus transfer learning models in identifying cyberbullying content on social media platforms. The research evaluates multiple algorithms using benchmark datasets, highlighting that transfer learning approaches – particularly those leveraging pre-trained language models – consistently outperformed traditional methods in accuracy and contextual understanding. The study emphasizes the importance of deep semantic analysis in addressing online abuse and contributes valuable insights towards building more effective and intelligent content moderation system.

### 5. The role of artificial intelligence and cyber security for social media

It examines how AI technologies can be leveraged to enhance cybersecurity in social media platforms. The study explores the dual role of AI in both enabling smarter threat detection (such as identifying fake news, bots, and malicious content) and reinforcing security mechanisms through automated responses and real-time monitoring. It emphasizes the growing need for intelligent systems to combat evolving cyber threats, while also addressing ethical and privacy concerns associated with AI-driven surveillance.

## 3. Proposed System

The proposed system integrates machine learning and Natural Language Processing (NLP) methodologies to detect offensive content in online interactions. It begins by collecting and preprocessing data from social media platforms, chat applications, and public databases. To categorize information as bullying or non-bullying, it first gathers and preprocesses data from public databases, messaging apps, and social media. In the process of identifying sarcasm, irony, and emotional nuance-all these are frequently employed in cyberbullying-the algorithm also integrates contextual awareness. When a message is detected as potentially dangerous, it is blocked and automatically responds with reports or warnings. This method guarantees real-time, scalable detection on multiple web platforms. Natural language processing(NLP) and automated cyberbullying detection are combined in the suggested method.

## Advantages of Proposed System

### Efficiency and Scalability:

The system's ability to process massive amounts of data in real-time allows for quick detection and intervention across the multiplicity of online platforms, that makes them economical and scalable.

### Contextual and nuanced recognition:

The system can precisely detect subtle types of cyberbullying, such as sarcasm and irony, by utilizing sophisticated natural language processing (NLP) and machine learning techniques (e.g., sentiment analysis, sarcasm recognition, BERT, and LSTM).

Regular feedback and retraining improve the system's accuracy over time, allowing it to adjust to new language trends and bullying strategies. Automated detection guarantees the consistent and objective identification of harmful information.

## 4. Methodology

### Data collection using twitter tweets:

The sentiment/tweets are collected from a set of 20 accounts. The data retrieval is done by using twitter API using OAuthAPI used to authenticate the open-source framework with the twitter application.

### Sentimental storage based on tweets:

The sentimental storage based on tweets is a process of storing data about the tweets into the relational storage in terms (TwitterId, TwitterDesc, UserId). Twitter Id is unique Id associated with the tweet, Twitter Desc is actual tweet and UserId is the Id associated with the user.

### Stop words:

These are the set of words which do not have any specific meaning. The data mining forum has defined set of keywords. Stop words are words which are filtered out before or after processing of natural language data (text). There is not one definite list of stop words which all tools use and such a filter is not always used.

### Data cleaning:

Data cleaning is used for removing the stop words from each of the tweets and clean them. After the cleaning process is completed, the clean data can be represented as a set (CleanId, clean data, userId). Clean Id is the unique Id associated with the Tweet, Clean Data is the clean data after removal of clean data and user Id is the unique Id associated with the user.

## Music Generation Model (MG):

The Music Generation Model takes into account both the user input and the deleted sentiment to dynamically adjust the generation of music. The model is extended to be sentiment-aware, considering sentiment information during the composition process. It is trained on a dataset of diverse musical compositions, possibly including sentiment-tagged examples.

### Real-time Preview:

This component provides users with an immediate preview of the music generated based on their input and the deleted sentiment. Users can assess the emotional resonance and characteristics of the music in real-time, influencing their future interactions.

### User Feedback Mechanism:

The feedback mechanism allows users to provide input on the generated compositions, sharing their thoughts and preferences. It could include a form within the GUI where users submit comments, ratings, or other feedback. User feedback is essential for refining the sentiment to music mapping and improving the overall system based on user preferences.

### Feedback Submission System:

This system is responsible for collecting and submitting user feedback received through the GUI. It collects feedback submissions, which may include qualitative comments and quantitative ratings.

### Feedback Processing:

This component processes the submitted feedback, extracting valuable insights and patterns from user responses. It involves analyzing feedback to understand user preferences, satisfaction, and areas for improvement. The processes feedback contributes to iterative updates and improvements in various aspects of the system.

### Music Player:

The music player component plays the final generated compositions for the user to listen to and evaluate. It produces the auditory output based on the dynamically adjusted music generated by the Music Generation Model.

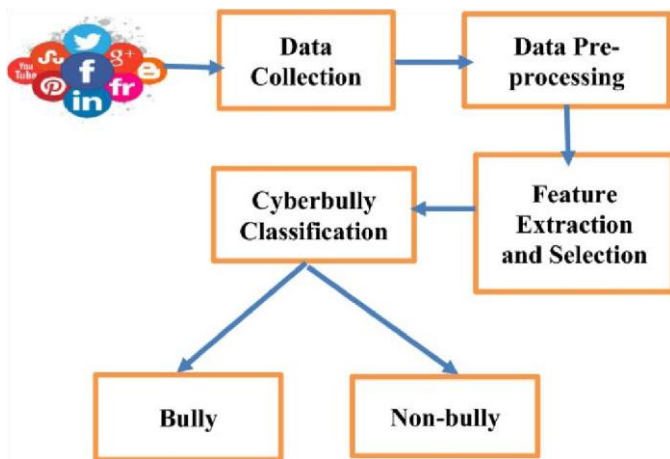


Chart -1: Methodology

## 5. CONCLUSION

This study presents a comprehensive approach to detecting cyberbullying by utilizing diverse datasets and deep learning techniques for feature extraction. The implementation of LSTM networks, enhanced by ReLU activation functions, has shown improved performance over traditional methods.

LSTM performance was improved over sigmoid by ReLU activation. To further enhance detection capabilities, future research proposes to include picture, video, and multilingual datasets.

The utilization of Long Short-Term Memory (LSTM) networks for cyberbullying detection represents a promising avenue, with current models demonstrating commendable performance in distinguishing harmful online behavior. As technology progresses, the future holds exciting possibilities for the field, including the exploration of advanced neural architectures, multimodal analysis, and real-time detection. Additionally, the development of context-aware models, personalized approaches, and the integration of behavioral analysis will contribute to more nuanced and accurate detection systems. However, ethical considerations, such as user privacy and fairness, must be at the forefront of development efforts. A collaborative, global approach involving researchers, policymakers, and educators is imperative to tackle the multifaceted challenges of cyberbullying effectively. As technology and society continue to evolve, the ongoing commitment to innovation, education, and ethical standards will be crucial in creating robust and responsible solutions for the detection and prevention of cyberbullying.

## REFERENCES

[1] Elaheh Raisi Bert Huang., "Cyber bullying Identification using Participant-Vocabulary Consistency" Virginia tech, Blacksburg, VA, 2016.

[2] Nebrase Elmrabbit Feixiang Zhou, Fengyin Li, Huivu Zhou., "Evaluation of Machine learning Algorithms for Anomaly Detection" 2018.

[3] Bhavani Thuraisingham., "The Role of Artificial Intelligence and Cyber Security for Social Media" Computer Science Dept. The University of Texas at Dallas Richardson, USA [bxt043000@utdallas.edu](mailto:bxt043000@utdallas.edu) 2020.

[4] Zaheer Abbass, Zain Ali, Mubashir Ali, Bilal Akbar Ahsan Saleem., "A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning" Department of computer Science University of Lahore, Guirat campus, Pakistan 2020.

[5] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe., "Detecting A Twitter Cyber bullying Using Machine Learning" Department of Information Technology Sardar Patel Institute of Technology Mumbai, India 2020.