CYBERBULLYING DETECTION IN SOCIAL MEDIA USING MACHINE LEARNING AND DEEP LEARNING APPROACHES

C.Kanimozhi[1], J.Anton Aakash[2]. S.Saravanan [2] and D.Venkatesh [2]
Assistant Professor[1], Final year[2]
Department of Information Technology
St.Joseph College of Engineering
Sriperumbudur , Chennai-602 117

Abstract— **Cyberbullying has become a significant concern in online platforms especially on social media where users can anonymously post harmful content. This project aims to develop a cyberbullying detection system that analyzes social media posts and identifies offensive or bullying content. A text dataset collected from Kaggle is used to train and compare multiple machine learning and deep learning model including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB) and Random Forest (RF) s. The best-performing model is selected based on accuracy and efficiency for real-time cyberbullying detection. The system is then integrated into a web application that monitors Twitter-based user posts. If a post is detected as cyberbullying, it is flagged and reported to an admin. The admin can issue up to three warnings to the user and if violations persist, the post is automatically deleted. The proposed system enhances online safety by reducing harmful interactions and promoting a respectful digital environment.**

I. INTRODUCTION

Cyberbullying has emerged as a serious issue in today's digital world, particularly on social media platforms, where users often interact anonymously. The anonymity provided by these platforms can sometimes encourage individuals to post harmful or offensive content, targeting others in a malicious manner. As the volume of social media content continues to grow, it becomes increasingly difficult for moderators to manually detect and address these incidents effectively. The project aims to develop an automated system capable of identifying cyberbullying in social media posts. By utilizing machine learning and deep learning models, the system will analyze textual content from platforms like Twitter to detect abusive language or harmful behavior. With this tool, it is possible to automate the detection process, providing faster response times and reducing the spread of harmful content online.

II. EXISTING AND PROPOSING SYSTEM

In the current scenario, cyberbullying detection on social media platforms is primarily managed through manual moderation and user reporting. Many social media companies rely on users to report offensive content, which is then reviewed by human moderators. However, this approach is time-consuming, inconsistent, and often ineffective in preventing the rapid spread of harmful messages. Some platforms use basic keyword-based filtering methods, which lack the ability to understand the context of a post, leading to false positives and negatives. Additionally, traditional rule-based systems struggle to detect sophisticated cyberbullying tactics, such as indirect harassment, sarcasm, or coded language. The absence of real-time detection and intervention allows harmful content to remain visible for extended periods, negatively impacting victims. Due to these limitations, there is a growing need for an intelligent, automated solution that can accurately identify cyberbullying patterns using machine learning and deep learning techniques, ensuring a safer online environment.

The proposed system aims to develop an intelligent and automated cyberbullying detection model using advanced machine learning and deep learning techniques. Unlike the existing system, which relies on manual moderation and keyword-based filtering, this system leverages models such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), and Random Forest (RF) to accurately analyze and classify social media posts. The best-performing algorithm is selected for real-time cyberbullying detection. The system is integrated into a web-based platform that continuously monitors user posts on Twitter. When a post containing cyberbullying content is detected, it is flagged and reported to an admin. The admin has the authority to issue up to three warnings to the user, and if the violations persist, the system automatically deletes the offensive post. This approach ensures a proactive response to cyberbullying, minimizing human effort while improving detection accuracy. By incorporating deep learning models, the system can understand the context of messages, detect indirect harassment, and adapt to evolving bullying patterns. The proposed system enhances online safety by providing an efficient and scalable solution to combat cyberbullying in real time.

III. SYSTEM STUDY

1) NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language. The goal of NLP is to enable machines to read, understand, and interpret human language in a way that is both meaningful and useful. NLP techniques are widely used in applications such as language translation, sentiment analysis, chatbots, and cyberbullying detection, as seen in the project mentioned.

a) Tokenization

Tokenization is one of the most fundamental steps in NLP, where a stream of text is split into individual units, known as tokens. Tokens can be words, phrases, or even characters, depending on the granularity needed. In the context of cyberbullying detection, tokenization breaks down social media posts into tokens such as individual words or sentences to be processed. This enables the model to analyze each unit for meaning and context.

b) Stop Word Removal

Stop words are common words such as "is," "in," "the," "and," and "a," which are often removed during preprocessing. These words usually carry little meaning and do not contribute significantly to the analysis of the text. By eliminating stop words, NLP models can focus on the most relevant parts of the text. For instance, in a post containing harmful language, the words "hate" or "bully" are more significant than common words like "the" or "is."

c) Stemming and Lemmatization

Both stemming and lemmatization are techniques used to reduce words to their base or root form. Stemming is a more aggressive technique that removes suffixes to find the root word (e.g., "running" becomes "run"). Lemmatization, on the other hand, uses a dictionary to return the base form of a word (e.g., "better" becomes "good"). These processes help reduce the complexity of the text, making it easier for machine learning models to interpret and analyze.

d) Part-of-Speech Tagging

Part-of-speech (POS) tagging is an important technique in NLP that involves identifying the grammatical components of a sentence, such as nouns, verbs, adjectives, etc. For example, in the sentence "He is bullying me," POS tagging would identify "He" as a pronoun, "is" as a verb, and "bullying" as a gerund. This technique helps in understanding the syntactic structure of sentences, which is crucial in detecting context in cyberbullying posts.

e) Named Entity Recognition (NER)

Named Entity Recognition (NER) is a technique that identifies and classifies named entities in text, such as people, organizations, dates, and locations. NER can be used to identify key individuals or entities mentioned in social media posts, which could be important in understanding the context of cyberbullying. For example, a post that mentions a specific individual may be flagged if it contains offensive language targeted at that person.

f) Word Embeddings

Word embeddings are a type of representation for text in which words are mapped to high-dimensional vectors. These vectors capture the semantic meaning of words based on their usage in context. Techniques such as Word2Vec, GloVe, and FastText create dense vector representations of words, allowing the model to understand relationships between words. For example, the words "bully" and "harassment" would be closer in the vector space, making it easier for the system to detect harmful language.

g) Sentiment Analysis

Sentiment analysis is a technique used to determine the emotional tone of a piece of text, whether it is positive, neutral, or negative. In cyberbullying detection, sentiment analysis can help identify posts with negative sentiments, such as anger, frustration, or hatred, which are often indicative of bullying behavior. By analyzing the sentiment of social media posts, the system can flag potentially harmful content for further review.

h) Text Classification

Text classification is a core NLP technique that involves categorizing text into predefined categories. In the context of cyberbullying detection, text classification models can classify posts into categories such as "bullying," "neutral," or "non-bullying." Machine learning algorithms like Naïve Bayes, Support Vector Machines (SVM), and deep learning models are commonly used for text classification. By training on labeled datasets, the model can learn to classify new posts based on their features.

i) Contextual Understanding with Transformers

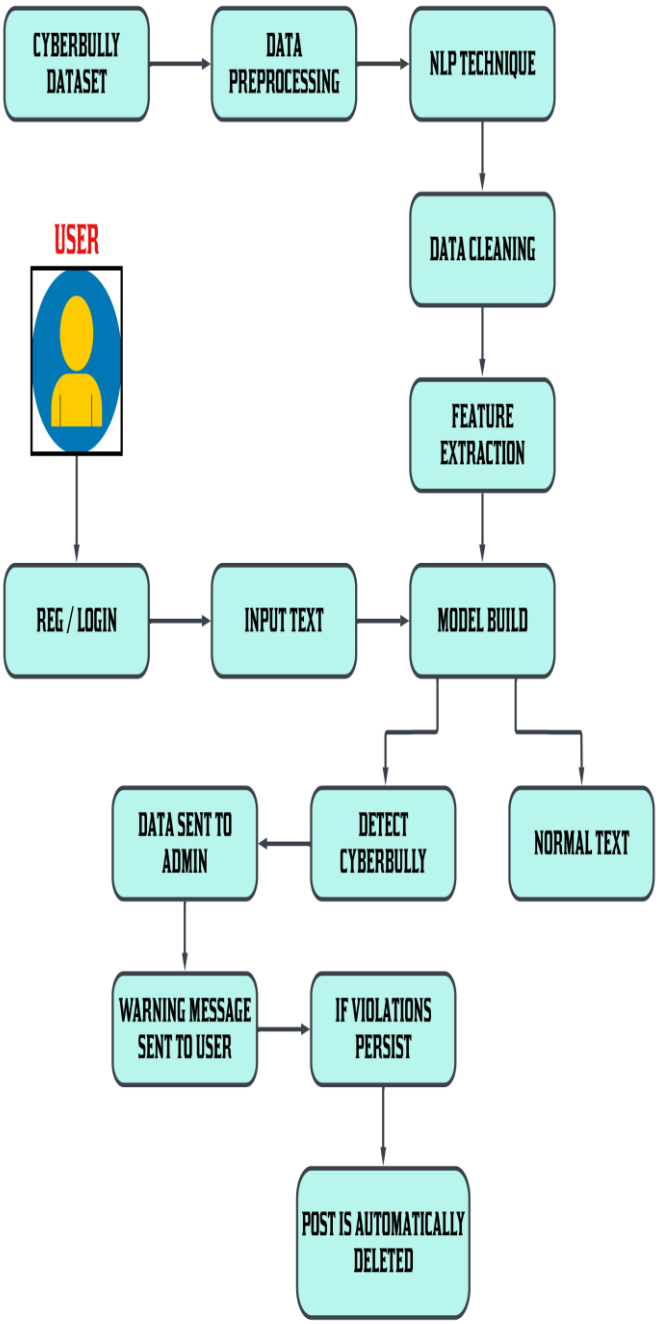
Transformers are a type of deep learning model that have revolutionized NLP. Unlike traditional models, transformers do not process text sequentially but instead look at the entire context of a sentence at once. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) have significantly improved performance in NLP tasks, including cyberbullying detection. These models are pre-trained on vast amounts of text data and can capture contextual relationships, making them highly effective for understanding the nuances of language, sarcasm, and implicit meaning in online posts.

In summary, NLP techniques are essential for enabling machines to process and analyze human language, particularly in applications like cyberbullying detection. The combination of tokenization, word embeddings, sentiment analysis, and advanced models like

transformers allows systems to accurately interpret and classify text, identifying harmful content in social media posts.

IV. ARCHITECTURE DIAGRAM

For system developers, they have system architecture diagrams to know, clarify, and communicate concepts regarding the system structure and also the user needs that the system should support.It's a basic framework may be used at the system designing section serving to partners perceive the architecture, discuss changes, and communicate intentions clearly.



V. SOFTWARE TESTING

UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

FUNCTIONAL TEST

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

SYSTEM TEST

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot "see" into it. The test provides inputs and responds to outputs without considering how the software works.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

1) TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VI. LIST OF MODULES

- Data collection
- Data preprocessing
- Feature extraction
- NLP Techniques
- Convolutional Neural Network Model Data

DATA COLLECTION:

The dataset used for this study is downloaded from website called kaggle.com. The dataset contains two types of set which are bullying text and non-bullying text. The goal is to identify all the bullying text.The training dataset used consisted of 1066 samples contained in a CSV File. Each sample is a sentence followed by the corresponding target label. The target label is 'pos' for a Non-Cyberbullying Sentence and 'neg' for a Cyberbullying Sentence. The data is then fed to the model for training.

DATA PREPROCESSING:

Text Blob, a python library that aims to provide access to common text-processing operations, is used for Sentiment Analysis. When a user sends a message, the message is checked for the phrase's polarity. The parameter is the phrase which has to be checked for profanity. The polarity will be any float bounded by -1 and 1 where 1 indicates the sentence is most negative and -1 indicates its most positive. We have to convert them into some other form like numbers or vectors before applying machine learning algorithm to them. In this way the data is converted by Bag-of-Words (BOW) so that it can be ready to use in next round. TF-IDF: One of the important features to be considered is this. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure to know the importance that a word carries in a document.

FEATURE EXTRACTION:

We wanted to develop a model based on textual features. This section describes the identification and extraction of features from each Form spring post. We were determined to avoid a bag-of- words approach for several reasons. First, the feature space with a bag-of-words approach is very large. Second, we wanted to be able to reproduce the model in code, and having each term as a feature would make that impractical. Third, we wanted to be able to understand why a post was considered as containing cyber bullying as this will inform the development of a communicative model for Cyber bullying detection.

NLP TECHNIQUES:

Natural language processing (NLP) is a study of artificial intelligence that helps machines and computers understand, interpret, and manipulate simple human language. Natural language processing helps developers organize knowledge to perform tasks such as translation,

summarization, named entity recognition, relationship extraction, voice recognition, topic segmentation, etc.Natural language processing is a way that computers analyze, understand, and derive meaning from day to day human language.

CONVOLUTIONAL NEURAL NETWORK MODEL:

It consists of many layered computations which are performed together. A neural network has layers known as: hidden layers, input layers and output layers and in case if the hidden layer is two or more the two than we can call it as deep neural network. It can be considered as the improved version of CNN (Convolutional Neural Network). This model has recently become very popular due to its accuracy over other algorithms. While training the dataset on DNN model an input vector is need to be collected. The training consists of two passes forward pass and backward pass. In forward pass a non-linear activation layer is calculated from input to output layer one by one. In backward pass we move in reverse order from output layer to input layer while calculating the error function.

VII. SCOPE FOR FUTURE DEVELOPMENT

The system supports users by creating a secure space where they can express themselves without the fear of harassment. Educational institutions and organizations can also utilize this tool to monitor and prevent cyberbullying among students and employees, fostering a positive digital culture. Furthermore, law enforcement agencies and policymakers can leverage this technology to analyze trends in cyberbullying and develop more effective regulations. By integrating advanced machine learning models, the project ensures high accuracy in detection, making it a valuable asset in combating online harassment and promoting digital well-being.

VIII.CONCLUSION

This project presents an advanced cyberbullying detection system designed to create a safer and more respectful social media environment. By integrating machine learning and deep learning algorithms such as CNN, RNN, SVM, MNB, and RF, the system efficiently identifies and mitigates harmful online behaviour. Unlike conventional approaches that depend on manual moderation and user reports, this system enables real-time monitoring, immediate detection, and automated intervention. The structured warning system ensures that users are given opportunities to correct their behaviour before their posts are removed, fostering a more controlled and ethical digital space. Additionally, the system's ability to adapt to new cyberbullying trends enhances its long-term effectiveness. By reducing reliance on human moderation and improving detection accuracy, this project offers a scalable and proactive solution to combating online harassment, ultimately promoting a more positive and inclusive social media experience.

IX. 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 IEEE Access, vol. 10, pp. 25857-25871, 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