International Journal of Advance Research in Science and Engineering
Volume. No. 15, Issue No. 05, May, 2026
ISSN: 2319-8354

Cyber Bullying Detection Using Machine Learning

Mrs.Velicharla Prathima
Assistant Professor
Tirumala Engineering College
Andhra Pradesh, India
prathima34.v@gmail.com

Shaik Shabeena Parveen
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
shabeenaparveen761@gmail.com

Bayyana Hima Sri
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
bayyanahimasri@gmail.com

Shaik Anwar Basha
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
skanwar7733@gmail.com

Kanugula Venkata Bhargavi
Department of IT
Tirumala Engineering College
Andhra Pradesh, India
22ne1a1261@gmail.com

***Abstract*—Cyber Bullying has emerged as a major challenge in online communication platforms, negatively affecting individuals' mental health and well-being. With the rapid growth of social me-dia, detecting harmful and abusive content manually has become difficult and inefficient. This project presents a machine learning-based web application for real-time Cyber Bullying detection using the Support Vector Machine (SVM) algorithm. The system utilizes Natural Language Processing(NLP) techniques such as text preprocessing, tokenization, stop-word removal, and stemming to prepare the input data for classification. A datasetof approximately 20,000 labeled comments is used to train the model, enabling it to accurately classify text into two categories: safe and cyberbullying. The trained model is integrated into a Flask-based web application that provides an interactive user interface. The application supports both single comment analysis and batch processing of comments from social media URLs. It displays results with clear feedback, including classification labels and percentage statistics. The proposed system offers improved accuracy, faster processing, and ease of use compared totraditionalmethods.Thissolutioncanbeeffectivelyused in educational institutions, social media platforms, and online communities to promote safer digital environments.**

***Index Terms*—**Cyberbullying Detection, Machine Learning, Natural Language Processing (NLP), Support Vector Machine (SVM), Text Classification, Social Media Analysis, Text Preprocessing,TF-IDF,WebApplication,Real-TimeDetection,Content Moderation, Data Mining

I. Introduction

In today's digital era, online communication through social media platforms has become an integral part of daily life. While these platforms provide opportunities for interactionand information sharing, they have also led to the rise of harmful activities such as cyberbullying. Cyberbullying refers to the use of digital platforms to harass, threaten, or humiliate individuals through abusive language, hate speech, or offensive comments. This issue has become a serious concern, especially among teenagers and young adults, as it negatively impacts mental health, self-esteem, and overall well-being. With the exponential growth of user-generated content on platforms such as Facebook, Twitter, and YouTube, manual monitoring of harmfull content has become extremely difficult and in ef-ficient. Traditional moderation techniques rely heavily on human intervention, which is time-consuming, inconsistent, and not scalable to handle large volumes of data generated every second. To address these challenges, automated cyberbullying detection systems using Machine Learning (ML) have gained significant attention. These systems analyze textual data and identify patterns associated with abusive or harmful language. By leveraging Natural Language Processing(NLP)techniques, such as tokenization, stop-word removal, and stemming, textual data can be effectively processed and transformed into meaningful features for classification. This project presents a Cyberbullying Detection Web Application that uses a Support Vector Machine (SVM) algorithm to classify user comments intosafeorcyberbullycategories.Thesystemisdesignedto perform real-time analysis through a user-friendly web interface developed using Flask. It supports both individual comment evaluation and batch processing of comments from social media sources, providing quick and accurate results.

II. Dataset Description

The dataset used for cyberbullying detection consists of textual data collected from publicly available sources such as socialmediaplatforms,onlineforums,andlabeleddatasets. It contains user-generated comments that are classified into categories such as safe and cyberbullying.

Key features of the dataset include the text content of comments, along with extracted linguistic features such as word frequency, term importance (TF-IDF), and contextual patterns. The dataset may also include metadata such as commentlabels,whichindicatewhetherthecontentisharmful or non-harmful.

Before training the model, the dataset undergoes several preprocessing steps to improve data quality. These includetext cleaning (removal of special characters and punctuation), tokenization,stop-wordremoval,andstemming.Additionally, the textual data is converted into numerical form using feature extraction techniques such as TF-IDF vectorization. These preprocessing steps ensurethatthe input data is structured and

suitable for machine learning algorithms, thereby enhancing the accuracy and performance of the cyberbullying detection system.

III. Literature Survey

Several research works have been carried out in the field of cyberbullying detection using artificial intelligence and machine learning techniques. Al-Marghilani (2022) proposed AI-enabled systemsfor cyberbullying prevention insmart city environments, improving user safety but facing scalability issues with large real-time data. Asfia Sabahath et al. (2024) introduced a hybrid deep learning frame work for image-based cyber bullying detection, achieving higher accuracy at the cost of increased computational requirements. Similarly, Belal Abdullah Hezam Murshed et al. (2022) presented a hybrid RNN-based model for Twitter data, which effectively captures sequential patterns but requires longer training time. Manuel F. Lo´pez-Vizca´ıno et al. (2022) focused on early detection of cyberbullying using machine learning, though the model struggles with contextual understanding and sarcasm. M.H. Obaidaetal.(2024)compared deep learning algorithms such as CNN and LSTM, demonstrating improved accuracy but requiring large datasets and high processing power. J. Sathya and F. M. H. Fernandez (2024) proposed an ontology-based approach combined with NLP to enhance semantic understanding, although it increases system complexity and maintenance difficulty. Earlier work by Dadvar and Eckert (2018) highlighted the effectiveness of deep learning models in improving detection accuracy, but emphasized the need for extensive training data. Milosevic et al. (2022) discussed the broader role of AI in addressing cyberbullying, providing conceptual insights but lacking practical implementation details. Gomez et al. (2022) focused on dataset curation through human-AI collaboration, though the process remains time-intensive.SaravananKarthikeyanetal.(2024)exploredmulti-modal learning by combining text and image data, improving detection performance but increasing system complexity. Al-Garadi et al. (2016) evaluated various machine learning techniques, offering strong baseline results but lacking real-time detection capabilities. Finally, Dinakar et al. (2011) presented early machine learning approaches for textual cyberbullying detection, which are limited in handling modern social media language and evolving communication patterns.

IV. ExistingSystem

Cyberbullying detection has been an active area of research, with several systems developed to identify harmful content on online platforms. Most existing approaches rely on machine learning and deep learning techniques to classify textual or multimedia data into safe and abusive categories. Traditional systems often use Natural Language Processing (NLP)methods such as keyword matching, bag-of-words, and basic feature extraction to analyze text data. While these methods provide moderate accuracy, they fail to capture the contextual meaning of sentences, especially in cases involving sarcasm, indirect language, or hidden intent. Recent advancements have incorporated deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks(RNN)toimprovedetectionaccuracy.Somesystems also combine multiple models, such as integrating machine learning algorithms like XGBoost with deep learning techniques for multimodal analysis (text and images). Although these approaches enhance performance, they introduce higher computational complexity and require large datasets and extensive training time. Despite these improvements, existing systems suffer from several limitations when applied to real-time cyber bullying detection. One of the major drawbacks is the high rate off al sepositives, where normal comments are incorrectly classified as harmful due to over-reliance on specific keywords. Additionally, many systems lack real-time processing capabilities and instead rely on batch processing, which delays the identification of harmful content. Another challenge is the high computational requirement of deep learning models, making them unsuitable for lightweight and scalable deployment. Furthermore, existing systems struggle with understanding context, sarcasm, and evolving language patterns used in social media communication. They also face scalability issues when handling large volumes of continuously generated data. These limitations highlight the need for a simple ,efficient, and real-time cyber bullying detection system that can provide accurate results with minimal computational overhead.

A. *Disadvantages of Existing System*

- Existing systems often misclassify normal comments as cyberbullying due to lack of proper context understanding.
- Use of complex machine learning and deep learning models requires high processing power, memory, and time.
- Lack of Real-Time Detection Many systems rely on batch processing, causing delays inidentifying harmful content.
- Poor Scalability and Context Understanding Systems struggle to handle large volumes of social media data and fail to detect sarcasm or indirect bullying effectively.

V. SystemArchitecture

ThesystemarchitectureoftheCyberbullyingDetectionWeb Application represents a structured workflow for processing and classifying user-generated content using machine learning techniques. The architecture is designed in a modular and efficient manner to ensure smooth data flow and real-time prediction. As illustrated in the system architecture diagram (page 15), the process begins with user input, where the user can either enter a single comment or provide a social media URL for batch analysis. The input data is first passed to the text preprocessing module, where it undergoes cleaning and normalization. This stage includes removing unwanted characters, eliminating stopwords, and performing tokenization

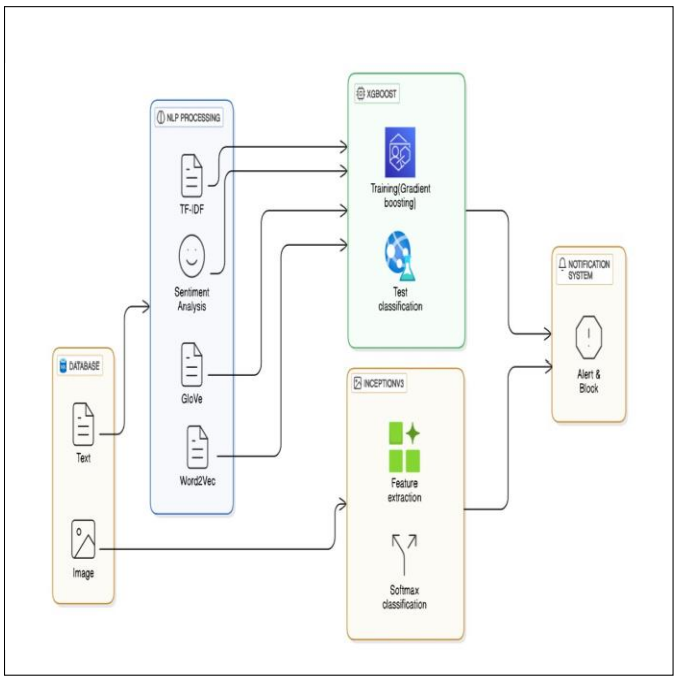

Fig.1.System Architecture

and stemming to convert raw text into a structured format. After preprocessing, the refined data is forwarded to the feature extraction module, where important textual featuresare identified and transformed into numerical representations using techniques such as vectorization. These extracted features are then provided to the core component of the system, the Support Vector Machine (SVM) model, which performs classification .The model analyzes the input and categorizes it into either safe or cyberbullying content based on learned patterns from the training dataset. The prediction output is generated accordingly and passed to the web interface. The final results are displayed through a user-friendly we binter face developed using Flask, where users can view classification results along with additional insights such as batch analysis statistics. The system also maintains a dataset component that supports training and continuous improvement of the model. Overall, the architecture ensures efficient processing, accurate classification, and real-time detection of cyberbullying content in an accessible and scalable manner.

VI. PROPOSEDSYSTEM

The proposed system is designed to provide an efficient and real-time solution for detecting cyberbullying in online content using machine learning techniques. It utilizes the Support Vector Machine(SVM) algorithm combined with Natural Language Processing (NLP) methods to classify user-generated comments into two categories: safe and cyberbullying. The system focuses on achieving high accuracy while maintaining low computational complexity ,making it suitable for real-time applications.

The system begins by accepting input from the user in the form of a single comment or a social media URL forbatch analysis. The input text is then processed through a preprocessing module, where it undergoes cleaning, tokenization, stop-word removal, and stemming to remove noise and standardize the data. After preprocessing, feature extraction techniques such as TF-IDF vectorization are applied to convert textual data into numerical representations suitable for machine learning.

The processed data is then passed to the SVM classifier, which acts as the core component of the system. The trained model analyzes the input and predicts whether the content is harmfull or safe based on learned patterns from the dataset.The prediction results are then displayed through a Flask-based web application, providing users with clear and immediate feedback.

Additionally, the system supports both individual comment analysis and batch processing of multiple comments, improving usability and practicality. It is designed to be lightweight, scalable, and user-friendly, allowing easy deployment without requiring high-end hardware. By enabling real-time detection and reducing reliance on manual moderation, the proposed systemhelpsinidentifyingharmfulcontentearlyandpromotes a safer online environment.

A. Advantages of Proposed System

- HighAccuracy
- Real-TimeDetection
- User-FriendlyInterface
- LowComputationalCost

VII. MACHINELEARNINGALGORITHMUSED

A. Support Vector Machine(SVM)

The proposed cyberbullying detection system uses the Support Vector Machine (SVM) algorithm for classifying user-generated text into safe and cyber bullying categories.SVM is a supervised machine learning algorithm that is highly effective for binary classification problems, especially in text classification tasks.

SVM works by finding an optimal boundary, known as a hyper plane, that separates data points of different classes with maximum margin. In this project, the input text is first converted into numerical feature vectors using techniques such as TF-IDF (Term Frequency–Inverse Document Frequency). These vectors are then used by the SVM model to learn patterns and distinguish between harmful and non-harmful content.

The choice of SVM is motivated by its high accuracy, efficiency, and ability to handle high-dimensional data, which is common in textual datasets. It also performs well even with limited training data and reduces the chances of overfitting compared to some other models. Due to these advantages, SVM is well-suited for real-time cyber bullying detection systems.

VIII. Experimental Setup

The experimental evaluation of the proposed cyberbullying detection system was conducted using Python-based machine learning libraries such as Scikit-learn, Pandas, and NumPy. The dataset, consisting of labeled user comments, was divided into training and testing sets using an 80:20ratio. The Support Vector Machine (SVM) classifier was trained on the training dataset and evaluated using unseen test data. Performance metrics such as accuracy, precision, recall, and F1-score were used to measure the effectiveness]softhe model .These met-rics provide a comprehensive evaluation of the classification performance in identifying safe and cyberbullying content.To ensure the reliability and robustness of the model, cross-validation techniques were applied during training. The ex-perimental results indicate that the proposed system achieves consistent performance across different data splits and is capable of accurately detecting cyberbullying content in real-time scenarios.

IX. Results

The experimental results of the proposed cyberbullying detection system demonstrate its effectiveness in accurately classifying user-generated contentint as a feandcy berbully-ing categories. The system was tested using both individual comments and batch inputs from social media URLs, and it produced reliable and consistent predictions.

The model achieved good accuracy in identifying harmful content, with correct classification observed in most test cases. The system also successfully handled different types of inputs, including normal text, abusive comments, and invalid inputs, providing appropriate responses in each case. Additionally,the batch analysis feature effectively calculated and displayed statistics such as the number and percentage of safe and harmful comments.

The results show that the system performs efficiently in real-time, with quick response time and minimal processing delay. The integration of preprocessing techniques and the SVM clas-sifier contributed to improved the prediction performance. Overall, the system proves to be a reliable and practical solution for detecting cyberbullying content and supporting safer online communication data.

X. ResultsandDiscussion

The performance of the proposed cyberbullying detection system is evaluated using a labeled dataset of user-generated comments collected from online platforms. To assess the ef-fectiveness of the machine learning model ,standard evaluation metrics such as accuracy, precision, recall, and F1-score are used. These metrics provide a comprehensive understandingof the system's ability to correctly classify content as safe or cyberbullying.

Experimental results indicate that the proposed system achieves high classification accuracy, demonstrating its capa-bility to effectively identify harmful and non-harmful com-ments. The use of the Support Vector Machine (SVM) al-gorithm, combined withNLP-based preprocessing techniques,

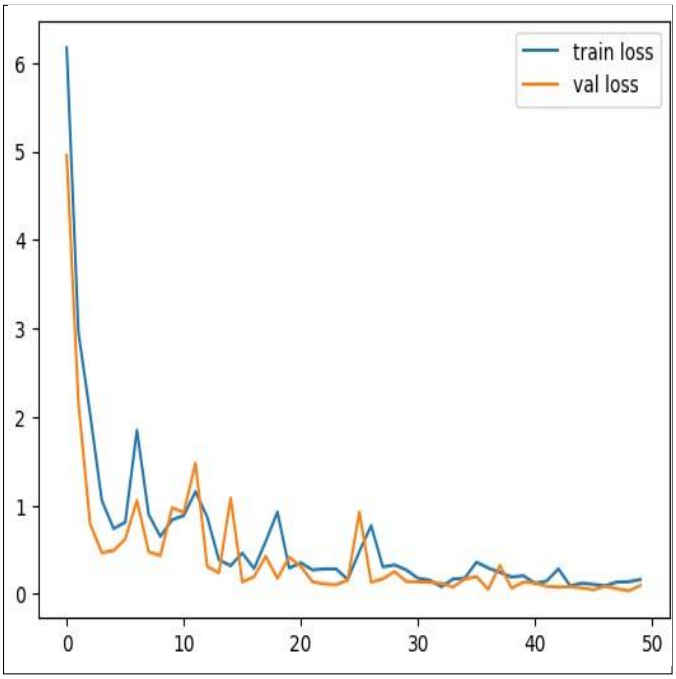

Fig.2.Training and Validation Loss

contributest o improved prediction performance.The model shows strong precision and recall values, indicating that it successfully reduces both false positive and false negative predictions, which is essential for reliable content moderation. Furthermore, the system performs efficiently in both single comment analysis and batch processing of social media data, providing consistentand real-timeresults. The integration of text preprocessing steps such as tokenization ,stop-word re-moval, and TF-IDF vectorization enhances feature quality and improves model performance .Overall ,the results demonstrate that the proposed system is accurate, efficient, and suitable for real-world applications indetecting and preventing cyberbul-lying, there by contributing oa safer online environment.

XI. Conclusion

TheCyberbullyingDetectionWebApplicationsuccessfully demonstrates the use of machine learning techniques to iden-tify harmful content in online communication. By utilizing the Support Vector Machine (SVM) algorithm along with Natural Language Processing (NLP) techniques, the system is able to classify user comments as safe or cyberbullying with good accuracy. The preprocessing steps such as tokenization, stop-word removal, and stemming play a crucial role in improving the performance of the model.The integration of the machine learning model with a Flask-based web application enables real-time analysis of user input. The system supports both single comment evaluation and batch processing of comments from social media sources, making it practical and user-friendly. The results are displayed clearly, allowing users to easily understand and interpret the predictions. Overall, the proposed system provides an efficient, scalable, and easy-to-

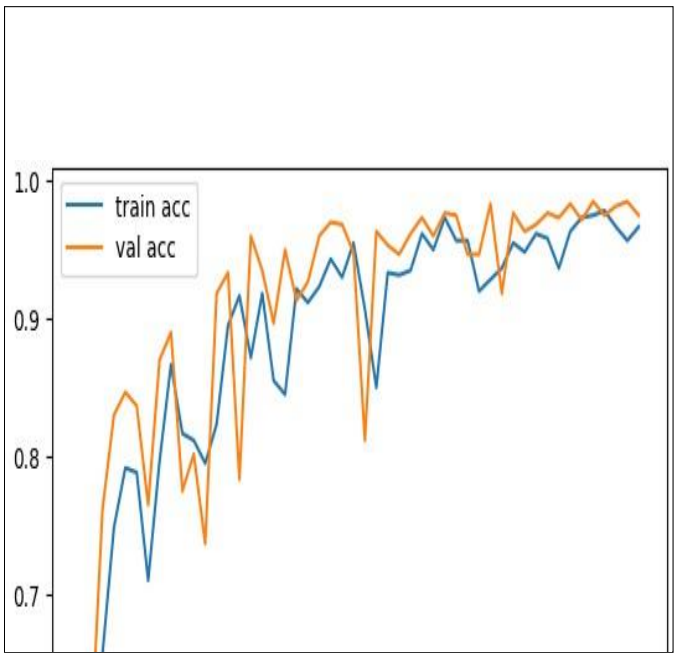

Fig.3.Training and Validation Accuracy

use solution for detecting cyberbullying. It helps in promoting safer online environments by reducing harmful interactions and enabling timely identification of abusive content. This project demonstrates how machine learning can be effectively applied to solve real-world problems and improve digital communication safety.

XII. FUTUREENHANCEMENTS

Although the proposed cyberbullying detection system achieves promising results, several enhancements can be incorporated to further improve its performance and real-world applicability.

A. Integratin of Real-Time Social Media Data

Future versions of the system can integrate real-time data from social media platforms through APIs. This enhancement will enable continuous monitoring of online content and support instant detection of cyberbullying activities.

B. Application of Deep Learning Techniques

Advanced deep learning models such as Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM), and BERT can be explored to better understand context, sarcasm, and complex language patterns. These models can further improved detectionaccuracyfordiverseandlarge-scaledatasets.

C. Incorporation of Multilingual Support

The system can be enhanced to support multiple languages, enabling detection of cyberbullying across different linguistic groups. This will make the system more versatile and applicable to global users.

D. Cloud-Based Deployment

Deploying the system on a cloud platform can improve scalability ,storage, and accessibility. Cloud integration allows efficient processing of large volumes of data and enables remote access for users and administrators.

E. Mobile Application Development

The system can be extended into a mobile application, allowing users to analyze comments and receive real-time alertsdirectlyonsmartphones.Thisimprovesaccessibilityand user engagement.

F. ExplainableAIforBetterDecisionSupport

FutureimprovementscanincludeexplainableAItechniques to provide clear reasoning behind predictions. This will help users and moderators understand why a comment is classified as cyberbullying, increasing transparency and trust.

REFERENCES

[1] Al-Marghilani,A.,*etal.*,"ArtificialIntelligence-EnabledCyberbullyingFreeOnlineSocialNetworksinSmartCities,"2022.

[2] AsfiaSabahath*etal.*,"HybridFrameworkforImageCyberbullyingRecognition," IEEE, 2024.

[3] C.BelalAbdullahHezamMurshed*etal.*,"DEA-RNN:HybridDeepLearning Approach," IEEE, 2022.

[4] ManuelF.Lo´pez-Vizca´ıno*etal.*,"EarlyDetectionofCyberbullying,"IEEE, 2022.

[5] M.H.Obaidaetal.,"DeepLearningAlgorithmsforCyberbullyingDetection," IEEE Access, 2024.

[6] J.SathyaandF.M.H.Fernandez,*etal.*"Ontology-BasedCyberbullyingDetection," IEEE, 2024.

[7] M. Dadvar and K. Eckert, *et al.* "Cyberbullying Detection Using DeepLearning Models," 2018.

[8] T. Milosevic et al., *et al.* "Artificial Intelligence to Address Cyberbully-ing," 2022.

[9] Christopher E. Gomez et al., *et al.* "Curating Cyberbullying Datasets,"2022.

[10] M.SaravananKarthikeyanetal.,*etal.*"MultimodalLearningforCyberbullying Detection," IEEE, 2024.

[11] A. Al-Garadi ,*et al.* "Machine Learning for Cyberbullying Detection,"2016.

[12] N. Dinakar ,*et al.* "Modeling the Detection of Textual Cyberbullying,"2011.

[13] R. Zhao ,*et al.* "Cyberbullying Detection Based on Semantic-EnhancedMarginalized Denoising Auto-Encoder," 2016.

[14] K.Reynolds,*etal.*"UsingMachineLearningtoDetectCyberbullying,"2011.