IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

An Automatic Method to Prevent and Classify Cyberbullying Incidents using Machine Learning Approach

Sheetal J[1], P Vinay Kumar[2], Vishal Raj[3], Vishwa Teja[4]
Associate Professor, Department of Computer Science & Engineering[1]
Final Year Students, Department of Computer Science & Engineering[2,3,4]
Ballari Institute of Technology and Management, Ballari, India.
[1]sjanthakal@yahoo.co.in, [2]vinaykumar15p@gmail.com,
[3]mvishalraj2002@gmail.com, [4]vishwat358@gmail.com

Abstract: *The technological advancements and the increasing popularity of social networking platforms, the sharing of personal information among online users has become widespread. This sharing occurs effortlessly through various devices such as computers and mobile phones. Cyberbullying can manifest through SMS, text messages, and various applications, as well as online platforms like social media and forums, where individuals can view, engage with, or distribute content.The project offers a comprehensive understanding of Cyberbullying incidents and their corresponding offences combining a series of approaches reported in relevant Work. The implementation provides the opportunity to systematically combine various element or Cyberbullying characteristics. Additionally, a comprehensive list of Cyberbullying-related offences is put forward. The offenses are ordered in a Deep Neural Network classification system based on specific criteria to assist in better classification and correlation of their respective incidents. This enables a thorough understanding of the repeating and underlying criminal activities. This study focuses on classifying user posts and image content into bullying or non bullying through reputation score.*

Keywords: cyber bullying, neutral networks, classification, bullying and non bullying

I. INTRODUCTION

In the contemporary digital landscape, cyberbullying emerges as a pressing concern, profoundly impacting the well-being and safety of individuals within online communities. Despite concerted efforts, conventional methods for detecting and addressing cyberbullying confront formidable challenges, particularly in effectively parsing textual content concealed within images. This limitation impedes the timely execution of intervention and prevention measures, necessitating innovative solutions to bridge this gap.

To address these challenges, this research introduces a pioneering approach centered on leveraging Deep Neural Network (DNN) architectures for cyberbullying detection within image-textual content. Unlike existing methodologies, which often rely on text inputs alone, our proposed system extends its reach to accommodate image-based content, thereby broadening the scope of cyberbullying detection. Through the utilization of advanced image-to-text extraction techniques, textual information embedded within images is seamlessly retrieved, enabling comprehensive analysis.

At the core of our methodology lies a sophisticated DNN-based classification model, meticulously designed to discern instances of bullying language within the extracted textual data. This model operates in real-time, dynamically adjusting users' reputation scores based on the presence of bullying or non-bullying words. By instituting this adaptive feedback mechanism, our system endeavors to cultivate a safer online environment, fostering mutual respect and civility among users.

Moreover, our proposed system incorporates an ordered offense classification mechanism within the DNN framework, facilitating the systematic categorization and correlation of cyberbullying incidents. By delineating specific criteria, this

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

classification system enables stakeholders to gain deeper insights into recurring patterns and underlying criminal activities, thereby informing the development of more targeted intervention strategies.

Through rigorous experimental evaluations, we demonstrate the efficacy and efficiency of our approach in detecting and mitigating cyberbullying incidents. By contributing novel insights and practical solutions to the ongoing discourse on online safety mechanisms, this research endeavors to advance the collective efforts aimed at combatting cyberbullying and fostering a more inclusive and harmonious digital ecosystem.

II. LITERATURE REVIEW

In [1], a victims often feel surrounded by cyberbullying, as the internet is readily accessible with just a click. This form of harassment can have profound mental, physical, and emotional repercussions. Typically occurring through text or images on social media platforms, distinguishing bullying content from non-bullying material is crucial for effective intervention. An efficient cyberbullying detection system can play a vital role in social media websites and messaging applications, helping to combat these attacks and mitigate the prevalence of cyberbullying incidents.

In [2], the cyberbullying is a prevalent issue on social media, often leading individuals to experience mental distress rather than confronting the bully. Detecting such situations automatically across most social networks requires intelligent systems. To tackle this, we've introduced a cyberbullying detection system. Our approach involves a deep learning framework designed to analyze real-time Twitter tweets or social media posts, accurately identifying any cyberbullying content within them. Recent studies indicate that deep neural network-based methods are more efficient than traditional techniques in detecting cyberbullying texts.

In [3],cyberbullying's impact is hard to gauge because it's subjective – what's hurtful to one person might not bother another. This makes spotting bullying content tricky, especially with images, which haven't gotten much attention in research. We're working on a model to tackle image-based cyberbullying on social media using deep learning methods like convolutional neural networks.

Authors of [4] specifies that usage in digital/social media is increasing day by day with the advancement of technology. People in the twenty-first centuryare being raised in an internet-enabled world with social media. This paper describes a system for automatic detection and prevention cyberbullying considering the main characteristics of cyberbullying such as Intention to harm an individual, Repeatedly and over time and using abusive curl language or hate speech using 8 supervised machine learning.

The study of [5]to find cyberbullying on different social media sites and sort it by how serious it is. We'll compare different types of machine learning and deep learning methods to see which works best. Plus, we'll share three datasets labeled with four levels of cyberbullying severity (none, low, medium, and high) for others to use in their research.

In [6],using online platforms, malicious individuals carry out unethical and deceitful actions to inflict emotional distress and harm the reputation of others. Recently, cyberbullying has emerged as a significant concern in social media. Cyberbullying, also termed cyber-harassment, encompasses electronic forms of bullying or harassment. As digital platforms expand and technology advances, cyberbullying has become increasingly prevalent, particularly among adolescents.

Authors of [7]specifies that the usage and use of the internet and social media is increasing day-by-day and consequently cyberbully vulnerabilities are also growing, Cyberbullying is an aggressive, planned behavior carried out by a group or individual. It is happening by sending, posting, sharing negative, harmful, untrue contents in online. It leads to psychiatric and emotional disorders for those affected. Hence, there is a critical requirement to develop automated 30 methods for cyberbullying detection and prevention.

In[8] the internet serves as the largest platform globally for communication and idea-sharing, with nearly 4 billion users across various social media platforms such as Twitter and WhatsApp. However, the prevalence of online abuse, harassment, trolling, and cyberbullying is escalating. Victims of such behavior often experience depression, engage in self-harm, and tragically, some even resort to suicide. Consequently, identifying bullying texts or messages on social media can significantly mitigate these harmful consequences.

The study of [9] specifies thatonline social networks have opened up new ways for cyberbullies, allowing them to target places and countries previously out of reach. We're using Support Vector Machines (SVM) to spot cyberbullying on Twitter. Our goals are listed in the objective section. Also, we'll use Optical Character Recognition (OCR) to detect

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

image-based cyberbullying. We'll assess the impact on individuals using a dummy system. By applying machine learning and natural language processing, we'll identify cyberbullying patterns and automatically detect them by comparing text data to known traits.

The study of [10]specifies that social networking is expected to have a huge impact, with over 3.02 billion active users on social media every month globally by 2021, which is about a third of the world's population. Among the many social networks, Twitter is crucial for researchers as it's a popular platform for real-time microblogging where news often breaks before official sources. With its short message limit of 280 characters and unfiltered feed, Twitter is widely used for quick communication and sharing of information.

In [11], purpose of cyberbullying is characterized by aggressive and negative behavior that is unwanted and repeated, occurring through digital devices such as cell phones, computers, and tablets. It can manifest across various platforms including email, social media, gaming, instant messaging, and photo sharing. Tactics employed by cyberbullies include threats and the dissemination of abbreviated forms of personal information, leading to online harassment aimed at exacting revenge, threatening, and compromising the privacy of individuals. This may involve making private information such as social security numbers, credit card details, phone numbers, links to social media profiles, and other personal data public.

The study of [12] specifies that cyberbullying (CB) represents a significant issue demanding attention across various social media platforms. This form of aggression, defined as repetitive and purposeful behavior, occurs through the utilization of information and communication technology (ICT) platforms like social media, the internet, and cell phones. Hate messages are commonly disseminated via email, chat rooms, and social media platforms accessed through computers and mobile devices. Given the prevalence of cyberbullying, employing deep learning (DL) models for its detection and categorization within social networks is imperative. A promising approach in this regard is Feature Subset Selection with Deep Learning-based Cyberbullying Detection and Categorization (FSSDL-CBDC), which integrates deep learning techniques with feature subset selection methods to address this growing concern in social media landscapes.

III. METHODOLOGY

This paper introduces a method for detecting cyberbullying on social media, employing a reputation score that dynamically adjusts based on the presence of bullying or non-bullying words. This scoring mechanism goes beyond mere sentimental analysis, incorporating considerations of the syntactic, semantic, and sarcastic nuances within sentences before labeling them as hate speech. The methodology commences with traditional sentiment analysis, involving contextual text mining to uncover subjective information and comprehend opinions, emotions, or attitudes towards the topic. Subsequently, a set of "social" features is introduced to enhance and guide the cyberbullying detection process. We have divided all the features we have extracted into three categories:

- Sentimental Features
- Sarcastic Features
- Syntactic Features

These features have been organized following a comprehensive review of existing systems in the literature, ensuring each feature provides a distinct identification of the text. Selecting informative, descriptive, and independent features is vital to enhancing the effectiveness of algorithms in pattern recognition and classification tasks.

A. System Architecture of cyberbullying

The Fig.1. represents the system architecture of cyberbullying

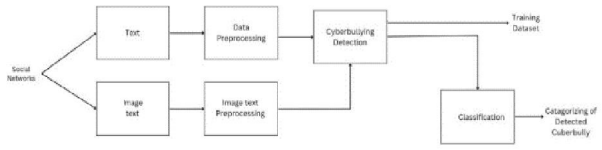

Fig.1. System Architecture of Cyberbullying

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

B. Implementation

- Data Collection: Gather diverse datasets containing text, images, and videos from various online platforms, ensuring a representative sample of cyberbullying instances.
- Preprocessing: Clean and preprocess the data, including text normalization, image resizing, and video frame extraction, to prepare it for analysis.
- Feature Extraction: Extract relevant features from text, images, and videos, including linguistic patterns, visual cues, and audio attributes.
- Multimodal Fusion: Combine multimodal features using fusion techniques, such as late fusion or deep learning-based fusion, to create comprehensive representations of content.
- Machine Learning Models: Develop a range of machine learning models, incorporating both deep neural networks and ensemble methods. These models will be trained using combined features specifically tailored for detecting instances of cyberbullying.
- Contextual Analysis: Develop algorithms to consider conversational context, user interactions, and platform-specific norms to enhance detection accuracy.
- Behavioral Analytics: Incorporate behavioral analysis methods to identify patterns in user behavior that indicate potential bullies or victims.
- Explainable AI: Implement explainability techniques to provide clear rationales for detection decisions, improving transparency.
- User Feedback Integration: Design a user-friendly interface for real-time alerts and user feedback, enabling community engagement and refinement of the system.
- Testing and Evaluation: Conductcomprehensive testing and evaluation using standard metrics, such as precision, recall, and F1-score, on diverse datasets and social media platforms to validate the system's effectiveness and usability.

C. Flowchart

The Fig.2. represents the flowchart of cyberbullying

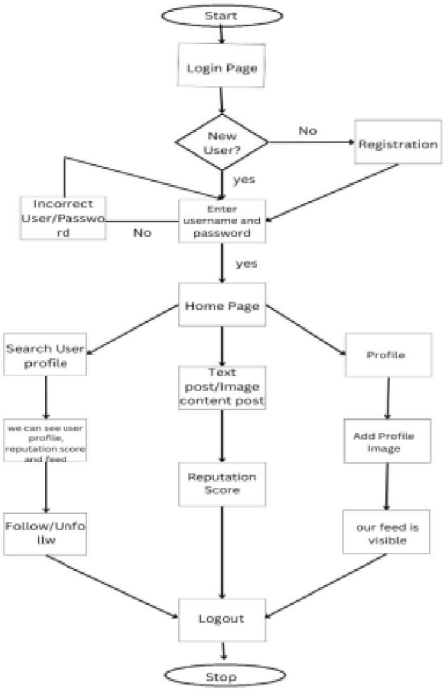

Fig.2. Flow Chart of Cyberbullying

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

D. User Case diagram

The Fig.3.represent the User Case diagram of Cyberbullying

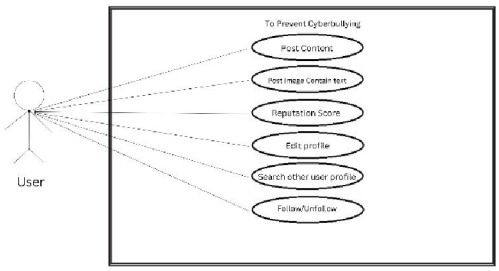

Fig.3. User Case diagram of Cyberbullying

IV. UML DIAGRAM

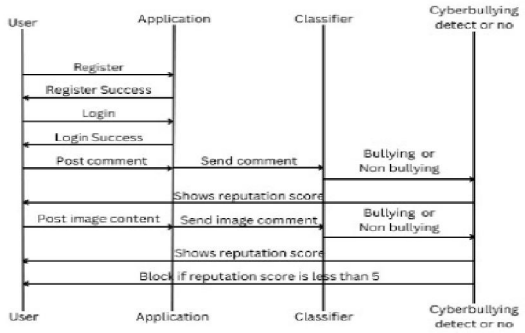

Fig.4. UML Diagram of Cyberbullying

E. Machine Learning Algorithms Applied

a. Long Short-Term Memory

LSTMs, or Long Short-Term Memory Networks, are a type of Recurrent Neural Network designed to tackle tasks involving sequences. They're great at remembering long-term patterns, unlike traditional RNNs that struggle with vanishing gradients. LSTMs work by using memory cells and gates to control information flow, allowing them to hold onto important details and discard less relevant ones. They're commonly used in things like language processing, speech recognition, and predicting future trends in data like stock prices or weather patterns. See Fig.5 for a visual of how LSTMs function.

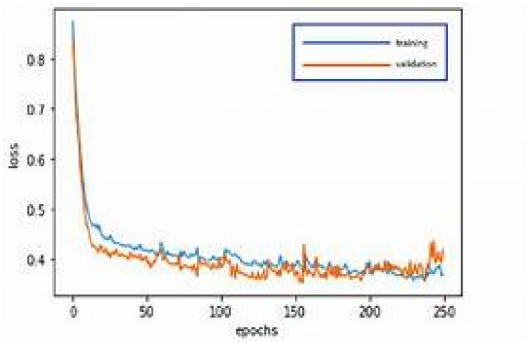

Fig.5 Long Short-Term Memory graph

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

b. Random Forest Algorithm

Random forests are crafted by constructing individual trees using a random subset of the dataset, evaluating a random subset of features at each partition. This inherent randomness fosters diversity among trees, curbing overfitting tendencies and enhancing predictive accuracy. During prediction, the algorithm amalgamates outcomes from all trees through voting (for classification) or averaging (for regression), culminating in a robust decision-making process. Leveraging insights from multiple trees, this collaborative approach yields stable and precise results. Renowned for their prowess in handling intricate datasets, mitigating overfitting risks, and delivering dependable forecasts across various scenarios, random forests find widespread utility in both classification and regression tasks. Refer to Fig.6. for a visual representation of the Random Forest algorithm.

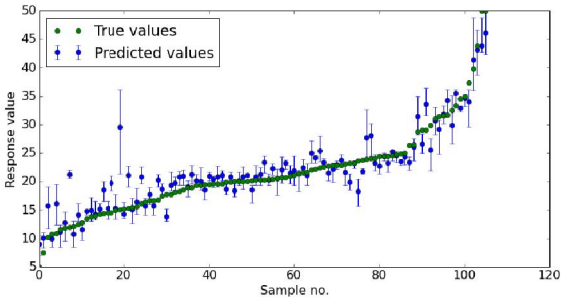

Fig.6. Random Forest graph

c. SVC

The Support Vector Machine (SVM) is a robust machine learning algorithm renowned for its versatility in handling both linear and nonlinear classification, regression, and outlier detection tasks. SVMs find applications in diverse fields such as text and image classification, spam filtering, handwriting recognition, genetic analysis, facial recognition, and anomaly detection. Their efficacy lies in their ability to effectively handle high-dimensional data and complex relationships. The primary objective of the SVM algorithm is to construct an optimal decision boundary, termed as a hyperplane, which effectively separates the n-dimensional space into distinct classes. This hyperplane facilitates the accurate categorization of new data points in the future. This best decision boundary is called a hyperplane.

How does SVM work?

The main aim is to split the dataset well. The margin, which is the distance between the closest points, is key. We want to choose a hyperplane that maximizes this margin between the support vectors in the dataset. SVM looks for the hyperplane that gives us the biggest gap between the data points from different classes. SVM searches for the maximum marginal hyperplane in the following steps:

- Generate hyperplanes which segregates the classes in the best way.
- Left-hand side figure showing three hyperplanes black, blue and orange.
- Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
- Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.

The Fig.7. shows the graph of SVC.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

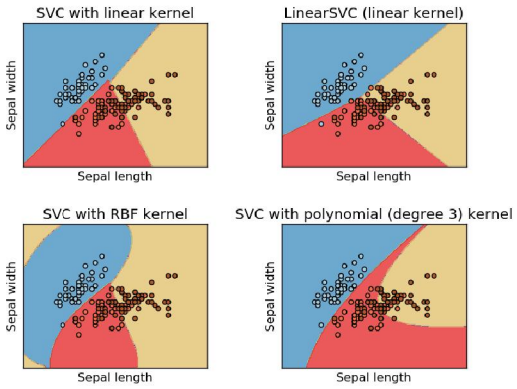

Fig.7. SVC graph

d. XGBoost Algorithm

XGBoost, short for Extreme Gradient Boosting, is a highly efficient and accurate machine learning algorithm. It offers parallel tree boosting and is considered the forefront library for addressing regression, classification, and ranking challenges. Renowned for its outstanding predictive capabilities, XGBoost is widely recognized as the premier choice in ensemble learning, particularly for gradient-boosting algorithms. It develops a series of weak learners one after the other to produce a reliable and accurate predictive model. Fundamentally,XGBoost builds a strong predictive model by aggregating the predictions of several weak learners, usually decision trees.

It uses a boosting technique to create an extremely accurate ensemble model by having each weak learner after it correct the mistakes of its predecessors.XGBoost falls under the boosting suite of algorithms which in turn is part of the ensemble learning method.XGBoost (Extreme Gradient Boosting) is a powerful tool in machine learning that's like having a team of experts make predictions together. Think of it like a group of cricket experts helping you decide whether to select a particular player in a team based on his recent form, strike rate, runs the batsman scored and his average. Each expert focuses on a specific detail and gives their opinion.XGBoost combines all their opinions to make a super accurate prediction. It's great at many tasks, like guessing movie ratings based on reviews.Just like your friends learn from their mistakes, XGBoost learns and gets better at guessing with each round.The Fig.8.shows the graph of XGBoost.

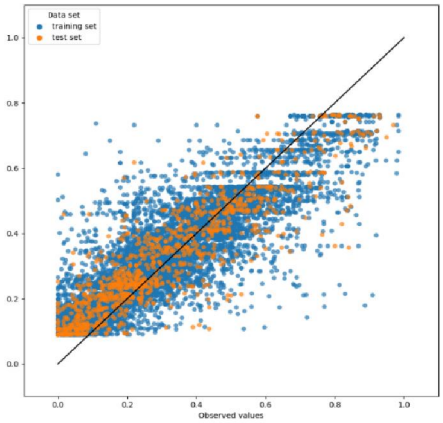

Fig.8.XGBoost graph

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

V. RESULTS

In our testing, we processed input and made predictions using our model. We tested it with text containing instances of bullying. In Table 1,"0" means Bullying, and "1" means Non-Bullying. The accuracy of the Long Short-Term Memory (LSTM) model was 67.25%. In Table 2, the accuracy of XGBoost, Random Forest, and SVC was 71.35%, 66.82%, and 71.50%, respectively.

Table 1. Accuracy of LSTM

Classifier	Accuracy in percentage
Long Short-Term Memory	67.25%

Table2. Accuracy of XGBoost, Random Forest, SVC

Classifier	Accuracy in percentage
XGBoost	71.35%
Random Forest	66.82%
SVC	71.50%

Screenshots of the project

1.Cyber Bullying page without any bullying words:The Fig.9. represents the reputation score of the user where doesn't decrease because feed doesn't contain any bullying words.

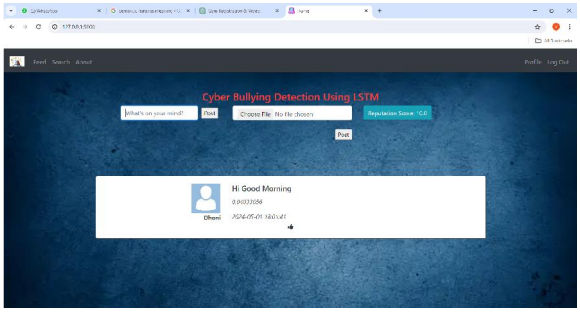

Fig 9: Non bullying

2. Cyber Bullying page where feeds contain bullying word:The Fig.10. represents where the user reputation score decrease's because feed contain any bullying words.

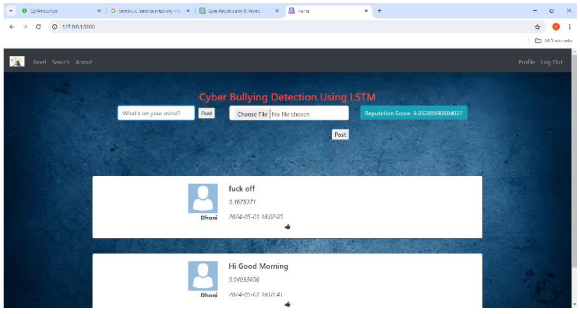

Fig 10: Bullying

3.User getting blocked page: The Fig.11. represents account of user get blocked automatically if reputation score of user is below five.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

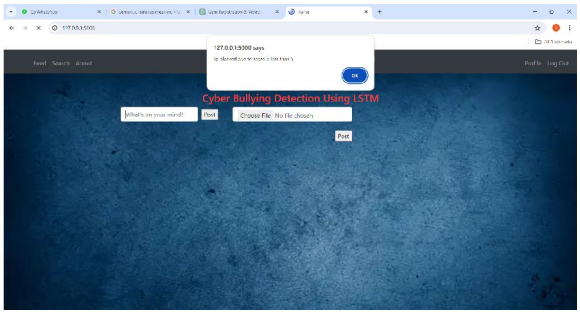

Fig 11: Block account

VI. CONCLUSION

The system effectively detects cyberbullying in text and image captions using advanced techniques. We compute a reputation score for each user, aiding in identifying potential cyberbullying behavior. Leveraging XGBoost, Random Forest, SVC, and LSTM, we achieve robust detection across various user-generated content. Our project emphasizes the importance of technology in combating cyberbullying, promoting responsible online conduct, and safeguarding users' mental well-being.

REFERENCES

[1]. S. K. K. a. R. D. Aditya Desai, "Cyber Bullying Detection on Social Media using Machine Learning," in ITM Web of Conferences 40, 2021.

[2]. S. a. R. Mitushi Raj, "An Application to Detect Cyberbullying Using Machine Learning," SN Computer Science, pp. 1-13, 2022.

[3]. P. K. R. a. F. U. Mali, "Cyberbullying detection using deep transfer learning," Complex & Intelligent Systems, 2022.

[4]. P. a. P. Fernandob, "Accurate Cyberbullying Detection and Prevention on Social Media," in International Conference on Health and Social Care Information Systems and Technologies, 2020.

[5]. K. M. a. A. C. Akshita Aggarwal, "Comparative Study for Predicting the Severity of Cyberbullying Across Multiple Social Media Platforms," IEEE, 2020.

[6]. L. a. Harshini.M, "Cyberbullying Detection using machine learning," IRJEdT, vol. 5, no. 4, 2023.

[7]. D. H. P. D. a. A. P. Dr. Vijayakumar V, "Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms," International Journal of Applied Engineering Research , vol. 16, pp. 568-574, 2021.

[8]. K. R. K. S. a. C. P. Ninad Mehendale, "A Review on Cyberbullying Detection Using Machine Learning," SSRN, pp. 1-5, 2022.

[9]. M. W. R. B. G. V. S. M. S. U. D. a. W. .. Miss. Jafri Sayeedaaliza Abutorab, "DETECTION OF CYBERBULLYING ON SOCIAL MEDIA," IRJMETS, vol. 4, no. 5, 2022.

[10]. M. 1. a. S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," MDPI, 2020.

[11]. M. Saharan, "Cyber Bullying Detection on Social Media Using Machine Learning," SSRN, p. 6, 2023.

[12]. S. M. S. C. M. K. A. K. S. P. S. R. a. T. L. Neelakandan S, "Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media," Hindawi, 2022.