

Cyberbullying Detection and Prevention System

Richie Kishan¹, Shaswhat Reshme² and Lalruatmawia Tochohawn³

¹⁻³Fr. Conceicao Rodrigues Institute of Technology, Vashi, Computer Department, Navi Mumbai, India
Email: richiekishan9@gmail.com

Abstract—The era of social networking began the moment the internet was invented. No one may have anticipated that the internet would eventually be home to a wide range of incredible services like social media networking. Online applications, as of now, and social networking sites have integrated into our daily lives. a person's life. Numerous individuals of various ages spend hours on such websites every day. Even though humans are emotionally volatile Social media connections between these facilities bring they carry significant risks with them like cyberattacks, which Cyberbullying and online grooming One of the most current social media sins is cyberbullying. The freedom of expression is being abused as social media usage grows. According to statistics, 36.5 percentage of people believe they have experienced cyberbullying at some point in their lives. These figures are up from 2018–19 and are more than twice what they were in 2007, which suggests we are going in the wrong direction. There are currently products on the market that can help to some degree with this problem. However, they have usage restrictions or don't employ effective algorithms.

Index Terms— Social media networking, cyberbullying, blocking, prevention, model.

I. INTRODUCTION

The number of social media users on the Internet has seen a steady growth for the last 6 years. This growth has further increased due to the pandemic in 2020 and 2021 which saw 520 million additional users on multiple social media sites. While this growth has led to many benefits in the internet space, many of the negative and harmful activities that were already taking place have grown as well. One of these harmful activities that has had a huge effect on the youth of India is Cyberbullying. Cyberbullying or cyber harassment is a form of bullying or harassment using electronic means. It is in many cases an extension of already existing traditional bullying. In this paper, we present a novel cyberbullying detection and prevention system that utilizes machine learning algorithms and natural language processing techniques to analyze online communication data and identify instances of cyberbullying. Our system uses a combination of keyword-based and context-based approaches to detect potential cyberbullying behavior, and then applies various prevention strategies, including automated warnings and blocking of abusive users. Identify applicable funding agencies here. If none, delete this. We evaluate the effectiveness of our system on a dataset of real-world cyberbullying instances, and show that our system achieves high accuracy in identifying cyberbullying behavior, while minimizing false positives and false negatives. Our system represents a significant step forward in the development of effective cyberbullying detection and prevention tools, and has the potential to make a meaningful impact in reducing the harm caused by cyberbullying.

II. LITERATURE SURVEY

A. Survey of Research Papers

As a part of our project, we conducted a thorough literature survey of nine research papers related to cyberbullying detection using machine learning techniques. We focused on papers that used Naive Bayes algorithm for cyberbullying detection in social media data. The papers were from various reputed conferences and journals in the field of machine learning and natural language processing. The survey helped us gain a deeper understanding of the different approaches used in cyberbullying detection and the challenges faced by researchers in this domain. It also helped us identify the gaps in the existing literature and the opportunities for further research in this field. The knowledge gained from the survey was instrumental in guiding our project and making informed decisions related to the selection of datasets, feature extraction techniques, and evaluation metrics for our model. Conducting a literature survey of 9 research papers has given us a broad perspective on the current state of cyberbullying detection methods and the various techniques used for this purpose. It has also helped us identify the limitations of existing approaches and the scope for improvement. We have gained insights into the importance of feature selection, the role of machine learning algorithms in cyberbullying detection, and the need to address the issue of bias in the training data. Furthermore, the survey has helped us understand the potential impact of our project and its implications for social media platforms and their users. Overall, the literature survey has been an enriching experience that has provided us with a solid foundation for our project.

TABLE I. TABULAR SUMMARIZATION OF OUR LITERATURE SURVEY

Paper No.	Aim/Objective	Algorithm/ Model used	Target	Architecture details	Result
1	To detect cyberbullying which uses machine learning algorithms to detect bullying messages for English, Hindi and Marathi.	Multinomial Naive Bayes (MNB), Logistics Regression (LR), and Stochastics Gradient Descent (SGD).	To detect cyberbullying in different languages.	MNB, LR, and SGD were selected as they perform well on Topic Modeling and Text Classification as well as in literature.	A model for detecting cyberbullying in the three languages was developed and Logistic Regression was found to outperform all other algorithms on the datasets.
2	A cyberbullying detection framework that leverages human-like behavioral patterns using Reinforcement Learning and combining several Natural Language Processing techniques to guide the detection process.	Reinforcement Learning	To provide a more accurate model to detect cyberbullying.	Reinforcement Learning approach incorporates trial & error to learn and accomplish the goal it has been assigned.	The model outperforms several other baseline models when the dataset is dynamic and populated with never before seen words and deliberately misspelled words.
3	To investigate the automated identification of posts on social media related to cyberbullying by considering two features Bag of Words (BoW) and term frequency-inverse text frequency (TF-IDF).	Decision Tree (DT), Random Forest, Support Vector Machine, Naive Bayes	To use different machine learning models and see the results.	DT, RF, SVM, NB are used to classify comments as bullying and non-bullying.	It has been observed that SVM exhibits better performance accuracy compared to the other classifier algorithms.
4	To explore the current Twitter cyberbullying detection techniques, and develop a new classification method based on deep learning.	Convolution Neural Network (CNN)	To develop a more sophisticated method for classifying text data.	Convolution Neural Network (CNN) classifies tweets in a more intelligent way than traditional classification algorithms.	CNN was found to classify tweets more intelligently than traditional algorithms.
5	To develop a model for detecting cyberbullying using machine learning.	CNN	To make a more precise text examination tool using CNN for detection of cyberbullying.	CNN is used with multiple layers which provide a process of iterative analysis over different layers to provide an efficient and accurate analysis.	The system uses a accurate method of CNN implementation using keras and helps in achieving precise results.

Paper No.	Aim/Objective	Algorithm/ Model used	Target	Architecture details	Result
6	To identify a piece of data as cyberbullying or not, and then compare five machine learning techniques to discover the most accurate one.	Naive Bayes, Decision Tree, Random Forest, SVM, DNN Model.	To identify the best machine learning approach to detect cyberbullying.	Python for backend. HTML, CSS for frontend.	DNN model performs best with an accuracy of 0.990145 along with the second best performing, Random Forest algorithm with an accuracy of 0.986897.
Paper No.	Aim/Objective	Algorithm/ Model used	Target	Architecture details	Result
7	To use the Sentiments Analysis (SA) approach to categorize tweets on cyberbullying as positive, negative, or neutral.	Support Vector Machine (SVM) and Naïve Bayes (NB).	To eliminate negative messages/ statements made on social media platforms.	Python NLTK	SVM classifiers outperformed NB on almost every performance measure.
8	To detect offensive messages on Arabic Twitter-streams and classify them according to their strength by assigning weights to each bullying message.	Twitter API	To classify Arabic slurs to stop cyber bullying on Arabic Twitter-streams.	JavaCript, JSON, HTML, CSS, PHP	The product system performs appropriate actions according to predefined settings.
9	To design a novel scheme for analyzing Bangla content on social media by combining text analytics and machine learning algorithms and to compare the performance of the module with other available techniques.	NB, SVM, Decision Tree and KNN, REST API	To create a reliable Bangla content analyzer to prevent cyberbullying.	java crawler, WEKA, java.	Result indicates that the SVM based method achieves best accuracy and the performance improves if user specific data can be included.

III. MOTIVATION

The McAfee Cyberbullying Report, which surveyed ten countries, including India, has revealed some critical insights into the trends and prevalence of cyberbullying. This report provides information on the types of cyberbullying being reported, data around who is most likely to be a perpetrator or victim of bullying online, and the differences in how parents and children perceive cyberbullying behavior. Disturbingly, India has been ranked as the country facing the highest rates of cyberbullying in the Asia Pacific region, surpassing Australia and Japan. The report indicates that almost 8 out of 10 individuals in India have experienced various types of cyberbullying, and around 63% of them are children. These statistics are staggering and indicate a pressing need for measures to be taken to prevent cyberbullying before it occurs. This is essential to safeguard the welfare of thousands of children and make the Internet a safer place for them. Given the high prevalence of cyberbullying in India, it is vital to develop and implement advanced technologies that can detect and prevent cyberbullying. Such technologies could help identify cyberbullying behavior and enable prompt action to protect victims and prevent future incidents. By reducing the number of cyberbullying cases, we can create a safer online environment for all, and this is a crucial step in protecting the welfare and mental health of young people in India.

IV. PROBLEM DOMAIN

Machine Learning: Machine learning is an exciting field that has revolutionized the way we approach problem-solving in various domains. However, it also poses several challenges that must be overcome to achieve effective solutions. One significant challenge in machine learning is the selection of appropriate algorithms for a given problem domain. The choice of algorithm significantly impacts the accuracy and efficiency of the model, and with the wide variety of machine learning algorithms available, it can be challenging to choose the most suitable one for a specific problem. This requires in-depth knowledge of the strengths and limitations of each algorithm, as well as the ability to evaluate their performance. Furthermore, selecting the right algorithm is not a one-time process; it may require multiple iterations and evaluations to identify the most effective algorithm for the given

problem. Another challenge in machine learning is the availability and quality of data. The accuracy and efficacy of a machine learning model are highly dependent on the quality and quantity of data used to train it. In many cases, collecting sufficient and relevant data can be a significant challenge, and ensuring its quality and accuracy can be even more challenging. This is particularly true in domains where data is scarce, and data collection is expensive. Furthermore, the data used to train a model must be representative of the problem domain; otherwise, the model's performance will suffer. Thus, data pre-processing, data cleaning, and data augmentation techniques are used to mitigate data quality issues and ensure the accuracy of the model. Machine learning models are also susceptible to bias, which can have a significant impact on their performance. Bias can occur at different stages of the machine learning pipeline, including data collection, pre-processing, and model training. This is especially true when training data is not representative of the broader population or when the model's output is used to make decisions that can affect people's lives. Addressing bias in machine learning models requires careful attention and ethical considerations. It is crucial to ensure that the training data is diverse and unbiased, and the model's output is transparent and explainable. Finally, the interpretability of machine learning models is a growing concern in many domains. As models become more complex, understanding how they arrive at their predictions becomes more challenging. This can make it difficult to identify the reasons behind a model's decisions, leading to a lack of transparency and accountability. Several techniques, such as feature importance analysis and model visualization, are used to address this challenge. These techniques help explain the model's behavior and provide insights into how the model can be improved. In summary, while machine learning presents exciting opportunities for solving problems in various domains, it also poses numerous challenges that must be addressed. These include selecting appropriate algorithms, ensuring data quality, mitigating bias, and promoting interpretability. By addressing these challenges, we can harness the full potential of machine learning to solve real-world problems effectively.

IV. PROBLEM STATEMENT

Social media platforms are confronted with a significant challenge of cyberbullying, which negatively affects the mental health of the victims and can lead to a decrease in user activity. The issue of cyberbullying also tarnished the reputation of social media sites, including Twitter, which has a large number of users generating an enormous volume of tweets. To regulate and control offensive and derogatory tweets, there is a pressing need to implement a system that can detect and prevent cyberbullying in real-time. To address the problem of cyberbullying on social media applications, a system that can detect and prevent such behavior is essential. This system should have the ability to identify and respond to cases of cyberbullying promptly, to prevent it from happening and to provide a safer and more secure environment for all users. By implementing such a system, social media sites can create a better online experience for their users and help to eliminate cyberbullying from their platforms.

Statement: Using a machine learning model to identify instances of cyberbullying on social media platforms.

V. PROBLEM FORMULATION OR REPRESENTATION OR DESIGN

Naive Bayes's Algorithm: Naive Bayes is a simple yet effective algorithm that can be used in cyber bullying detection systems. Naive Bayes is a probabilistic algorithm that calculates the probability of a particular data point belonging to a certain class. In cyber bullying detection, the algorithm is trained using a set of labeled data points that represent instances of cyber bullying or non-cyber bullying. The algorithm then uses this training data to build a model that can classify new data points as either cyber bullying or non-cyber bullying. To apply the Naive Bayes algorithm to cyber bullying detection, text-based features such as word frequency, sentence structure, and tone are extracted from social media posts or messages. These features are then used as input to the Naive Bayes algorithm, which calculates the probability that a message or post belongs to the class of cyber bullying or non-cyber bullying. The Naive Bayes algorithm assumes that the features are independent, which may not always be the case in real-world scenarios. However, in practice, this assumption often holds true and does not significantly affect the performance of the algorithm. One advantage of Naive Bayes is that it can work well even with limited training data, which is often the case in cyber bullying detection systems. Additionally, Naive Bayes is computationally efficient and can process large amounts of data in real-time. Overall, Naive Bayes can be an effective algorithm for detecting cyber bullying in text-based content. However, it should be noted that no algorithm is perfect, and a combination of different techniques may be required to achieve optimal performance in cyber bullying detection systems. For the time being we have implemented an admin login as a loophole to certain inefficiencies that we couldn't overcome. The purpose of the admin login is quintessential

human intervention for review of appealed posts and users which have been banned fortuitously. A systematic flowchart of our system is shown below:

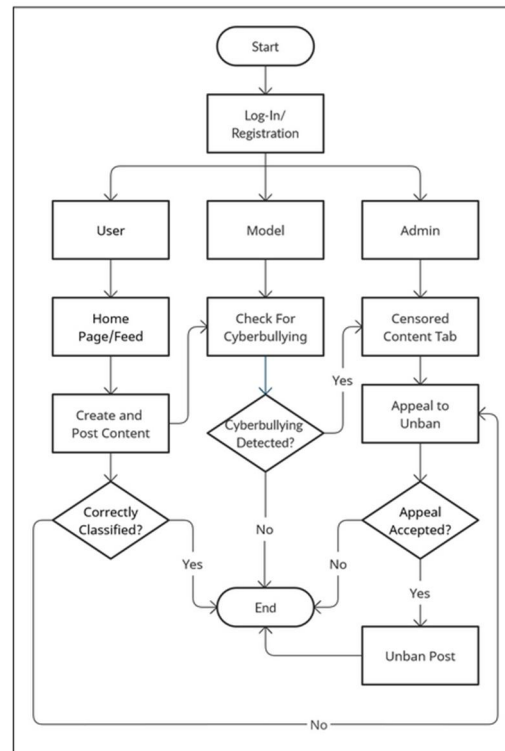


Figure 1. Flow Chart of our System

VI. SOLUTION METHODOLOGIES OR PROBLEM SOLVING

The Naive Bayes algorithm can use the TF-IDF approach for cyberbullying detection. The term frequency-inverse document frequency (TF-IDF) is a statistical method used to evaluate the importance of a word in a document based on its frequency and the frequency of the word across all documents in a corpus. In the context of cyberbullying detection, the TF-IDF approach can be used to identify words or phrases that are indicative of cyberbullying. First, the algorithm will extract the text data from social media posts, comments, or messages. Next, it will perform text preprocessing steps such as tokenization, stop-word removal, and stemming to clean the text data. Once the text data is preprocessed, the algorithm will calculate the TF-IDF score for each word in the text. This score reflects how important a word is in the document and how frequently it appears in the corpus. Words with a high TF-IDF score are considered more important than words with a low score. After calculating the TF-IDF scores, the algorithm can use them as input features for the Naive Bayes algorithm. The Naive Bayes algorithm will then use these features to classify the text as either cyberbullying or non-cyberbullying. The algorithm can be trained on a labeled dataset of cyberbullying and non-cyberbullying messages to learn the patterns in the data and improve its accuracy. In summary, the Naive Bayes algorithm can use the TF-IDF approach to identify important words and phrases in text data that are indicative of cyberbullying. By using these features as input, the algorithm can accurately classify messages as cyberbullying or non-cyberbullying, helping to detect and prevent cyberbullying on social media platforms.

VII. DATA MODEL

In order to train our model to detect cyberbullying, we utilized a dataset of 1847 examples of cyberbullying and non-cyberbullying tweets. This dataset was carefully curated to ensure that it contained a representative sample of different types of cyberbullying instances, such as harassment, threats, and derogatory comments. Additionally, the dataset included examples of non-cyberbullying tweets to ensure that our model could accurately distinguish between the two types of content. By using this dataset to train our model, we were able to develop an algorithm that could accurately classify tweets as either cyberbullying or non-cyberbullying with a

high degree of precision. This allowed us to create a system that can help to identify and prevent cyberbullying, making social media platforms safer and more enjoyable for all users.

Unnamed: 0			q_a	label
0	0	Justin bieber. He's a		1
1	1	its also funny how u stalked my whole twitte...		1
2	2	Hello ! How's ur day going? Good I hope :...		1
3	3	Hey. Why you such a		1
4	4	how old are u ? And how do you know michael?		1
...	...	...	...	...
1842	1842	Nahhh. only that the ocean is dispicably tur...		0
1843	1843	Not really. I wanna go back to gatliburg..i g...		0
1844	1844	Not that im aware of thankfully		0
1845	1845	a hemp bracelet my sister made me on my left ...		0
1846	11	Nope		0

1847 rows x 5 columns

Figure 2. Data Model

VIII. CONCLUSION

We will design an enhanced version of our system in the future that keeps polishing it's accuracy further. Based on the posts that the admin reviews and the incorrect classifications made by the system, these corrections are stored and reused as labelled data again to further add to the model's knowledge base. This process is automated and thus the model gets to learn more and more nuanced data on the go.

FUTURE WORK

In the future, we plan to incorporate natural language processing (NLP) and sentiment analysis into our cyberbullying detection system. By adding these techniques, we hope to improve the accuracy of our model by better understanding the context and tone of tweets. NLP can help identify specific patterns and features in text that may indicate cyberbullying behavior, while sentiment analysis can help determine the emotional tone of a tweet, such as whether it contains negative or aggressive language. These additional tools can enhance the effectiveness of our cyberbullying detection system and make it more adept at identifying instances of cyberbullying. Ultimately, we hope that this will lead to a safer and more positive social media experience for users.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude and appreciation to Prof. Mritunjay Ojha for his invaluable guidance, support, and encouragement throughout the course of our project. His expertise, insight, and willingness to provide constructive feedback have been instrumental in shaping our work and pushing us to achieve our best. We would also like to thank the computer department of our college for their assistance and resources, which have been essential in completing this project. Their commitment to providing a high-quality learning environment and state-of-the-art facilities has been instrumental in our success. Without their support and guidance, we would not have been able to complete this project successfully. We are truly grateful for their contribution and look forward to continuing our work with their continued support and guidance. We would also like to express our heartfelt gratitude to Prof. Lata Ragha, Head of the Computer Department, for her constant support and guidance throughout the project. We are also grateful to all the staff members of our college who provided us with the necessary resources and facilities to complete this project. Their support and encouragement were instrumental in the successful completion of this project.

REFERENCES

- [1] Rohit Pawar, Rajeev R. Raje, Multilingual Cyberbullying Detection System. Department of Computer and Information Science Indiana University-Purdue University Indianapolis.IEEE 2019.
- [2] Monirah A. Al-Ajlan, Mourad Ykhlef, Optimized Twitter Cyberbullying Detection based on Deep Learning. 2018 IEEE.
- [3] Prof. Vidya Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques. Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020).

- [4] Md Manowarul Islam, Md Ashraf Uddin, et. al. ,Cyberbullying Detection on Social Networks Using Machine Learning Approaches.2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering(CSDE).
- [5] Adya Bansal, Akash Baliyan, et. al., Cyberbullying Detection on Social Networks Using Machine Learning Approaches. International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 05 May 2022.
- [6] Jalal Omer Atoum, Cyberbullying Detection Through Sentiment Analysis. 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020 IEEE.
- [7] Alwin T. Aind, Akashdeep Ramnaney, et. al., Q-Bully: A Reinforcement Learning based Cyberbullying Detection Framework, 2008
- [8] Djedjiga Mouheb, Masa Hilal Abushamleh, et. al., Real-time Detection of Cyberbullying in Arabic Twitter Streams. 2019 IEEE.