International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Detection of Cyberbullying with Advanced Security

Mrs. R. Deepa[1], Ms S Yuvasri[2]

[1]Assistant Professor, Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India.
[2]II - MCA, Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur, Tamil Nadu, India.

ABSTRACT
Social media is a platform where many young people are getting bullied. As social networking sites are increasing, cyberbullying is increasing day by day. To identify word similarities in the tweets made by bullies and make use of machine learning and can develop an ML model automatically detect social media bullying actions. However, many social media bullying detection techniques have been implemented, but many of them were textual based. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naïve Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM Out performs Naïve Bayes of simila r work on the same dataset. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not.

KEYWORDS: Natural Language Processing, Support Vector Machine (SVM), Naïve Bayes, Tweet Analysis, Bullying Detection

1. INTRODUCTION

Social Media is a group of Internet based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content. Via social media, people can enjoy enormous information, convenient communication experience and so on. However, social media may have some side effects such as cyber bullying, which may have negative impacts on the life of people, especially children and teenagers. Cyber bullying can be defined as aggressive, intentional actions performed by an individual or a group of people via digital communication methods such as sending messages and posting comments against a victim. Different from traditional bullying that usually occurs at school during face to face communication, cyber bullying on social media can take place anywhere at any time. For bullies, they are free to hurt their peers' feelings because they do not need to face someone and can hide behind the Internet .For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media. As reported in [2], cyber bullying victimization rate ranges from 10% to 40%. In the United States, approximately 43% of teenagers were ever bullied on social media [3]. One way to address the cyber bullying problem is to automatically detect and promptly report bullying messages so

that proper measures can be taken to prevent possible tragedies. To add up a social media called Twitter, Social media a powerful platform where you can have full freedom on what one wants to express or say; whether a negative or a positive one. Suicide is the act of taking one's own life. Suicide is the second leading causeofdeathgloballyamongpeople15to
29 years of age, according to the 2014global report on preventing suicide by the World Health Organization [3]. Close to 800,000 people die due to suicide everyyear. For every suicide, there are more people who attempt suicide every year. A prior suicide attempt is the most important risk factor for suicide in the general population. The age-standardized suicide rate in the Philippines is 5.8 formale, 1.9 for females, and 3.8 for both sexes. The rate is based from the number of cases affected per sample size of 100,000 people [2]. It is a misconception that suicide and depression affect mostly the poor. Stories abound of the growing prevalence of serious depression and suicide incidents in colleges attended by middle-class and rich kids[4].

2. LITERATURE SURVEY

Deep KNN Based Text Classification for Cyber bullying Tweet Detection
AUTHORS: M.Nisha; J.Jebathangam Year:2022
Nowadays, cyber bullying is a serious issue that many businesses must deal with. Existing technology makes use of machine learning to automatically detect cyber bullying. Deep learning-based algorithms have been shown to achieve higher accuracy in text classification than existing methods. In this paper, we develop a cyber bullying tweets detection using machine learning algorithm. The proposed method reads the tweets and then classifies the texts relating to cyber bullying and blocks the users. The study uses a k-NN classifier integrated with Deep Learning and show how effective the model is over large text datasets than other methods. The results of simulation shows that the proposed method has higher rate of classification accuracy than the other existing methods.

1. Analysis of Tweets for Cyber bullying Detection

AUTHORS: Shipra Anil Mathur; Shivam Isarka; Bhuvaneswar Dharmasivam; Jaidhar
C. D. YEAR:2023
Cyber bullying takes place online on gadgets like smart phones and computers. Cyber bullying can occur through social media platforms. This paper presents a real-time cyber-bullying detection system for Twitter using Natural Language Processing (NLP)and Machine Learning(ML).Th system is trained on a dataset of cyber bullying tweets using several ML algorithms and their performance is compared. Random Forest was found to provide the best results after tuning. To achieve real-time analysis, Selenium was used to scrape tweets from a given Twitter account and store the timestamp of the already checked tweets. Additionally, an image captioning model was employed to generate descriptions for images posted on the account and compare them with user-written captions to filter out spam tweets. The proposed work aims to prevent cyber bullying and provides a valuable tool for online platforms to detect and remove harmful content. The results of this study have shown that the selection of appropriate ML algorithms and pre processing techniques significantly impact the performance of cyber bullying detection on Twitter. Our model sheds light on the appropriateness of different ML algorithms for the detection of cyber bullying.

3. EXISTING SYSTEM

In an effort to model the cyber bullying, Kelly Reynolds and April Kontosthatis, 2011[1] used machine

learning to train the data collected from From Spring. me, a social networking site, the data was labeled using Amazon Web service called Turk. The number of bad words were used as a feature to train model. In a study by Dinakar et al [2], states that individual topic sensitive classifiers are more effective to detect cyber bullying. They experimented on a large corpus of comments collected from Youtube.com website. Ellen Spertus [3]tried to detect the insult present in comments, they used static dictionary approach and defined some patterns on socio-linguistic observation to build feature vector which had a disadvantage of high false positive rate and low coverage rate.

4. METHODOLOGY

1. Requirement Analysis

1. Problem Definition

The rapid growth of social media platforms and online communication has increased the occurrence of cyber bullying activities. Harmful comments, abusive messages, and offensive posts negatively affect users, especially students and young individuals. Traditional monitoring methods are not sufficient to identify and prevent cyber bullying effectively in real time. To address this issue, the proposed system is designed to detect cyber bullying content automatically and provide advanced security mechanisms using Machine Learning techniques and intelligent text analysis methods.

2. Functional Requirements

The system is expected to perform the following key operations:

- Collect and store textual data from social media posts, comments, chats, and online messages.
- Process and analyze the collected text to identify bullying-related patterns and abusive language.
- Classify the content as cyber bullying or non-cyber bullying using Machine Learning algorithms.
- Generate immediate alerts and notifications when harmful content is detected.
- Provide secure login and authentication features for users and administrators.
- Maintain reports and prediction history for monitoring and analysis.

3. Non-Functional Requirements

The system should satisfy the following performance and quality requirements:

- Ensure accurate and reliable cyber bullying detection results.
- Support fast processing of large volumes of social media data.
- Maintain user privacy and secure storage of sensitive information.
- Provide a user-friendly and responsive interface.
- Ensure system scalability and efficient performance during real-time analysis.
- Follow advanced security practices for protecting application data.

4. Data Requirements

For effective operation, the system requires access to:

- Publicly available cyber bullying datasets collected from social media platforms and online communities.
- Textual data including tweets, comments, posts, chat messages, and abusive content examples.
- Labeled datasets containing bullying and non-bullying categories for model training and testing.
- User activity records and prediction logs for monitoring and future analysis. Top of Form

5. PROPOSED SYSTEM

Twitter dataset may easier to extracted compared to other mediums such as Facebook , Instagram, and YouTube. Even though statistic brain. come aforementioned stated that cyber bullying occurred most in Facebook but only data from public profiles could be extracted easily such as Twitterthat the data is publicly available. The main function was to extract social media public data using available API. Then next step isto data cleaning and Pre-processing. As the extracted data had multilingual unstructured content along with lot of emoji, it was required to clean the data for higher accuracy. Several supervised machine learning algorithms were compared to identify the best one. Frequent use SVM by researches shows that SVM is popular among other classifiers in supervised learning approach. SVM is suitable for high-skew text classification such as to detect cyber bullying using content based features. Any circumstances such as missing data, type of feature and computer performance, SVM still perform other classifier.

6. SYSTEM ARCHITECTURE AND DESIGN

The system architecture represents a sequential workflow for detecting cyber bullying content with advanced security mechanisms. The process begins with collecting input data from social media platforms, online chats, comments, and user-generated posts. After preprocessing, relevant textual features are extracted and fed into Machine Learning models such as Support Vector Machine (SVM) and Naive Bayes for classification. Finally, the system produces output in the form of predictions by identifying whether the content is normal or cyberbullying and generates alerts for harmful activities to ensure user safety and security.

7. MODULES AND FUNCTIONALITIES

- The system follows a structured machine learning workflow for detecting and preventing cyber bullying activities on social media platforms.
- It is divided into multiple modules, where each module is responsible for a specific stage of model development and security management.

1. Dataset Preparation and Initialization

- Collect historical cyber bullying datasets from social media platforms, online forums, and chat applications.
- The dataset includes features such as comments, posts, messages, abusive words, user behavior, and bullying labels.
- Initialize the dataset for training and testing purposes.

2. Data Preprocessing

- Clean the dataset by removing unwanted symbols, punctuation marks, and duplicate records.
- Handle missing values and incomplete data entries.
- Convert text into lowercase format for uniform processing.
- Remove stop words and apply tokenization and stemming techniques.
- Convert textual data into numerical form using encoding and vectorization methods.

3. Feature Engineering

- Identify important textual features using TF-IDF Vectorization and frequency-based analysis.
- Select bullying-related keywords and meaningful patterns from the dataset.

- Apply statistical and correlation-based methods to improve feature selection.
- Use domain knowledge to enhance the detection accuracy of the system.

4. Model Training and Optimization

- Train machine learning models using the prepared dataset.
- Algorithms used include:
- Support Vector Machine (SVM)
- Naive Bayes
- Split the dataset into training and testing sets.
- Tune model parameters to improve prediction accuracy and overall system performance.
- Reduce false predictions through optimization techniques.

5. Model Evaluation and Selection

- Evaluate the performance of models using metrics such as:
- Accuracy
- Precision
- Recall
- F1-Score
- Compare the results of different machine learning algorithms.
- Select the best-performing model for cyberbullying detection.
- Test the selected model using unseen data to ensure reliability and efficiency.

6. Security and Alert Module

- Provide secure user authentication for system users and administrators.
- Encrypt sensitive user data and stored information.
- Generate alerts when cyberbullying content is detected.
- Maintain reports and logs for monitoring harmful online activities.
- Prevent unauthorized access through advanced security mechanisms.

7. Output Generation Module

- Display prediction results as:
- Normal Content
- Cyberbullying Content
- Generate warning notifications for detected bullying activities.
- Store prediction reports securely for future analysis and monitoring.
- Help administrators take necessary actions against harmful content.

8. IMPLEMENTATION AND INTEGRATION

1. Frontend Implementation

- The frontend of the system is designed using HTML5, CSS, Bootstrap, and JavaScript to provide a clean and user-friendly interface.
- It enables users to enter comments, messages, or social media posts for cyberbullying analysis.
- The interface is designed to be simple, responsive, and accessible for both technical and non-technical users.
- Users can securely log in to the system using authentication features.

- The prediction result is displayed instantly, showing whether the content is classified as cyberbullying or normal content.
- Alert messages and notifications are also displayed when harmful content is detected.

2. Backend Implementation

- The backend is developed using Python with the Django framework to manage application logic and system communication.
- Machine Learning models such as Support Vector Machine (SVM) and Naive Bayes are trained using cyberbullying datasets containing abusive and non-abusive text data.
- When a user submits text, the backend processes the request through preprocessing techniques such as tokenization, stop-word removal, and stemming.
- TF-IDF Vectorization is used for feature extraction before feeding the data into the trained machine learning model.
- Django handles routing, request processing, user authentication, and integration between the frontend and the machine learning model efficiently.
- Advanced security mechanisms such as encrypted password storage and secure login validation are implemented to protect user data.

3. Database Integration

- SQLite is used as the database to store and manage application data efficiently.
- The database records user-submitted posts, prediction results, timestamps, and generated alerts.
- User account details and login credentials are securely maintained within the database.
- The database ensures data persistence and maintains a history of analyzed content for future reference and monitoring.
- Stored data can be used for performance evaluation, future analysis, and improving the cyberbullying detection model.
- SQLite is selected because of its lightweight nature, simplicity, fast performance, and easy integration with the Django framework.

9. EVALUATION AND RESULTS

Cyberbullying detection uses advanced machine learning techniques to identify harmful content in online platforms. It involves natural language processing and text analysis to understand user messages and detect abusive language. The system classifies data into bullying and non bullying categories for accurate prediction. Sentiment analysis is also used to analyze emotions in the text. Continuous monitoring helps in early detection of cyberbullying activities. Security evaluation is performed using metrics like accuracy, precision, and recall. These measures ensure the system is reliable and effective. Overall, it helps create a safer online environment.

10. CONCLUSION

An approach is proposed for detecting and preventing Twitter cyberbullying using Supervised Binary classification Machine Learning algorithms. Our model is evaluated on both Support Vector Machine and Naive Bayes, also for feature extraction, used the TFIDF vectorizer. As the results show us that the accuracy for detecting cyberbullying content has also been great for Support Vector Machine of around 71.25% which is better than Naive Bayes. Our model will help people from the attacks of social media

bullies.

REFERENCES

1. Rice, Eric, et al. “Cyber bullying perpetration and victimization among middle-school students.” American Journal of Public Health (ajph), pp. e66-e72, Washington, 2015.
2. Bangladesh Telecommunication Regulatory Commission, http://www.btrc.gov.bd/content/internet-subscribers-Bangladeshjanuary-2018, [Last Accessed on 18 Mar 2018].
3. Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla
4. Rice, Eric, et al. “Cyber bullying perpetration and victimization among middle-school students.” American Journal of Public Health (ajph), pp. e66-e72, Washington, 2015.
5. Bangladesh Telecommunication Regulatory Commission, http://www.btrc.gov.bd/content/internet-subscribers-Bangladeshjanuary-2018, [Last Accessed on 18 Mar 2018].
6. Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508
7. Dani Harsh, Jundong Li, and Huan Liu, “Sentiment Informed Cyberbullying Detection in Social Media” Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
8. Spinger,Cham,2017
9. Dinakar, Karthik, Roi Reichart, and Henry Lieberman. “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web 11.02(2011):11-17
10. K. Dinkar, R. Reichart and H. Liebernman, “Modeling the Detection of Textual Cyberbullying,” MIT. International Conference on Weblognd Social Media. Barcelona, Spain, 2011.
11. web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5,