Automated Cyberbullying Activity Detection using Machine Learning Algorithm

Vedadri Yoganand Bharadwaj1, Vasamsetti Likhitha1, Vootnoori Vardhini1, Adari Uma Sree Asritha1, Saurabh Dhyani2, M. Lakshmi Kanth3

1 Department of CSE (AI & ML), GRIET, Hyderabad, Telangana State, India
2 Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India
3 KG Reddy College of Engineering & Technology, Hyderabad, India

Abstract. Cyberbullying is the use of technology to harass, intimidate, or harm another person by making hurtful comments, sending threatening messages to humiliate someone in social media. It is important to recognize the signs of cyberbullying activities and take steps to prevent it. Automated Machine Learning algorithms and Text mining concepts for detecting and classifying bullying messages in social media environment. The abusive texts are clustered using Multinomial Naive Bayes, LinearSVC, Logistic Regression, K -Nearest neighbour to build a classifier from training datasets. Implementation uses Suspicious-communications-on-social-platforms dataset.

1 Introduction

In today’s digital age with the emergence of social media platforms and online communication, cyberbullying has grown to be a serious problem. It describes the practice of intimidating, harassing, or harming people or organizations through technology and online platforms. Among the serious effects of cyberbullying are emotional pain, social isolation, and even self-harm or suicide. Machine learning techniques have become important resources for the detection of cyberbullying in response to this urgent issue. In order to spot trends and indicators of cyberbullying behaviour, machine learning algorithms can analyse enormous volumes of online data, including text messages, comments, and social media posts. Machine Learning models can assist in preventing harm to people and fostering a safer online environment by automatically identifying and highlighting instances of cyberbullying. Automated Machine learning models are able to successfully identify harmful language on social media and to safeguard teenagers in order to overcome these limitations.

2 Existing methods

The paper [1] to enable automatic detection and prevent cyberbullying on social media platforms is the technology created by Andrea Perera and Pumudu Fernando's

* Corresponding Author: yoganand.bharadwaj@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
3 Proposed method

3.1 Problem Statement

Cyberbullying is a problem in the current world that negatively impacts many people, particularly young people. The availability of online content and the fluidity of online interactions make it difficult to recognize and address instances of cyberbullying using conventional methods. A system that can rapidly and precisely identify cyberbullying and take action to stop it is urgently needed as a result of this. Therefore, the goal of the paper is a machine learning based system that accurately detects instances of cyberbullying in a variety of online communication channels while taking into account contextual complexity, evolving language, biases, and privacy concerns. The objective is to offer a strong and expandable solution that can support prompt intervention and foster a safer online environment for people who are susceptible to cyberbullying.

3.2 Objectives

1. To understand the dynamics of cyberbullying, and to identify effective strategies for preventing and mitigating its effects.
2. To implement Automated Machine Learning algorithms that can effectively detect cyberbullying.
3. To prevent cyber bullying and promoting wellbeing by reducing depression, anxiety and other mental issues.

4. To enhance online safety and security making the internet a safer and more positive environment for everyone.

3.3 Connectivity diagram

![Connectivity Diagram]

Fig. 1. Connectivity diagram.

1. **Data collection module**: Data collection is a fundamental step in this process of gathering and analyzing information for the purpose of decision-making. Data Collection is responsible for collecting data from datasets suspicious communications on social platforms dataset.

2. **Data processing module**: Data processing refers to the manipulation, transformation, and analysis of collected data to extract useful information, identify patterns, and draw conclusions. In preprocessing, datasets are cleaned and preprocessed to remove unwanted characters and making it free from outliers, noisy data, and incomplete data.

3. **Feature extraction module**: Feature extraction describes the procedure of parsing raw data into a collection of significant and representative features. Reduced dimensionality is achieved through feature extraction. To train the models, extract the features from the preprocessed data, like frequency of each word or n-gram.

4. **Tokenization module**: A text file is tokenized when it is broken up into smaller, tokenized parts. It is possible to carry out additional analysis, feature extraction, and modeling for NLP tasks with a correct tokenization procedure. It allows us to recognize and examine terms that could be indicators of online bullying.

5. **Classification module**: Classification depends on problem, the type and size of the data, computational efficiency, and data distribution. The classification process typically involves two main steps: training and prediction. Classification involves categorizing text data into different classes and determine whether text contain cyberbullying word or not.

6. **Training model module**: Training a model involves the process of teaching the model to recognize patterns and make accurate predictions or classifications based on the provided training data. Train Multinomial Naive Bayes, LinearSVC, Logistic
3.4 Architecture of the proposed work

Two system modules and two user modules are shown in this architecture diagram. We have a system and database in the system module, and we’ll be extracting the pre-trained dataset from kaggle.com. The database is directly connected to the system. The four main modules that we use are the admin module, User module, Database module, and Machine learning module. Admin can handle and make changes whereas user can login the page and they have access to select the model which are provided and the user can view the performance analysis. A database is another module that stores data along with outcomes of whether they are objectionable or not.

![Diagram of architecture](image)

**Fig. 2.** Proposed framework.

4 Methodology

The proposed work suggests a technique to spot automated cyber bullying on social media that considers both syntactic and emotional elements. Before labelling it consider the sentence’s semantic and satirical content.

- **Data Collection:** Data Collection is responsible for collecting data from datasets suspicious communication on social platforms dataset. Compile a broad dataset made up of text messages, comments, social media posts, and any other type of online communication that might include incidents of cyberbullying. To achieve balanced training, it should contain both cyberbullying and non-cyberbullying examples.

- **Data Preprocessing:** In preprocessing, datasets are cleaned and preprocessed to remove unwanted characters and making it free from outliers, noisy data and incomplete data. To prepare the raw text for analysis, it must first be cleaned up of any extraneous information, such as special characters or URLs, using methods like tokenization, stemming, and stop-word removal.

- **Feature Extraction:** Extract the features from preprocessed data, such as frequency of each word or n-gram in the text, to train the models. Word frequency, n-grams, sentiment analysis scores, syntactic patterns, and other linguistic or contextual characteristics may be used as features to identify cyberbullying behaviour.

- **Machine Learning Algorithms:** The model will learn probability of each feature given the class based on training data.
cyberbullying. It involves separating a continuous stream of textual information into tokens, which can be words, phrases, sentences, symbols, or other significant components.

- Classification: Classification involves categorizing text data into different classes and determine whether text contain cyberbullying word or not.
- Training Model: Train Multinomial Naive Bayes, LinearSVC, Logistic Regression, and K-Nearest Neighbour models on pre-processed data and features, model will learn probability of each feature given the class (offensive or non-offensive) based on training data. For supervised learning, the features found from the pre-processed data are utilised as input along with labels denoting instances of cyberbullying versus non-cyberbullying. Based on the patterns seen in the training data, the model learns to categorize new occurrences.
- Testing and Evaluation: Evaluate the performance of the trained model by testing it on a different set of data and measuring its accuracy, precision, recall, F1-score. To improve robustness and prevent overfitting, cross-validation techniques like k-fold validation may be used.

5 Results and discussions

5.1 Description about dataset

We have gathered data from Twitter and Facebook groups for this dataset, which is based on questionable behaviours like racism, prejudice, harsh language, and threatening behaviour most often seen in cyberbullying. Based on words that are often used in tweets and comments, the data is tagged. After data is scraped, suspicious data is manually assigned the identifier 1 and non-suspicious data the identifier 0. Up to 20,000 rows of sentiments make up the dataset. About 8,000 of the data have positive or neutral sentiment tags, showing that the data is not suspicious, while about 12,000 of the data have negative sentiment tags like racism, prejudice, or abuse.

Fig. 3.
5.2 Proposed method

The importance of cyberbullying detection using different machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and K-Nearest Neighbours (KNN), lies in their capacity to automatically identify and categorize cyberbullying instances in online content. In terms of detecting cyberbullying, the benefits and distinguishing features of each algorithm are listed below.

1. Linear SVC: Binary classification challenges are handled by the machine learning algorithm linear SVC (Support Vector Classifier). It is a special adaptation of the Support Vector Machine (SVM) technique for linearly separable datasets. The kernel trick is a method used by SVC to translate data to a higher-dimensional space where it can be linearly separable when the data is not linearly separable. As the name implies, linear SVC is created especially for datasets that can be separated into linear segments. It is based on the idea that the classes can be divided by a hyperplane or a straight line. To determine the ideal hyperplane, a linear kernel, which is a dot product of feature vectors, is used.

2. Multinomial Naïve Bayes: A supervised machine learning algorithm is Multinomial Naive Bayes. It is a probabilistic machine learning technique that is employed for classification problems, especially in applications involving natural language processing (NLP). Despite this presumption, it has been discovered that Naive Bayes classifiers perform well in actual use, particularly for text classification problems.

3. Logistic regression: For binary classification tasks, the widely used supervised learning algorithm logistic regression is utilized. Despite its name, logistic regression is more frequently employed for classification than for regression tasks. It employs a logistic function to model the relationship between a number of independent variables (features) and a binary dependent variable (target).

4. K-Nearest Neighbours classifier: For classification problems, supervised machine learning algorithms like the k-Nearest Neighbour (KNN) classifier are used. The number of neighbours used for classification in the KNN method is determined by the value of k. The decision boundaries are smoothed down by a higher number of k, but if the classes are not well-separated, it may also result in more misclassifications. In contrast, a smaller value of k may be more noise-sensitive but can still capture local patterns.

5.3 Experimental results

5.3.1 Target classes in the dataset

![Fig. 4. Tagging.](https://doi.org/10.1051/e3sconf/202343001039)
5.3.2 Training and testing

To get started, the code imports the train_test_split function from the sklearn.model_selection package. Dividing the Data: In order to divide the data into subsets for training and testing, we use the train_test_split function. The four distinct variables X_train, X_test, y_train, and y_test are each assigned a portion of the input data. The two input datasets being divided are X and Y. By setting the test_size parameter to 0.2, 20% of the data will be allocated to the testing subset, while the remaining 80% will be allotted to the training subset.

5.3.3 Model evaluation

Assessing a machine learning model's effectiveness and quality is the process of "model evaluation." It entails assessing the model's ability to generalise to previously unobserved data as well as its success in completing the tasks or goals for which it was created. Evaluation of a model is a useful tool for identifying its advantages, disadvantages, and overall efficiency in addressing a specific issue. We employ metrics like F1 score, recall, accuracy, and precision.

5.3.4 User interface

The following figures show the flow UI design and steps of cyberbullying detection, after successful login, the interface displays login successful. Then user needs to select display on the left page and select the model. User can compare the performance measures of all these models by selecting Performance analysis page on left side.

Fig. 5. User interface.

Fig. 6. Login successful.
6 Conclusion

Improved Accuracy of Detection
One of the outcomes of applying automated machine learning to cyberbullying detection is the appreciable increase in accuracy. Due to the increased accuracy, incidences of online harassment are more consistently identified, resulting in fewer false positives and false negatives. Real-time observation and prompt action systems for detecting cyberbullying based on machine learning allow for real-time monitoring of online information, offering a proactive strategy to combat cyberbullying. These tools can quickly spot instances of cyberbullying by automatically scanning and analyzing text as it is shared or posted. This real-time capacity enables prompt replies to incidents.
safeguard potential victims and lessen the negative impacts of cyberbullying, enabling timely intervention.

Efficiency and Scalability

Machine learning methods are very scalable for the identification of cyberbullying since they can quickly process and analyse massive volumes of data. It is difficult to carefully monitor and identify cases of cyberbullying given the volume of internet content being produced. Machine learning is an essential tool in combating the pervasiveness of cyberbullying across the digital world because of its scalability and efficiency.

References


