IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Cyberbullying Messages on Social Media

[1] Mr. Bhushan Nandwalkar, [2] Ms. Nikita Hire, [3] Ms. Damini Mahale,[4] Ms. Pallavi Patil

[1] Assistant Professor, [2] BTech Student, [3] BTech Students, [4] BTech Students

[1] Dept. Computer Engineering,

[1]SVKM's Institute of Technology, Dhule, India

Abstract: Social channels have expanded in popularity as a result of the rapid advancement of internet technology, yet they have built a robust to notoriety as the most major streaming platforms in the twentieth century. An enormous amount of data and information is being returned from social networks, which is being used to develop a variety of exploratory research design for several methods of studies, such as human social behaviour, system security, and sociology. Cyberbullying is a concern that affects both college students on the internet. It has led to traumas such as suicides and depressions. The desire for content governance on social media networks is growing. Cyberbullying frequently causes chronic and disabling discomfort, particularly among women and children, and can even lead to suicidal ideation. Because of its strong negative social impact, Cyberbullying draws attention. Various cases of online bullying have happened recently around the world, such as the sharing of private chats, accusations, and vulgar insults. As a result, experts are bringing awareness to the detection of bullying speech or messages on social media. Therefore we proposed a methodology to detect the most common type of social media crimes such as Cyberbullying or online abuses that involve the exploitation of social media data. By combining linguistic communication with machine learning, the framework aims to design and build a good technique for observing online abusive and bullying texts. The goal of this study is effective applications of machine learning to develop a suitable methodology for detecting. In order for the suggested system to produce higher accuracy results, various like Naive Bayes, Random Forest, Linear Regression, and Svm Classification techniques are applied.

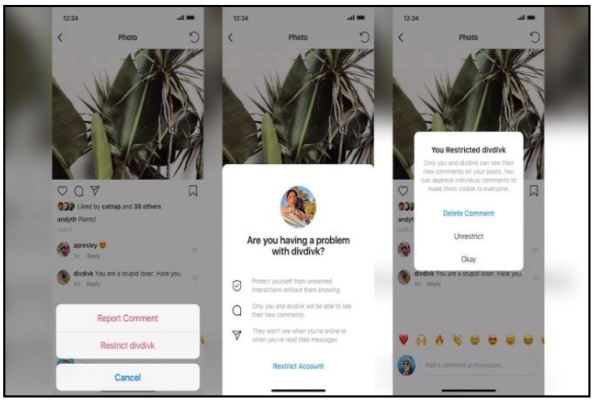
Index Terms **- Cyberbullying, Natural Language Processing, Machine Learning, Chat Application.**

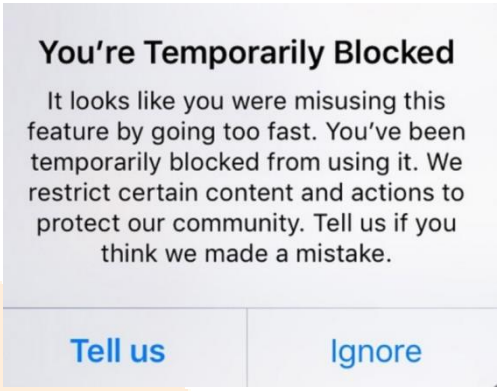
I. INTRODUCTION

Nowadays, social media platforms such as Facebook, blogs, wikis, Instagram micro blogging, and Twitter play an important part in formal communication and unstructured communication. The tremendous popularity of social networking sites on the internet sites raises a number of key questions about their effectiveness Safety and usefulness have an impact on people's social lives. A growing number of online users are abusing the system. In addition to harassing, threatening, and frightening other users, disseminating false information, is resulting in a flood of misinformation. Incidents of Cyberbullying Sites for social networking are fantastic resources. Individual communication is important. Making use of social media however, in general, it has become more widespread over time. People come up with immoral and unethical strategies to do unpleasant things. This is something we see between teenagers and, on sometimes, between adults.

In this study, we emphasize on Instagram because it is the social media channel with the most users reporting Cyberbullying. Instagram is a popular social media platform. A platform that allows users to exchange photos and videos in a variety of ways they can communicate with their fans either publicly or privately. Instagram users have the ability to upload photographs or videos together with a text caption, geotags and hash tags to aid in the discovery of new places, their photographs. They can also follow the feeds of other users like or comment on the images of other people.

Crime occurs all across the world, for a variety of reasons increase in the number of crimes agencies of law enforcement are requiring modern information systems with the ability to contribute to the reduction of crime and the protection of society. The scientific study of crime is known as criminology. Users will now be asked if they are sure about submitting potentially harmful comments on the social media app. The site will also get a supplementary function called 'Restrict,' which will safeguard users from inappropriate encounters such that the community can 'stand up to some of this bullying. After multiple warnings for misbehaving Instagram blocks the user.





By gathering and analyzing data, we can learn about the causes of crimes. Natural Language Processing is a useful method for text analysis in this way.

II. LITERATURE SURVEY

There has been a lot of research on finding ways to detect cybercrime on social networking sites. Crimes occur all across the globe, and humanity always triumphs. The vulnerability of cybercriminals is rapidly increasing in our day, especially among the younger generations. As a result, over the years, researches have been done to better understand or classify crime data. In this section, we explain earlier but adequate responses to this problem succinctly. Here are a few examples of research papers:

The comprehensive study in this journal looked at a selection of cybercrimes and evaluated a number of studies on their occurrence rate as well as some of their shortcomings. In this study, the current state of the humanities was examined, and a comparison was done using columnar statistics to assess their outcomes and identify their individual strengths and flaws. [1]
The study examined into cyber behaviour and patterns of communication using Data from twitter and 3 predictive modeling classifiers with a variety of word models. The suggested system is made up of three parts. The first unit comprises tweet which was before, which is utilized as an input to the second module, which creates a classification classifier, and the third module, which is final module. [2]

In this study, multiple social media evidence was used to explore and evaluate the language in order to define it as a challenge or non-threat. They proposed a methodology that is simple to adopt and will contribute to development of abilities to detect new threats and acquire users in a variety of ways. [3]

The survey's main goal was to provide a novel sequencing relevant statistical formulation for tackling the problem of Online bullying detection by assessing features intelligently and effectively. They developed a framework that evaluates the possibility of a message being annoying with high precision while adjusting for the new framework effort in obtaining a highly accurate result. [4]

The researchers of this article proposed Cyberbullying detection architecture. They discussed the structure for two data sets on Twitter: abuse speech to text and Wikipedia personal assaults. Hate speech detection was found to be accurate using Natural Language Processing methods. [5]

This study employs a scalable approach for progressively creating decision trees for detecting cybercrime, with the classification model set changing dynamically. Bootstrapping builds different tiers of the tree in a single scan of the training dataset, resulting in a significant performance improvement over previous decision tree techniques. The proposed approach has a 94.67 percent accuracy rate and effectively detects fake rate anomalies. This study focused on detecting anomalies and misuse at the user level. To accomplish extremely secure transactions in the future, we will develop this technology to identify cybercrime at a distributed level by profiling system activity. [6]
They detail the team's work in identifying hate and inflammatory literature in English, Hindi, and Marathi in this study. On English, Marathi, as well as Hindi language dataset contributed by the organizers, the problem of recognizing and objectionable

content is investigated experimentally. Our team achieved successful performance for both subtasks in all three languages and used an ensemble of classification models: Nave Bayes, Random Forest, SVM, and Logistic Regression using soft voting. [7]

They explored into existing software tools for detecting profanity in textual content in this article. Profanity-filter, for example, uses more refined algorithms that function more accurately but have a performance bottleneck (time complexity). In this benchmarking, the proposed technique is somewhere between 300 and 4000 instances faster than profanity-filter. They also examined the methods for detecting profanity. [8]

An intelligent chatbot with text-based Deep Learning Algorithm Cyberbullying prevention was developed in this paper. For text-based Cyberbullying detection, long short-term memory is recommended. This model's projected outcomes range from 90% to 94 percent, and it identifies more accurately than machine learning algorithms. The DNN model which is based on LSTM performs better in prediction. [9]

III. METHODOLOGY

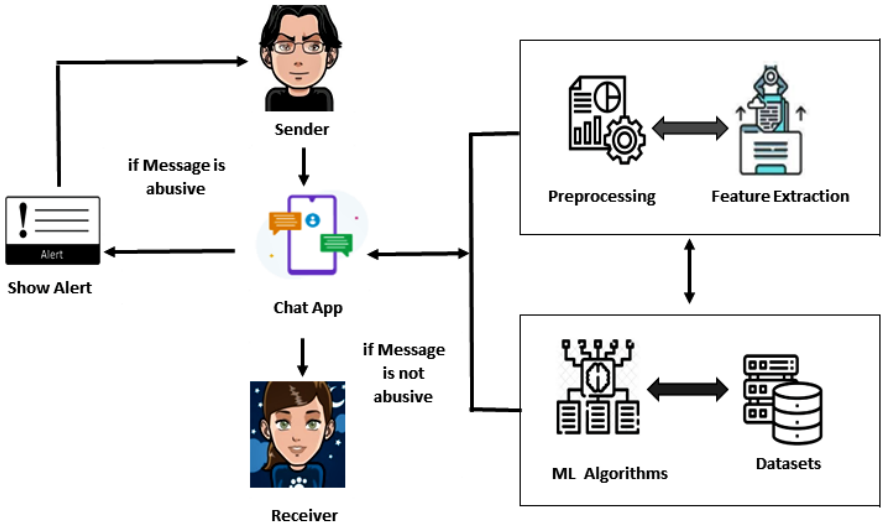


Figure 1 Architecture

Age groups and genders use modern information and communication technologies on a regular basis, which raises the possibility of disruptive behaviour like bullying. Abuse is one of the most devastating activities that people can have, especially as a child. Childhood, adolescence, and women are more likely to be bullied. Bullying has the capacity to affect people's mental and emotional well-being as well as change their personalities. Victim may get threatening behaviour twitter messages, texts, or posts promoting hate, harassing them, or putting their lives in danger.

We develop a method of dealing with these issues. Whenever user uses this chat application to send messages it will first check the message is whether offensive or not if the message is found offensive that is whether message contains abusive content it will automatically show an alert showing “it is very insulting” to the user who send the message. Vice versa if the message is found non offensive with no abusive content it will our chat app will send the normal message directly to receiver. The implementation's core format is split into three essential aspects.

1) The data collecting module is the first. This module is about We will gather all currently accessible data from social media. Facebook and Twitter or Kaggle

2) The Data Cleaning module comes next. This includes all of the data collected is processed to remove any extraneous information

3) The data is categorized in the classifier and sorted into user-defined classes based on the training data.

Once these three modules have been completed, the information can be displayed in the form alert box

3.1. Dataset :-

Through Kaggle, one can choose from a range of datasets. Our Cyberbullying detection dataset is made up of social media messages, comments, and tweets. We used Kaggle to collect a dataset of 253 abusive comments and messages from Twitter and Instagram for training. For our experiment, we manually created abuse datasets in three languages: English, Marathi, and Hindi for our experiment as follows:

Abuse
fucking
lesbo
chutiya
madarachod
behenchod
ganja
erotic
ghatiya aurat
masturbation
Porn

Table 1 manually created abuse dataset

3.2. Data Preprocessing: -

Information from the physical world is frequently insufficient, unpredictable, and/or missing in specific acts or habits, as well as including multiple mistakes. Pre-processing data seems to be a considered trying way to solve such concerns. Data pre-processing is the practice of creating original details for future processing. A workflow is used to process the datasets. All text data is initially converted to lowercase. Then "What's" and "Can't" become "What is" and "Can't," respectively. The string library is also used to remove all punctuation. The Natural Language Toolkit is then used to apply the relevant Natural Language Processing techniques:

- **Tokenization:** It is the process of breaking down text data into proper words or tokens.
- **Stemming:** It is the conversion of a term into a base word or stem.
- **Stop word removal:** They are terms that add no meaning to a sentence, such as what, is, at, and an in the English language.

3.3. Feature Extraction -

Feature extraction is essential in Natural Language Processing. Text data must be converted to raw values since classifiers cannot classify it. Thus every document (within that case, communications) can be parameterized, which can subsequently be used to classify them. After vocabulary has been created, all that is needed is to use a method of assessing features to change all of the documents depending on the vocabulary. It doesn't take into account context or word order, which might make a big impact in some circumstances. In addition, due to the increased quantity of features in huge datasets, vocabulary design becomes more challenging. "Is it intriguing," for example, is not the same as "It is interesting." The feature extraction method was used in this research is TF-IDF.

- ***TF-IDF Model:*** - In almost the same approach that the context of text notion develops a vocabulary, the TF-IDF approach does as well. The TF-IDF algorithm handles a problem that wouldn't be well here in the datasets but is essential for optimal feature extraction. When the frequency of a term in the same text increases, the frequency of TF-IDF rises, and when the proportion of texts in the corpus decreases, the rate of TF-IDF drops.

3.4. Building Machine Learning model –

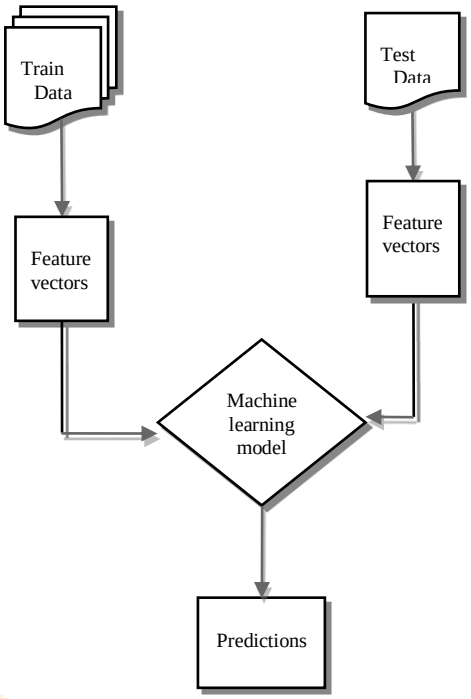


Figure 2 Proposed Model

Logistic Regression, Svm Classifiers, Random Forest, as well as Gaussian Naive Bayes are the four most common classifiers used in machine learning models.

Logistic Regression: It's a classifications model or perhaps a regression analysis. The parameter is a deterministic function being used model a real problems output.

$$sig(x) = \frac{1}{\{1 + \exp(-x)\}}$$

$$A = LT + C \qquad [5]$$

$$T(x) = sig(A) \qquad [5]$$

A simple regression analysis is used to assess the chances of passing or failing a program or scenario, which including passing/failing, winning/losing, or being alive/dead. The term "determine whether the data necessary is a flowers, vegetable, or fruit" could be used to indicate a wide range of operations, such as deciding yet if the collected data is indeed a species, vegetable, or fruit.

Random Forest: The Gaussian Mixture model is made up of several clustering algorithm classifiers. Thus every tree has a categorization hypothesis of its own. As a result, our classes are the most highly anticipated. This classification procedure is a training data method that generates accurate results since several decision trees are combined to obtain a conclusion. The random forest checks information estimates from each generating tree as well as selects the result simply based on the absolute majority of forecasts, rather than relying on a categorization algorithm. If there are two characteristics, A and B, and most decision trees estimated the component B of any occurrences, RF must choose the component B as follows:

F (x) = B is the majority decision of all trees. [5]

Support Vector Machine (SVM): In an essence, it's a svm classification technique that uses a tree model similar to classification. It is the only classification system that can discriminate between categories in n-dimensional space. As a result, SVM produces more accurate results with fewer incidences than other algorithms. In theory, SVM uses a kernel to adapt a data acquisition vector into the correct format by creating a series of separation hyper planes in an effectively infinite space. This approach, for instance, uses the standard prediction equation of any two cases as follows:

Sum(x xi) = K(x, xi) [5]

Naive Bayes classifier: The Bayesian theorem is used to create a machine learning method. The method generates assumptions based on the object's likelihood. The machine code is a type of computer code. As a result, using this strategy, problems with inter categorization can be simply overcome. It uses Bayes' Theorem to find the likelihood of an event happening given the likelihood of another already-occurring occurrence.

$$p(y|X) = \frac{p(X|y) \times p(y)}{X} \qquad [5]$$

The accompanying classifiers are implemented in Python using Scikit-learn, and the models and examples are classified using Scikit-learn. The purpose of this model's development is to determine the best separating hyper plane that maximises the training data's margin and aids classification. Because all these words come most commonly in offensive communications, the model learns whether word are "insulting" and how they are "insulting" through the training process. As a result, it's as though the training process removes the word "abusive" from all possible word permutations and employs it to make prospective predictions.

Our suggested algorithm successfully detects Cyberbullying text in specified messages on social media chat applications (messenger, twitter, Instagram, and Telegram). Individual abusive texts are identified from a batch of text messages. In a list of texts, sentiment analysis is accurate in identifying abuse and hatred. It assists in the classification of communications into good and bad categories. The goal is to classify a user's intents, which could take the form of messages, into two groups: "Offensive" and "Non-Offensive." whether the messages are harmful or not. Figure 5 depicts a sampling of output prediction.

Figure 3 Chat Application

Figure 4 Abusive Message

Figure 5 Alert

IV. RESULTS

Efficient screening metrics like as recall, precision, and F-measure are applied to quantify the effectiveness and integrity of any classification technique. Jupyter Notebook was used to conduct the experiments. The requirements for each classification model were tested on the test sets.

Accuracy (A): It's the weighted sum of estimates based on the number of accurate predictions.

$$A = \frac{\text{True positives}}{\text{Size of dataset}} \quad [5]$$

Precision (P): If any of the predictions made, apart from the learning algorithm, are genuinely positive?

$$P = \frac{\text{True positives}}{\text{True positives+ False positives}} \quad [5]$$

Recall(R): Number of times positive inputs were predicted true among all the positive inputs

$$R = \frac{\text{True positives}}{\text{True positives+ False negatives}} \quad [5]$$

F-measure (F): Evaluates the harmonic mean of precision to aid in comparisons of accuracy and recall.

$$F = \frac{2\times P\times R}{P+ R} \quad [5]$$

The following are the outcomes of our experiment:

4.1. Accuracy:-

Model	Test Accuracy	Precision	Recall	F1
Logistic Regression	82.35	0.82	0.82	0.82
Random Forest	80.39	0.80	0.80	0.80
Gaussian Naive Bayes	56.86	0.57	0.57	0.57
SVM	82.35	0.82	0.82	0.82

Table 2 Accuracy

With this we conclude our research with the following

1) Logistic Regression, Svm Classifiers, Random Forest, as well as other algorithms used in parameter-based Abusive message detection are 80-82 percent accurate, while Gaussian Nave Bayes is 56.86 percent accurate.

2) Except for naïve Bayes classifier, which has a precision of 0.57, all of these algorithms have a precision of 0.80 to 0.82, trying to make them a good fit for our requirements.

3) The above table compares the accuracy of all the algorithms we achieved with this model after implementation

4) And the below graph shows the comparison of accuracy of all the algorithms we achieved with this model after implementation:

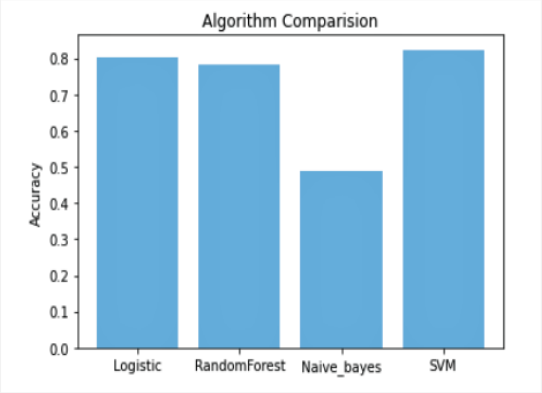


Figure 5 comparisons of algorithms

V. CONCLUSION

The essential aim of this application's training is to enhance societal behaviour toward using social platforms. Cybercrime has become much more frequent as a result of youngsters' increased usage of social media, compounding important social challenges. In order to reduce the negative consequences of cybercrime, it is vital to investigate the impact of something like the monitoring and management. We proposed a method for automatically detecting abusive messages. As a result, the overall project activity culminates in the monitoring of unethical behaviour, on social media, and indeed the reduction of danger or hazard. By overcoming this issue and ensuring the safety of society. In this framework, Machine learning techniques are applied generated greater accuracy results.

REFERENCES

[1] Wadha Abdullah al-chapter 1, Somaya al-maadeed 1, Abdulghani Ali Ahmed 2, Ali safaa Sadiq 3,4,and Muhammad Khurram khan 5, "Comprehensive Review of Cybercrime Detection Techniques", IEEE Conference 2020

[2] Zaheer Abbass, Zain Al, Mubashir Ali, Bilal Akbar, Ahsan Saleem, “A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning ”, IEEE Conference 2020

[3] Mahesh Mahat, “Detection of Cyber Crime on Social Media using Random Forest Algorithm ”, IEEE Conference 2019

[4] Mengfan Yao, Charalampos Chelmis, Daphney–Stavroula Zois, “Cyberbullying Detection on Instagram with Optimal Online Feature Selection” IEEE Conference 2018

[5] Varun Jain, “Cyberbullying Detection on Social Media using Machine Learning” IEEE Conference 2021.

[6] Md Manowarul Islam, “Cyberbullying Detection on Social Networking using Machine Learning Approaches” IEEE Conference 2018.

[7] Manveer Kaur 1, Sheveta Vashisht 2 , Kumar Saurabh 1, “Adaptive Algorithm for Cyber Crime Detection” 2012

[8] Anusha M D, H L Shashirekha “An Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages,2020

[9] Raktim Chatterjee, Sukanya Bhattacharya and Soumyajeet Kabi, “Profanity detection in social media text using a hybrid Approach of NLP and machine learning”, International Journal of Advance Research, Ideas and Innovations in Technology,2021

[10] Dr. Vijayakumar V and Dr Hari Prasad D , “Intelligent Chatbot Development for Text based Cyberbullying Prevention” International Journal of New Innovations in Engineering and Technology,2021