Cybercrimes on Social Media using Machine Learning

Nikita Hire
Computer Engineering
SVKM's Institute of Technology
Dhule, India
nikitashire@gmail.com

Prof. Bhushan Nandwalkar
Computer Engineering
SVKM's Institute of Technology
Dhule, India
nandwalkar.bhushan@gmail.com

Damini Mahale
Computer Engineering
SVKM's Institute of Technology
Dhule, India
daminimahale234@gmail.com

Pallavi Patil
Computer Engineering
SVKM's Institute of Technology
Dhule, India
palpatil1329@gmail.com

Jignesh Patel
Computer Engineering
SVKM's Institute of Technology
Dhule, India
jigneshdholu239@gmail.com

Abstract: Cybercrime is a serious concern that affects teens on the internet. The number of people who use these platforms is steadily expanding. The person who uses these platforms becomes increasingly exposed to the negative effects of having an online presence as their use increases. Cybercrime is one of the biggest negative repercussions of utilising social media, among many others. It has resulted in misfortunes such as suicide and depression. People are bullied and scammed online, which has a devastating impact on their mental health. Increased knowledge and information from social media will contribute to a variety of data patterns for various types of research, such as human social behaviour, system security, sociology, and so on. Despite the fact that social media platforms were created to share information and interact with family and friends, they have also been used to promote fake news, false narratives, hate speech, and abusive communications. We presented a methodology for predicting different forms of cybercrime on social media, such as Cyberstalking, Cyberbullying, cyber hacking, cyber harassment, and cyber frauds, using data gathered from social media.

Keyword: - Cybercrimes, Natural Language Processing, Machine Learning.

I. INTRODUCTION

Social media is a platform that allows people to share anything they want, such as images, videos, and documents, and to interact with others. People use their computers or smartphones to access social media. Facebook, Twitter, Instagram, and YouTube, among others, are the most common social networking platforms. Nowadays, social media is used in a variety of fields, including education, business, and charitable causes. In addition, social media is boosting the global economy by offering a slew of new work opportunities.

Although social media offers many advantages, it also has significant disadvantages. Psychopathic individuals utilize this medium to carry out illegal and misleading activities in order to hurt others' feelings and harm their reputation. Cybercrime has recently emerged as one of the most serious social media issues. Cyberbullying, often known as cyber-harassment, is a type of online harassment. Online crimes such as abuse and threat are emphasized. Online harassment has become more widespread as the virtual age has evolved and technology has advanced, particularly among teens. People develop a social media addiction. We have become increasingly reliant on social internet for both information and entertainment, and quitting is becoming increasingly difficult. According to studies, over 210 million people around the world from social media and the internet addictions, which is defined as an inability to unplug from the internet. Furthermore, no internet or social media platform is a utopia. False data is frequently used in some locations. The following are the most commonly targeted websites/apps for cybercrime:

1. Facebook
2. Instagram
3. Twitter
4. LinkedIn

Crime occurs all over the world, and security agencies are requesting advanced information systems to assist eliminate crime and safeguard society[1]. The empirical study of crime is criminology, which collects and investigates data to determine the reasons of crime. On social media, the following are the most common crimes: - Cyberbullying, stalking, and threats from the internet: People who are threatening, abusing, assaulting, and stalking others online are the most frequently reported and visible activities that emerge on social media.

Cyber-Bullying: Online bullying or suicides is a global public health issue and one of the leading causes of death amongst teenagers. Other factors that contribute to Cyberbullying include societal inequalities in demographics, cleanliness, and class.

Cyber-Stalking: Creating fake or imitated accounts to mislead others might be considered frauds depending on the actions taken by the fake profile user.

Cyber-Scams: Using social media is one of the most prevalent methods robbers exploit. It is thought to be harmful

if a person's leisure status updates are made public rather than constrained to friend groups.

Machine learning algorithms will be used to process the data collected from social media. For text analysis, we'll also leverage a Natural Language Processing technique...

II. LITERATURE SURVEY

Lots of research have been done to find solutions to detect cybercrimes on social media sites

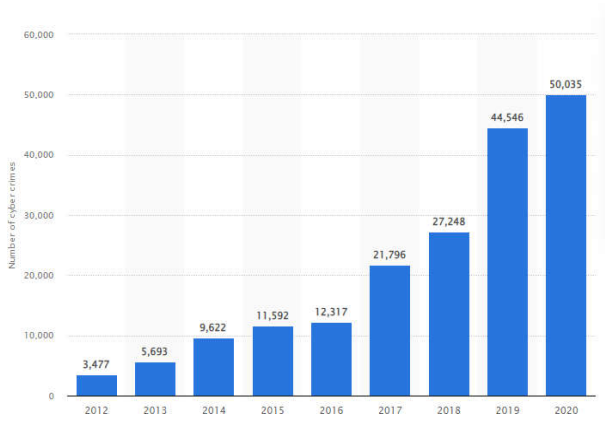


Fig 1. Number of Cybercrimes in India

Crimes happen all around the world, and humanity wins hands down. Particularly the younger generations, as the risk of cybercriminals is quickly increasing in our day. As a result, studies have been conducted over the years to better interpret or identify crime data. We explain briefly earlier and adequate responses on this topic in this part. Here are some research papers:

This journal's comprehensive review included a variety of cybercrimes and assessed a number of research on their prevalence rate along with some of the their flaws. The current status of the humanities has been analyzed in this research, and a comparison has been made using tabular statistics to determine their results and indicate their distinct merits and shortcomings. [1]

The goal of the study was to examine hacking behaviours and patterns of communication of hackers using Twitter data and three machine learning classifiers with such a bags of word models. Three modules make up the proposed system. The first unit is tweet pre-processing, which is being used as an input to a second module, which generates a trained model, and the third module is the final module. [2]

In this work, public social media data was taken into account and evaluated in order to investigate the language and classify it as a threat or non-threat. They suggested a model that is extremely easy to implement and will also be able to assist in the development of capabilities to classify new threats and capture users in different ways. [3]

The paper's primary idea was that a new sequence inferential statistical formulation was proposed to deal with that problem of Online bullying detection by analysing features intelligently and efficiently. They created a framework that calculates the likelihood of a communication being suggestive of harassing with high accuracy while compensating for the framework's effort in enhancing its chances of obtaining a highly correct result. [4]

The author of this research proposed an architecture for detecting Cyberbullying. They talked about the framework for two categories of data on Twitter: hate speech data and personal assaults on Wikipedia. Natural Language Processing algorithms were found to be accurate in detecting hate speech. [5]

Limitations in Existing systems:

1. Until present, the identification of cybercrime to be detected has not been done effectively in terms of the consequences.
2. There is no reliable method for identifying and detecting just taken in account abusive or offensive text.
3. Recognizing text-based content even when words in text are transformed by symbols.
4. Abuse detection is currently limited to the English vocabulary; content in other languages is not yet detected.

III. OBJECTIVE

The objective of this research is to generate a working model that can automatically find harassment and abusive behavior on social media and online forums by:

1. Collecting, extracting, and labeling the data set.
2. Improve accuracy by pre-processing, cleaning,and experimenting with various aspects.
3. Text, message, or post classification through oneof the several classes.
4. The best model's evaluation and analysis

The goal of this project is to learn how to apply Natural Language Processing and Machine Learning to a real-world problem, such as Cybercrimes and online harassment

IV. METHODOLOGY

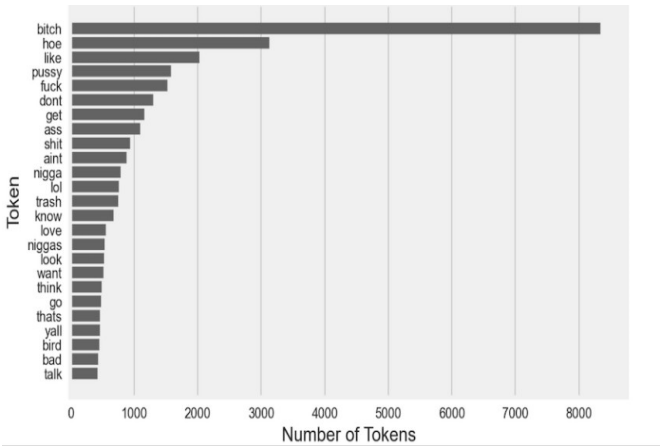
We propose an approach to manage these limitations. We'll collect data from social media sites like Twitter and Instagram, such as comments, sentiments, and links, among other things. Following the collection of this information, we will employ the acquired data to process the user's intentions. Following that, these expected objectives will be sorted into other categories of cybercrime, such as phishing and Cyberbullying. After classification, the outcome will be depicted in the form that identified cybercrime, its correctness, and whether or not it is offensive.

This approach evaluates texts and predicts offensive remarks using Natural Language Processing and Machine Learning. The framework includes data extraction from various online streams, pre-processing, and selection, as well as classification. Sentiment analysis is effective in detecting abuse and hostility in a list of comments. It aids in the categorizing of comments into positive and poor categories.

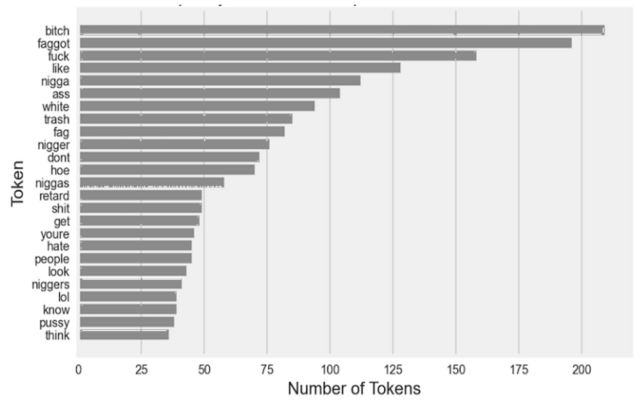
The purpose is to sort a user's content, which might be in the form of messages, into two categories: "Offensive" and "Non-Offensive." Also, figure out what kind of crime it is. The following phases are being used to develop a system for detecting cybercrime on social media:

1. Extracting, collecting, and labelling the data set:-

On Kaggle, there are a variety of datasets could choose from. The cybercrime detection dataset consists social media comments, posts, photos, videos, and phishing URLs.
These information will be provided input to the system for evaluation. We utilized Kaggle to generate a dataset from Twitter and Instagram that included 20000 abusive comments and 16k malicious links to predict the sort of cybercrime, such as bullying, cyber harassment, and phishing. The comments or responses were divided into two categories:
Non-Offensive Text: All of those comments or messages that are not offensive or positive. The comment "Your photo seems to be quite beautiful" is indeed an example of a positive and non-offensive comment.
Offensive Text: One such type relates to bully type comments or harassment's. "Go out bitch," for example, is a threatening text or message that we interpret to be a bad comment.



Frequency Distribution of Top 25 Tokens for Non-Offensive Text



Frequency Distribution of Top 25 Tokens for Offensive Text

'Bitch,' 'hoe,' 'pussy,' 'fuck,' 'nigga, "shit,' and 'ass' are among the top offensive words, but all of those identified with a lesser frequency are indeed represented in the non-offensive class. It indicates that the word 'bitch' is more frequently used mostly for insulting or neutral content than for hate text. In almost 20000 nasty remarks, the word 'bitch' emerges in a far higher percentage of non-hate material. In the non-hate subset, 'trash' appears without 'white,' while in the hate subset, 'white' and 'rubbish' appear in almost equal proportions. 'Niger(s)', 'white' ('trash'),'retard', 'homosexual', 'lgbt', 'fag' and 'nigger', and 'hatred' are virtually entirely in the non-offensive category. In simple terms, the data collected for machine learning is categorised into two levels: "1" for abusive offensive comments and "0" for non-offensive remarks. The end outcome would fall between [0, 1]. If the result is "1," the comment is categorically offensive. If it's zero, the comment is completely non-offensive

2. Pre-processing and cleaning using natural language processing: -

The data gathered from social media is unstructured and noisy. So we need to perform the following steps:-

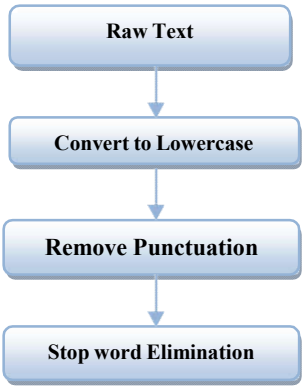


Fig 2 Data Preprocessing

The information gleaned via social media is unorganised and unreliable. As a result, we must take the following steps: -

Removal of Punctual Marks: - The elimination of punctual marks like the comma, semicolon, colon, question mark, punctuation mark, inverted commas, brackets, and hyphen from the dataset is the second stage in text processing.

Stop Eliminating Words: - Stop words are the most commonly used terms in the text. The value of terms that occur repeatedly in the text is quite low.

Feature Selection: - The selection of a appropriate feature produces a large amount of processed data. If efficient characteristics are chosen, it decreases training time, makes the model easier to read, improves accuracy, and improves the model's performance. Following the filtering, we prepare the texts' two most crucial features: -

i. TF-IDF: To pick characteristic terms from of the Social crime dataset, the TF-IDF approach is used. Phrase frequency-inverse Document Frequency (TF-IDF) denotes that if a term appears frequently in a document, it has less value and weighting in that document.

ii. Bag-of-Words: Machine learning algorithms can't operate with raw text directly. As a result, we must transform the algorithms to vector or numbers before using them. The processed data is then translated to a Bag-of-Words (BoW) format for the next step.

3. Classification using Machine Learning:-

To detect abusive messages, malicious links, and text, this module uses a variety of machine learning approaches such as Random Forest, Support Vector Machine, Nave Bayes, and Logistic Regression, as well as XGBoost. Various common ml algorithms using Python are described to detect cybercrimes using social media texts, with the classifiers with the precision predicting the final result.

Random Forest: Various decision tree classifiers make up the Random Forest algorithm. Each tree provides a class prediction on its own. Our final result is the highest amount of anticipated classes. Because numerous decision trees are blended to create the conclusion, this classification algorithm is a supervised learning model that produces correct results. Instead than depending on a decision tree classifier, the random forest considers the predictions out of each generating tree and selects the final output based on the absolute majority of forecasts. [6]

Support Vector Machine*:* The Support Vector Machine (SVM) is a supervised machine learning technique that, like such a decision tree, may be used for classification and regression. In n-dimensional space, it can differentiate the classes in a unique way. As a result, SVM gives a more precise result in less time than other algorithms. In practise, SVM creates a series of separating hyper plane in an effectively unlimited space, and it uses a kernel to transform a data input space into the appropriate format. [6]

Naive Bayes: A Naive bayes network is a condensed form of a comprehensive Bayesian learning network. As a result, it suffers from the same issue of being probabilistic. It's dubbed "naive" because it implies that the hypothesis is wrong to one another, which isn't the case in reality. [1] It is a Bayes theorem-based efficient machine learning algorithm. The programme makes predictions based on the likelihood of an object. This technique can be used to effectively answer consists of multi classification issues. [6]

Logistic Regression: The frequency of a given class or event existing, such as pass/fail, win/lose; alive/dead, or healthy/sick, is modelled using a logistic model. This can be used to represent a variety of events, such as assessing whether or not an input image a flower, vegetable, or fruit.

XGBoost: XGBoost the Algorithm uses a second-order estimate of the scoring function, which distinguishes it from previous gradient boosting algorithms. XGBoost may use this approximation to calculate the best "if" situation as well as its effects on performance. XGBoost. The Algorithms can and save the subsequent decision tree in its memory to avoid having recomputed. After Pre-processing and classification, the dataset is divided into two sets as 70% training and 30% testing Effectiveness and accuracy of each machine-learning algorithm to predict the accuracy, type of cybercrime, and class i.e. 'offensive' or 'non-offensive.

The suggested system is separated into two modules:

i. Hate Speech Detection

ii. Malicious Link Detection

Hate Speech Detection:

People of all ages and genders are constantly using social media and technology, which increases the risk of disruptive behaviour such as bullying. Abuse is among the most traumatic experiences a person can have, especially when they are young. Bullying is more common in childhood, teenagers, and women. Bullying has the potential to cause psychological and emotional harm as well as alter people's personalities. Victims may get threatening or abusive tweets, texts, or posts that advocate violence, harass, or endanger their lives.

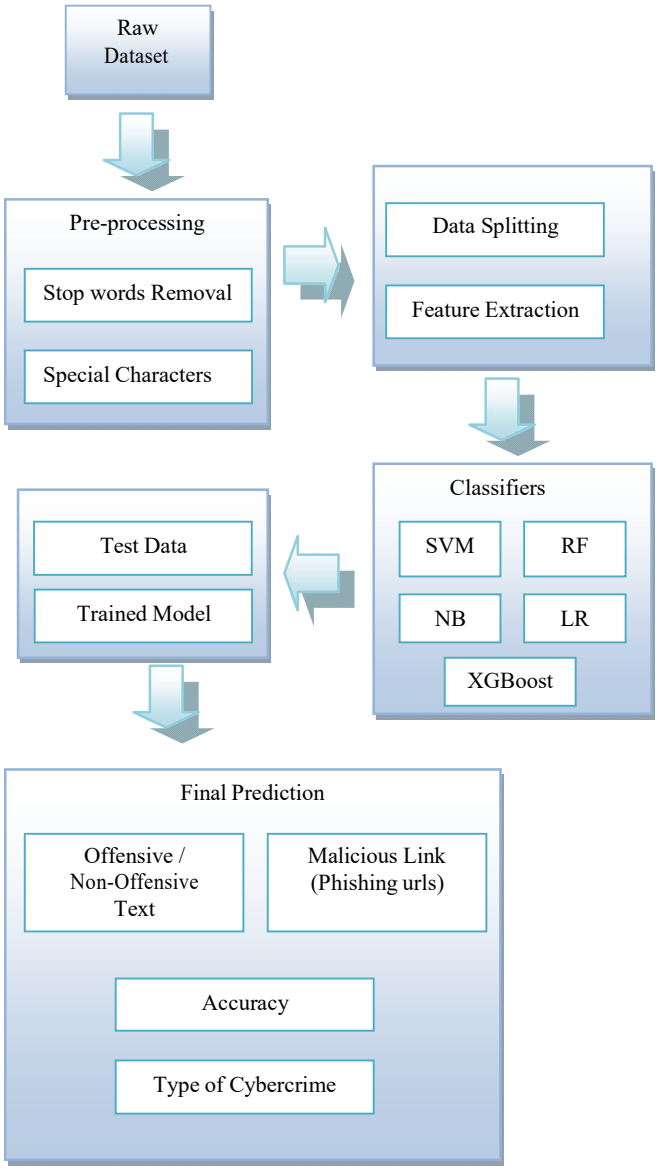


Fig3. Proposed Model

The algorithm for the proposed system is as follows:

START:

1. Gather the sample data, including comments and links.
2. Next, retrieve and label the data set you've gathered.
3. Next, natural language processing is used to perform pre-processing and cleaning.
4. Detect offensive messages, comments, and links using various machine learning algorithms such as Random Forest, Support Vector Machine, Nave Bayes, and Logistic Regression.
5. The final outcome will be predicted by the classification model with the best accuracy.
6. Go to step 9 if the final outcome is of the "offensive" kind.
7. Go to step 8 if the final outcome is of the type "non-offensive" class.
8. The data or comment is legitimate, and no cybercrime has occurred.
9. The comments or data is falsified, and cybercrime has occurred.
10. Determine the nature of the crime.

END

Detection of Malicious URL: Every social media user is now at risk from malicious links. It is a manipulative attempt to obtain sensitive information such as passwords, email addresses, passwords, and credit card numbers, among other things. The XGBoost algorithm is being used to identify phishing links. We chose the XGBoost method since it is both quick and accurate compared to other algorithms. For security reasons, multiple zeroes have been placed at the end of the each URL to prohibit it from being clicked. Phishers can obscure the dubious component of a URL in the urls bar by using a long URL. They are lured by the malicious site and fall victim. It's impossible to tell whether or not the connection is correct.
The proposed system's algorithm is as follows:

START:

1. Choose a sample data.
2. Next, retrieve and label the data set you've gathered.
3. Next, natural language processing is used to perform pre-processing and cleaning.
4. Now employ machine learning approach to harmful links, such as XGBoost.
5. The final outcome will be predicted by the classification model with the best accuracy.
6. Go to step 7 if the end outcome is of the type "Malicious link."
7. Determine the severity of the threat.

END

V. CONCLUSION

With the growing incidence of social media and increased internet use by teenagers, cybercrime has become more common and has proceeded to enhance significant social issues. It is critical to consider the impact of the monitoring system in order to avoid negative outcomes in cybercrime. We suggested a method for detecting cybercrime automatically. As a result, the total project activity concludes with the goal of monitoring unethical behaviour, preventing cybercrime on social media, and reducing risk or hazard. By resolving this issue and safeguarding society.

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.