Cyber Crime Analysis in Social Media Using Data Mining Technique

1M. Ganesan and 2P. Mayilvahanan
1Vels University, Chennai.
2Dept of Computer Sciences, Vels University, Chennai.

Abstract

The advancement in emerging internet technology and its widespread knowledge leads to security issues, cyber crime, internet hackers and intruders. The charisma of internet enhances the network structures that develop enormous online theft, fraud, are called as cyber attacks or cyber crime. In our paper we propose the cybercrime and the types of its illegal activities using the internet. Either national or international cybercrime creates huge negative impacts over society. We analyze the cybercrime web data due to social media using data mining techniques. Social networking websites are used as a medium of interaction and communication among all the people across the world. Due to over use and lack of awareness leads to increase the cyber crimes by several aspects. Data mining applied in acquiring negative impacts in social network sites. The data mining algorithm is used to analyze the crime data.

Key Words: Internet technology, cyber crime, social media, data mining technique, negative impact analysis.
1. **Introduction**

Internet and computer connects all over the world with web communication, information technology trends to digitalization. Generally Human weakness is exploited and addicted to the fascinating technology. Cyber crime, is defined as, a PC or hardware device used as a tool to dedicate the offense. It means criminal activity through the computer networks violating the rules & regulation and laws. The common examples of cyber crime are identifying theft, damage, transaction fraud, hacking, and software privacy. Cybercrimes are divided into two categories; Violent and Nonviolent cyber crimes. Most of the cyber crimes are nonviolent offenses, because it was occurring due to lacking of any physical contact interaction. Some of the nonviolent cyber crimes are cyber trespass, cyber theft and cyber fraud. The following diagram shows categories of cybercrime with various types are described in the below figure 1, The Web is an online medium that permits users from various bases to make a profile and connect with alternate users with similar sites. Person to person communication destinations, for example, Facebook, Myspace, Twitter, Orkut and so forth have turned out to be so famous among the general population that they have begun to share each and every snapshot of their lives on these sites.

Long range interpersonal communication in these sites is one of the least demanding types of correspondence nowadays and has turned into an unavoidable thing for younger generation. Each area of the general public is subject to these. In any case, social media have negative viewpoints moreover [1]. Social network media becomes a part of daily life experience in social networking sites and increases the users, which allows the users to share information, views with unknown (stranger) friends and can connect well with recognized friends. It can be extensively defined as internetbased social sites designed to facilitate conversation, collaboration and content sharing throughout networks of contacts, the maximum uses connects with current networks, making and increasing of friends creates an online presence for their users, viewing content, finding information, creating and customizing profiles and so on. It is in general view of how users are utilizing these sites like Facebook and other networking media.

![Diagram of Cyber Crime Categories](image-url)
Hackers without much of a stress get into and assemble their own and delicate data. Users are less aware and minimum worried about the security setting. The most efficient networking needs to be focus of protection and personality break. Absence of digital learning is the fundamental component of most important actualities because of which private photographs and individual data are shared between the systems. In this paper our fundamental Concentration is towards how to leave each of the issues of security in long range interpersonal communication. We have attempted to observe and crack information from each perspectives and feeling of vague users from instantly recognizable foundation.

The goal of individual vulnerability with protection and security framework is to check web violations related to the protection and character rupture on interpersonal interaction sites. Majority of the 170 respondents, 53 users (i.e. 31.18%) were using the Facebook for more than 5 years while the minimum number (3.52%) of the user has been using Facebook for one year and 5.29% users were using from less than 1 year, 8.23% of users were using for more than 2 years, 10.59% of users were using for more than 3 years, 15.88% of users were using more than 4 years, whereas 15.29% of users were using the Facebook for less than 5 years, and 10% of users were not using Facebook. By data mining technique we will be analyzing the cybercrime occurs through Facebook [2].

2. Literature Review

Many people can access this social networking site through iPhone, Android Phone, Tab, Laptop or other electronic gadgets. They can expertise their profile through posting any comment, uploading a photo, text or scrap posting, uploading of music and video in their profile to make the profile more attractive in front of their Facebook friends. By this site, users may choose to communicate through various digital objects are connected with friends who are staying far away from them.

Facebook users are used to access this social networking site regularly or time to create personal profile is very easy on the home page on the Facebook and there must no longer be allotted any registration charge whilst some of the new users need to create their profile or join with others in the Facebook community. Therefore, use of social networking site is very popular among the youngsters. The majority teenagers have contributed their time mostly in Facebook. Day by day social network site users are increasing all over the world. In 2015, nearly 2500 million users are available in India; they are having their profile account on this site. However, the maximum part of users is covered by teenagers in India. Impact of Facebook has a Social Networking Site (SNS).

Chew et al [3] targeted on how non-public data is being tormented by the internet and social media, and also discussed how the privacy end up at risk and a way to assign security recognition to prevent safety breaches. They highlighted the contemporary scenario for the use of the social community and
threats that can affect the customers. Finally, they stated some protection consciousness that may be practiced to be extra aware of social network threats.

Gangopdhay and Dhar have posted a document in which they have got noted that Social websites attract young adults and permit them opportunities to get along with regarded and unknown humans. Making pals with unknown human beings and including them to their pals list are probably taking into consideration as classy or as matters that can be confirmed off. So they targeted on how and to what extent the unveiling of private information by means of users is crazy. They additionally focused the security settings made by using the social networking sites like Facebook, Myspace, Orkut, Twitter and so on [4].

According to Information Technology Act, 2000 Cyber Crime is “the acts which is probably punishable via way of the Information Technology Act". It isn't exhaustive because the Indian Penal Code moreover offers many cyber crimes, together with email spoofing and cyber defamation, sending, threatening emails [5].

In Al-Janabi K. B. S., they, gives a proposed framework for the crime and crook records evaluation and detection using Decision tree Algorithms for records category and Simple K Means set of policies for information clustering. The paper has a tendency to assist professionals in coming across styles and developments, making forecasts, finding relationships and feasible reasons, mapping criminal networks and figuring out viable suspects. This type based on specific grouping the crimes, in line with the type, place, and time and different attributes; Clustering is based totally on locating relationships among exceptional Crime and Criminal attribute having a few formerly unknown commonplace traits [6].

F. Stutzman and J. Kramer-Duffield provide advice on how to enhance the privacy of users on social networking sites. To avoid identification of hacking, they advise making the user profiles, personal information for best friends, if you want to reduce the facts about security dangers on Social Networking websites [7].

A. Verma et al. proposed a decentralized and dispensed structure that preserves the privacy and protection of the customers in social networking sites. The higher the privacyits protection will be higher by the usage of a cryptographic approach like RSA and digital signature [8].

C. Marcum et al. suggested that users won't understand the risks associated with sharing non-public statistics or the potentiality to apply this record to predict exceedingly exclusive records like social security numbers [9].

Yabing Liu, et al., (2011) attempted to enhance defaults and provide better tools for keeping privacy. However, they lamented that the total extent of privacy problem remained unknown and there was little knowledge about the
prevalence of wrong privacy settings or the difficulties customers face when handling their privacy [10].

3. Privacy Setting in Facebook

As per result, 83% users have a real name on social networking sites whereas 5% users have partial name and 2.35% users have fake name and the majority (68.82%) of the users were not sharing personal details on social sites. The 23% of users were not satisfied with privacy setting and 77% of users were satisfied with the default privacy setting, however it need to be improved [9]. The trust, security and photo sharing are the drawbacks, people with fake id use the photographs and morphing technique is used for various issues. Spreading viruses, software threats etc. Individuals make fake profiles for social designing, online impersonation to slander a user, to defame a person, advertising and campaigning for an individual or a group of individuals.

4. Crimes in Facebook

Scams

Criminals had been making use of the social networking sites like the Facebook which involve international scams. They are especially effective in attracting people through highly tempting images or web pages to click on a hyperlink that could be liked almost by all people. It consists of an innocent-looking notification which you’ve received an unfastened prize like a gift card, or any astrology predictions about you is awaiting inside etc. After that to declare the prize, scammers request you to provide some information inclusive of a credit card quantity or Social Security number etc. This description may additionally make it appear like scams are clean to identify, however, even the most savvy social media person needs to be on the lookout for illegitimate requests for statistics.

Cyber Bullying

Cyber bullying is a common occurrence among teenagers on Facebook and one that may bring about serious criminal charges if it is going a long way enough. Cyber bullying on Facebook has contributed to the deaths of numerous young adults who either dedicated suicide or had been killed with the aid of a peer. Cyber bullying that includes hacking or password and identity robbery can be punishable underneath state and federal regulation. When adults have interaction on this kind of online behavior, then it is called as cyber-harassment or cyberstalking.

Stalking

The time period “stalking” is thrown round loads on Facebook, and the main thing is it regularly supposed as a shaggy dog story for regularly looking at someone’s profile. However, the actual act of cyber stalking is a common crime at the social networking web page and might bring about a serious offense.
Cyber stalking commonly involves harassing someone with messages, written threats, and other continual online conduct that endangers a person’s safety. Although cyber stalking can also look like not anything more than demanding conduct, but should have a valid purpose for problem in many cases and can even lead to in-individual stalking or endangerment if now not treated critically.

**Robbery**

It doesn’t take a whole lot for a thief to discover where you stay, pass to high school, paintings, or grasp out in case you make that records effortlessly available on Facebook. If you utilize Facebook’s registration or Google Maps trademark, then you can be in a stack of bother if a thief is paying interest. This individual isn't continually an entire outsider, it is possible that; they will be an old colleague or a man else you'd not the slightest bit foresee to come let you by the announcement.

**Identity Theft**

With the large amount of private records swarming round Facebook nowadays, it has emerged as pretty clean for criminals to use or borrow users’ identities. Hackers often spoil into users’ emails and make fake Facebook money owed. Mainly your name, pictures and your travel information will be copied and trapped for illegal activities. From there they could get admission to private and financial institution facts and purpose havoc for your sense of security. Protect yourself from identifying theft on Facebook by means of maintaining your profile very comfortable and unfastened of personal records that a crook would really like to have.

**Defamation**

An individual commits the crime of defamation when they talk a fake announcement to a third birthday party that paints another individual or entity in a bad mild. Facebook makes communicating defamatory statements frighteningly easy, and the exposure Facebook offers make it much more likely that groups or people can be harmed through the defamatory statement, and thus more likely to pursue legal remedies. Be cautious what you say on Facebook; you may be committing against the law without even knowing it.

**Harassment**

By using Facebook harassment occurs often. From sexual harassment to assault, threats there were a sizeable boom within the variety of harassment cases going on Facebook. It’s not unusual for sex offenders and sexual predators prey on unsuspecting sufferers on Facebook and even pose as a youngster or college scholar. Harassing messages, irrelevant comments, and different persistent behaviors must be stated to Facebook and your local police station.

5. **Methodology**

Data mining offers with the discovery of unpredicted patterns and hidden new ideas which can be in tremendous databases It serves as an automatic instrument
that uses multiple evolved computational strategies, together with synthetic intelligence (the usage of computer systems to perform logical services), to discover wholly and signify big knowledge units involving one or more information sources, opting for colossal, recognizable patterns, traits, and relationships now not easily detected via normal analytical procedures on my own. This knowledge then may support with quite a lot of functions, such because the prediction of future activities or behaviors. The development of new intelligent tools for automated data mining has led to design and construct the successful systems that show early promise in their ability to scale up to the handling of voluminous data sets. We are using clustering and k-means creates $k$ groups from a set of objects so that the members of a group are more similar. It’s a popular cluster analysis technique for exploring a data set.

**Statistical Survey of Facebook Users**

A total 153 (90%) users out of 170 were to be found Facebook user and 10% users were not having an account on Facebook. Means Facebook is quite popular among people, the below figure represents the users percentage.

![Fig. 2: Facebook Users](image)

The bar chart represents the percentage of users who utilize the Facebook for reading news feeds, chatting, browsing friend’s profile and media sharing and blog, browsing profiles etc.

![Fig. 3: Different Statistical Purposes Using Facebook](image)

The status update is the common happenings, the life event celebration and inviting of an event are extra privileges of using Facebook.
**Data Collection**

The data set is the collection of field in the data from web pages on the internet. The data set which consists of the text from web pages and the pictures, videos or sound format will be ignored.

*Preprocessing*

A data preprocessing is a process that consists of data cleaning, data integration and data transformation which is usually processed by a computer Program. It intends to reduce a few noises, incomplete and inconsistent statistics. The results from preprocessing step may be later intending by way of a statistics mining set of rules.

6. **Clustering**

The clustering techniques are used to fetch the information about hackers. Data clustering is a process of collecting similar data into groups. Clustering is used to create a set (cluster) of the facts in order that it may without problems discover the essential records. Clustering is a class of comparable gadgets into several exclusive businesses; it is normally applied in the evaluation of statistical information which may be utilized in various fields, as an instance, system getting to know, records mining, sample popularity, photograph evaluation and Bioinformatics.

A clustering algorithm walls facts set into several corporations such that the similarity within a collection is greater among clustering it can be considered the most important unsupervised gaining knowledge of the method; so, as each different problem of this kind, it deals with locating a shape in a set of unlabeled information. There are such a lot of strategies used in clusters, on this paper best K-means set of rules is used.

\[ J(\gamma) = \sum_{i=1}^{C} \sum_{j=1}^{\gamma_i} (||x_i - v_j||)^2 \]

Whereas,

\[ ||x_i - v_j|| \text{ is the Euclidean distance between } x_i \text{ and } v_j. \]

Ci is the number of data points in \( i^{th} \) cluster.

C is the number of cluster centers.

There are two types of machine language

1. Supervised Learning and
2. Unsupervised Learning.

This machine learning model model can make it easier for the Facebook to manage the huge number of profiles existing in the database. The machine learning algorithms that have been proposed to be used in this model are Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Networks (ANN) and Naïve Bayes (NB).
Sentiment Analysis Using Naïve Bayes

There are many types of supervised algorithms available, one of the most popular ones is the Naïve Bayes model which is often a good starting point for developers since it’s fairly easy to understand the underlying probabilistic model and easy to execute. Every feature of the data being classified is independent of all other features given the class. Naïve Bayes is a statistical classifier. It is used to predict class membership probabilities. Here a membership probability means the probability that a given tuple belongs to a particular class. Naïve Bayes classification is based on independence assumption. This classifier assumes that the presence or absence of particular feature is independent of the presence or absence of any other future. It’s easy to implement, stable classification results, high efficiency. However, this algorithm generally is assumed that every attribute is independent of each other.

\[
P(A/B) = \frac{P(B/A)P(A)}{P(B)}
\]

Where,

- \(P(A/B)\) is the posterior probability of the hypothesis.
- \(P(A)\) is the prior probability of \(A\).
- \(P(B)\) is constant or evidence.
- \(P(B/A)\) is the conditional probability of \(B\) given \(A\). It is also called the likelihood.

In a simple way it is represented as,

\[
\text{posterior} = \frac{(\text{prior}) \times (\text{likelihood})}{(\text{evidence})}
\]

K-Means Clustering Algorithm

The k-means algorithm is a popular unsupervised algorithm that makes no assumptions about the data’s meaning it uses random seeds and an iterative process that eventually converges. This unsupervised clustering algorithm uses a distance metric with the goal of minimizing the Euclidean distance from the data points to a centroid, reassuring and reassigning each data point to the centroid on all the iterations.

K-means algorithm mainly used to partition the clusters based on their means. Initially a number of objects are grouped and specified as \(K\) clusters. The algorithm clusters observations into \(K\) groups, where \(K\) is provided as an input parameter. It then assigns each observation to clusters based upon the observation proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. In this paper the use of the K-means algorithm is the process of getting a structured data from an unstructured data.
The working of an algorithm is explained as follows:

1. K: pre-determined number of clusters
2. Algorithm (Step 0: determines value of k)
   
   - Step 1: Randomly generate k random points as initial cluster centers
   - Step 2: Assign each point to the nearest cluster center
   - Step 3: Re-compute the new cluster centers

Repetition step: Repeat steps three and four till some convergence criterion is met (normally that the assignment of factors to clusters becomes stable).

The use of K-means clustering algorithm in this paper is to get the structured data from unstructured data from the database. The prediction of this retrieving information from the database is clear and understands to the user of the system. K-means can be used to pre-cluster a massive dataset, followed by a more expensive cluster analysis of the sub-clusters. K-means can also be used to rapidly “play” with K and explore whether there are overlooked patterns or relationships in the dataset.

The k-means algorithm is a simple yet effective approach to clustering. K points are (usually) randomly chosen as cluster centers, or centroids, and all dataset instances are plotted and added to the closest cluster. After all instances have been added to clusters, the centroids, representing the mean of the instances of each cluster are re-calculated, with these re-calculated centroids becoming the new centers of their respective clusters.

Cluster membership is then reset, and the new centroids are used to re-plot and all instances and add them to their closest centroid cluster. This iterative process continues until convergence centroid locations remain constant between successive clustering iterations is reached.

7. Conclusion

Crime data is a sensitive domain where effective data mining techniques play a vital role in crime analysis. In this paper we classify the cluster techniques used to analyze the crime data from database. This technique is faster to get the data through the web; the effective web mining is to get the unstructured data to structured data. The classifications of crime type are violent, arson, fraud, etc. and clustering of crime using k-means to retrieve the data. We believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Many future directions can be explored in this still young field. For example, more visual and intuitive criminal and intelligence, investigation techniques can be developed for crime pattern and network visualization.
References


