Web Phishing Detection using Machine Learning

Balajee A V, T S Aswin, S Balaji, Abisheik R

Abstract: Phishing is one kind of cyber-attack , it is a most dangerous and common attack to retrieve personal information, account details, credit card credentials, organizational details or password of a client to conduct transactions. Phishing websites seem to like the relevant ones and it is difficult to differentiate among those websites. It is one of the most threatening that every individuals and organization faced. URLs are known as web sites are by which users locate information on the internet. The review creates warning of phishing attacks, detection of phishing attacks and motivate the practice of phishing prevention among the readers. With the huge number of phishing emails or messages received now days, companies or individuals are not able to find all of them. Vicious Web sites significantly encourage Internet criminal activity and inhibit the growth of Web services. As a result, there has been a tremendous push to build a comprehensive solution to prevent users from accessing such websites. We suggest a literacy-based strategy to categorize Web sites into three categories: benign, spam, and malicious. Our technology merely examines the Uniform Resource Locator (URL) itself, not the content of Web pages. As a result, it removes run-time stillness and the risk of drug users being exposed to cyber surfer-based vulnerabilities. When compared to a blacklisting service, our approach performs better on generality and content since it uses learning techniques.

***Keywords*: Security; Web Services; URL; Vulnerabilities**

I. INTRODUCTION

Due to rapidly growing technology internet is now an integral part of our daily life. Most of activities in our daily life are depended on the use of the internet. Social networking sites have widely increased over the last few years. Due to the regular use of the internet, the users are under frequent and harmful threats; one of them is ‘Phishing’. Phishing can be defined as impersonation of a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most wide spread cybercrime used today. There are countless domains where phishing attack may happen in online payment sector, webmail, and finances, file hosting or cloud storage networks and many others. The webmail and online payment sector has been attacked by phishing more than in any other industry. Phishing can be done through email phishing scams and spear phishing, which a user should be aware of the consequences and should not give their all-hearted trust on common security application. Machine Learning is one of the most efficient techniques to detect phishing as it removes drawback of various existing approaches. The objectives, which is the most vital thing in proposed project, is to verity the validity of the website by capturing blacklisted URLs. To notify the user on blacklisted website through pop-up while they are trying to access the URL and a platform for an individual too check and validate the integrity of any URL they want to access. Cybercriminals, rovers, or non-malicious (fair-limited) bushwhackers and data theft are all capable of carrying out attacks. The goal is to get to the system or the content it stores or to retrieve specific data from various methods. Bushwhackers seek to obtain a lot of information and or plutocrat by contacting a wide range of target druggies. As according Kaspersky's analysis, the estimated price of an attacker by 2019 is between $ 108K and $ 1.4 billion (depending on the scale of the attack). Furthermore, the plutocrat spent roughly $ 124 billion on worldwide security products and services. Phishing is the most common sort of cybersecurity assault, and bushwhackers are no exception. Phishing assaults are typically simple since most victims are unaware of the complexities of web operations, computer networks, and related technology, making them ideal prey for being misled or caricatured. It's far easier to phish unwitting drug addicts using fake websites and entice them to click on the websites in exchange for a prize or offer than it is to target the information security system. A vicious site is created to have a similar look and feel, and it looks to be authentic in appearance because it uses the association's ensigns and other copyrighted content. As a result of many drug addicts unknowingly accessing the phishing website URLs, the person and the organization involved suffering significant financial and reputational losses. The most common and dangerous of these types of cyberattacks was phishing. Cybercriminals typically use a dispatch or other social networking communication channels in this type of attack. Bushwhackers approach the drug addicts by claiming that the money was transmitted through a reputable site, other than banking, an e-commerce platform, and something related. As a result, someone attempts and try crucial information from them. Bushwhackers also use this information to puncture their victims' accounts. As a result, it results in financial loss and irreversible harm.

II. LITERATURE SURVEY

The current circumstance is that the population's maturity has been wisecracked, causing them to unknowingly give their private information to hackers. In emerging technology, industry, which deeply influence today’s security problems, has given a headache to many employers and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. In these new times there are many security systems being enabled to ensure security is given the outmost priority and prevention to be taken from being hacked by those who are involved in cyber-offenses and essential prevention is taken as high importance in organization to ensure network security is not being compromised. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, etc. Customer face numerous security threats like cybercrime.

Many cybercrime is being casually executed for example spam, fraud, identity theft cyber terrorisms and phishing. Among this phishing is known as the most common cybercrime today. Phishing has become one amongst the top three most current methods of law breaking in line with recent reports, and both frequency of events and user weakness has increased in recent years, more combination of all these methods result in greater danger of economic damage and issues.

Phishing is a social engineering attack that targets and exploiting the weakness found in the system at the user's end. This paper proposes the Agile Unified Process (AUP) to detect duplicate websites that can potentially collect sensitive information about the user. The system checks the blacklisted sites in dataset and learns the patterns followed by the phishing websites and applies it to further given inputs. The system sends a pop-up and an e-mail notification to the user, if the user clicks on a phishing link and redirects to the site if it is a safe website. This system does not support real time detection of phishing sites; user has to supply the website link to the system developed with Microsoft Visual Studio 2010 Ultimate and MySQL stocks up data and to implement database in this system.

The recent approaches to prevent the attacks like heuristics approach, blacklist approach, fuzzy rule-based approach, machine learning approach etc. and finally filtering all detection techniques based on accuracy and performance proposed a framework to detect and prevent phishing attacks. A combination of supervised and unsupervised machine learning techniques is used to detect malicious attacks.

III. METHODOLOGY

Modules:

- Data Collection
- Data Pre-Processing
- Feature Extraction
- Deployment Model

A. Data Collection

The data for this project is a collection of records. This stage includes choosing a sample of all available information on which to work. Data, especially as the huge quantity of data whereby the target output has been established, is the starting point for machine learning challenges.

B. Data Pre-Processing

Organize the data we've chosen by formatting, cleaning, and sampling it. Three common data pre-processing steps are:

- **Formatting:** That information we've chosen will not be in an easy-to-work-with format. The data could have been in a relational database which we'd like to export to a flat file, or it could have been in a unique file format that you'd like to export to a relational database or a text description.
- **Cleaning:** Clearing information includes eliminating and replacing data that isn't present. There could be a situation when data is missing or imperfect, and we don't have all of the information we need to solve the problem. It is indeed likely that all these circumstances have to be removed. Moreover, a few of the characteristics might be sensitive data, which must be cleared or completely removed from the information.
- **Sampling:** There could be a lot more well-chosen data accessible than we need. Increased method execution durations and larger computational and storage requirements result from more information. We can choose a shorter sample size of the data sample before reviewing the complete dataset, which will allow us both to explore and develop ideas much faster.

C. Feature Extraction

The following stage is feature extraction, and that's an attribute extension that allows us to create more columns from URLs. Finally, we use a classifier algorithm to train our models. They take advantage of the obtained classified dataset. The remainder of our classified data would be used to validate the models. ML algorithms have been used to identify pre-processed data.

D. Deployment Model

The deployment of a model is a key step in its development. It helps us to figure out which model perfectly describes the data but also how this might perform as in years ahead. We build the trained model and publish it on Static Web Page using IBM Cloud facility. Here, We use Python Flask which is an API of Python that allows us to build up web-applications. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application. Database integration is also very simple and easy using Flask.

IV. CONCLUSION

This survey presented various algorithms and approaches to detect phishing websites by several researchers in Machine Learning. On reviewing the papers, we came to a conclusion that most of the work done by using familiar machine learning algorithms like Naïve Bayesian, SVM, Decision Tree and Random Forest. Some authors proposed a new system like PhishScore and PhishChecker for detection. The combinations of features with regards to accuracy, precision, recall etc. were used. As phishing websites increases day by day, some features may be included or replaced with new ones to detect them.

V. REFERENCES

1. Phishing Website Classification and Detection Using Machine Learning by Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, B.S. Bindhumadhava was published in the year 2020.
2. Machine Learning-Based Phishing Attack Detection by Sohrab Hossain, Dhiman Sarma, Rana Joythi Chakma published in the year 2020.
3. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013