Detecting Phishing Websites Using Machine Learning

Adwait Changan[1], Vaibhav Mahalle[2], Prafull Patil[3], Prabuddha Salve[4]

Department of Information Technology, Sinhgad College of Engineering, Pune, India. [1, 2, 3, 4]

Abstract – ***Today's generation is increasingly reliant on web technology for a variety of tasks such as banking, communication, and so on. As a result, users can encounter multiple security threats, with phishing being one of the most serious and prominent attacks. Phishing attacks attempt to steal sensitive information from users by impersonating a legitimate entity. The attacker employs a phishing attack to obtain the victims' credentials, such as their bank account number, passwords, or other sensitive information by impersonating a genuine website, and the victim is unaware of the phishing website. So, in this paper, a system is proposed to detect phishing sites using machine learning in real-time by utilizing a classifier that is trained on an exhaustive dataset with enriched features.***

Key Words: Phishing, Cyber Security, Random Forest, Malicious URL detection, Machine learning in Cyber security

1. INTRODUCTION

Today, the majority of individual and organizational communication and interaction takes place via the internet, and this trend is expected to continue and grow. People are currently heavily reliant on web technology, and most are unaware of the cyber threat due to a lack of technical knowledge. Numerous organizations and businesses have already been confronted with the threat and problem of cyber-attacks. Among different types of cyber-attacks present, Phishing is a significant cyber-attack that can threatens online users' identities. Phishing attacks typically involve an attacker who will act as a credible resource in order to steal sensitive data from victims. Victims of successful attacks visit phishing websites without recognising it. Once on the website, users can provide private information such as passwords or banking-related sensitive information, and they are also at risk of downloading malware that the attacker has placed on the site.

This paper's main objective is to present a technique for identifying phishing websites. There are approximately five ways to this problem, including the Blacklisting strategy, the Rule-based or Heuristics-based approach, the Content-based approach, and the Machine Learning approach, which is then improved by the Hybrid approach [9]. The suggested method uses a model that is trained on the website's contents and URL-based features to determine if the site is a legitimate website or a phishing website.

2. LITERATURE SURVEY

1) Mehek Thakera, Mihir Parikhb, Preetika Shettyc Vinit Neogid, Shree Jaswale(2018)

This paper proposes a system that will detect old and newly generated phishing URLs using Data Mining. A cloud-based classifier is developed which takes features of URL as an input. The model will be deployed using the chrome extension. The model will be trained with an exhaustive dataset and uses URL-based and Domain-based Features to ensure maximum accuracy [1].

2) Srushti Patil, Sudhir Dhage (2019)

A comparative study of the important anti-phishing tools was completed and their limitations were pointed out. This paper analyzed the URL-based features used in the past and improved their definitions as per the current scenario [9]. There is a full implementation of the anti-phishing tool shown in the paper. Also, the accuracy and observation of the developed tool are given [9].

3) Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri

In this paper, a real-time anti-phishing system, using seven different classification algorithms and natural language processing (NLP) based features, is proposed [4]. The system has the following distinguishing properties: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers [4]. New dataset is constructed for measuring the performance of the system and the experimental results are tested on it. According to the experimental results from the implemented classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs [4].

4) Dharmaraj R. Patil, and Jayantrao B. Patil (2018)

This paper gives a methodology to detect malicious URLS and the type of attacks based on multi-class classification. In this work, they proposed 42 new features of spam, phishing and malware URLS. These features were not considered in the earlier studies for phishing URLs detection and attack types identification [3]. The training data for the developed tool was created with help of 26041 benign and 23894 malicious URLs containing 11297 malwares, 8976 phishing and 3621 spam URLS. Experiments are performed on the created dataset using machine learning classifiers [3].

5) Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi (2019)

The purpose of this study is to provide a thorough overview and a structural knowledge of machine learning-based malicious URL detection methods. They outline the formal definition of malicious URL detection as a machine learning challenge, classify, and evaluate the contributions of literature works that address various aspects of this issue. Paper offers numerous URL- and content-based capabilities that can be utilized to improve model training [2].

3. PROPOSED SYSTEM

The suggested system will have a client-server design. On the client side, a chrome extension will be used to send the Uniform Resource Locator (URL) and Web page source attribute to the server that the user is presently visiting. A cloud-based model for phishing site detection will be constructed on the server side which is trained using random forest algorithm. [1]

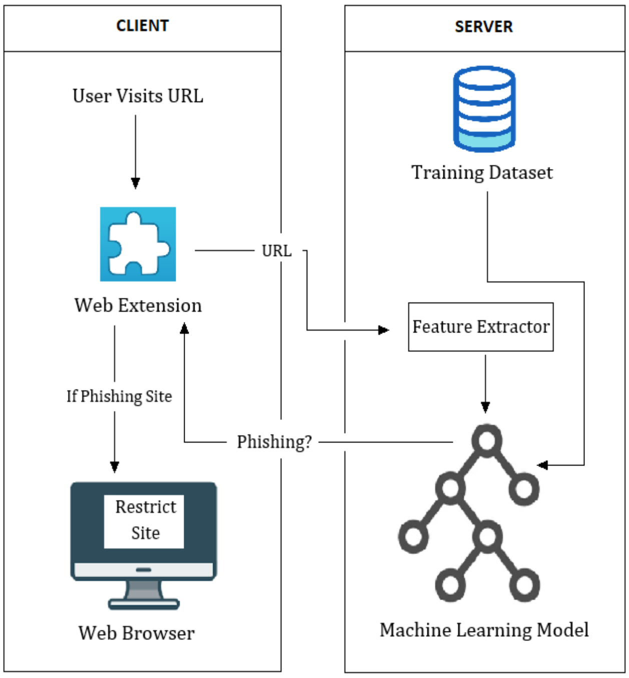

Fig -1: Proposed System Architecture

3.1 Training Dataset:

The dataset needed to train the model should be large, with two distinct classes: legitimate and phishing. Furthermore, the dataset should include a balanced mix of legitimate and phishing sites. Phish Tank will mostly be the source of the phishing URLs. [5] For legitimate pages, Alexa, Statista, and Similarweb can be used to get pages with high traffic and better ranking as these pages will have a very low possibility to be phishing web pages. It is because malicious websites will have less traffic and lower ranking on search engines due to their limited life span. As a result, a dataset with 40000 URLs can be formed.

3.2 Feature Selection:

Feature selection is an important process because it has a large impact on model accuracy in the real world. The process of selecting the best set of features for model training is known as feature selection. In proposed system features used are URL based and Content based features.

3.2.1 : URL based features:

These are features that are obtained from the URL that which user is currently visiting.

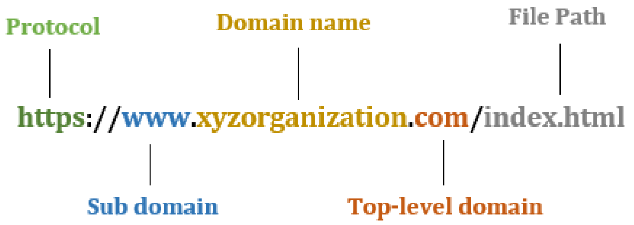

Fig -2: URL Components

1)Protocol check: To check if protocal used is "https".

2)Word count: After parsing URL through special characters words are counted.

3)Average word length : Average length of words obtained after parsing.

4)Character count: Total number of characters present in URL.

5)Digit count: Total number of digits present in URL.

6)Special characters count: Total number of special characters present in URL.

7)Keyword count: Keywords like login, gift, secure, etc. count.

8)Brand name count: Keywords like facebook, gmail, etc. count.

9)Look alike keywords count: Keywords like logiin, seccure, etc.

10)Look alike brand name count: Keywords like faceb00k, instagrom.

11)Random words : Words which are not keywords and brand name.

12)Length of file path : The length calculated for path.

13)Top level domain check: Verify if Top level domain is most widely used domian like : com, edu, org, etc.

14)Occurance of Subdomian : Occurance of Subdomian are usually more in malacious site.

3.2.2 : Content based features:

These features can be derived from the source code of the page which user wants to access. Feature are:

1) Word count: Total number of text words present on web page.

2) Average word length: Average length of text words present on web page.

3) Links Count: Total number of links present in web page.

4) Iframe tag count: Total number of Iframe tags present in web page.

5) Embed tag count: Total number of embed tags present in web page.

6) Common Phishing word count: Words like pay, bonus, free, access, log, etc. count.

3.3 Data Pre-processing:

Raw data is transformed into usable formats during data preprocessing. Decomposers can be used on URLs to split it and extract the necessary parts in order to obtain attributes such as brand name, portocol, etc. The most well-known and frequently used brand names and keywords are gathered and checked for their presence in URLs. To extract URL-based features, the URL visited by the user is split into words using special characters. After that, brand name and keyword checks are performed on the obtained words. If a splited word is not found in both dictionaries, it is sent to a word decomposer, which can split two adjacent words in a string into two separate words. Word decomposer firstly creates substrings of the input string passed. Then a dictonary check is made on the obtained sub strings to know the words present. If it is unable to separate, then the word's similarity to available brand names and keywords is examined. Still, if there is no similarity,then it is treated as a random word. Depending on the status of a word under review appriopriate features are incremented. For content based feature web crawling can be done to get the value for the features.

3.4 Classifiers:

With the data set created, multiple classifiers were trained. The Random Forest algorithm was the most accurate. Random forests are an ensemble learning method for Classification, Regression, and other tasks that operate by constructing a multitude of decision trees at training time. The trained model will be deployed using cloud services in the proposed system.

Table -1: Accuracy of Classifiers trained

Classifier	Accuracy
Naive Bayes	91.9832
Support Vector Machine	94.5901
Neural Network	96.3394
Random Forest	**97.3659**
K-Nearest Neighbor	97.1384

4. CONCLUSIONS

Phishing website is one of the challenging security problems faced recently due to the rise of web pages worldwide and the detection of these websites as legitimate and phishing is one of the challenging aspects. The Detection and Prevention of Phishing Websites system offers security for the user who can easily fall into a trap due to a lack of awareness or technical knowledge. So, a system is developed with enriched Features and a Random Forest classifier is used to achieve better accuracy. Trained model will be deployed using cloud services. On client side, browser extension which is a small software module for customizing a web browser will be used [11]. The extension will send the URL that the user is attempting to access to a cloud-based feature extractor, which will then supply the extracted feature to the model for detection. Furthermore, the classification result will be updated to an extension to restrict user access to the page if it is a phishing site.

The proposed system will have advantages:

1. Real-time Execution.

2. Huge Size of Phishing and Legitimate Data.

3. Detection of new Websites

4. Independence from Third-Party Services

5. Use of Enriched Features.

REFERENCES

[1] Mehek Thakera, Mihir Parikhb, Preetika Shettyc, Vinit Neogid, Shree Jaswale. "Detecting phishing websites using Data Mining" 2018.

[2] Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi. "Malicious URL detection using Machine Leaming: A Survey" 2019.

[3] Dharmaraj R. Patil, and Jayantrao B. Patil. "Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification" July 2018, Volume 10, Number 2.

[4] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri "Machine learning based phishing detection from URLs" 25 July 2018.

[5] Ebubekir Buber, Banu Diri, and Ozgur Koray Sahingoz. "NLP based phishing attack detection from URLs." November 2007. [6]

[6]Varsharani Hawanna, V. Y. Kulkarni, R. A. Rane. "A novel algorithm to detect phishing URLs" 2016.

[7] Yu Zhou, Yongzheng Zhang, Jun Xiaon,Yipeng Wang, Weiyao Lin "Visual similarity based anti-phishing with the combination of local and global features" 2014.

[8] M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani. "Hybrid model to detect phishing sites using Supervised Learning Algorithms" 2016. [9] Srushti Patil, Sudhir Dhage. "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework" 2019.

[10] Microsoft Contributors. Phishing[online] Available:

https://docs.microsoft.com/en-us/windows/security/threat-protection/intelligence/phishing

[11] Wikipedia Contributors. Browser Extension[online] Available:

https://en.wikipedia.org/wiki/Browser_extension