International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Detecting Phishing Websites using Data Mining

Ashna Antony M[1]

[1]*DDMCA Student, Department of MCA, Sree Narayana Guru Institute of Science and Technology, Kerala, India*

Abstract – *Phishing is one of the serious cyber threats now, where the victims' credentials are acquired by an illegitimate website. Phishing sites which expects to take the victims confidential data by distracting them to surf a fake website page that resembles a honest to goodness one is another type of criminal develop through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. In general, Phishing is a type of extensive fraud that appears when a malicious website serve like a real one. Phishing site detection is truly an uncertain and element issue along with numerous components and criteria that are not stable. This paper introduce a system which will detect and block old as well as newly generated phishing URLs that have completely no past behaviours to judge upon, using Data Mining. A cloud-based classification model will be designed for the same wherein various extracted attributes through the URL will be used as input data. The model will be trained with an exhaustive dataset so as to provide maximum accuracy. Here we applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we block the website in our system.*

***Key Words*: Phishing Detection, Data Mining, Machine Learning**

1. INTRODUCTION

Phishing is a technique used by hackers or attackers to trick the users into inserting their sensitive credentials such as usernames, passwords, and credit cards details into a nongenuine entity such as a website. In this type of attack, unauthentic entities disguise themselves as genuine and trustworthy entities. The users are thus, tricked by the look and feel of the fake website which is almost identical to the legitimate one. Generally, attackers handle banking and payment sites, social media sites and E-Commerce sites to fake potential victims. In 2016, a variety of variations in spam flows with an increase in the number of malicious mass E-mails involve links to phishing sites was observed [6]. Until recently, PhishTank has verified and confirmed 2,259,845 websites as phishing sites [10]. Hence, phishing has now become the leading delivery vehicle for ransom ware and other malware [6]. So, there is a heavy use for expanding a very effective anti-phishing solution.

Phishing is a type of extensive fraud that develops when a malicious website act like a real one keeping in mind that the end goal to gain touchy data, for example, passwords, account points of interest, or MasterCard numbers. Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing generates utilization of spoof messages that are built to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations. The misleading sites are intended to emulate the look of a genuine organization site page[11].We proposing a system that detects the Malicious URLs websites hosting phishing, spam etc. by using Machine Learning. The system should be useful in preventing online frauds leading to leakage of important and private user data. Detection and Prevention of Phishing Websites using Machine Learning Approach is a mechanism that is proposed in order to ensure high security[7]. In this mechanism we deals with the URLs and the URL check with machine learning technique and predict is it is phishing website or not. Here we create a web browser for browsing. Each time we browse a site the corresponding URL of site will be checked with machine learning technique. If the predicted result will be negative then that site will be blocked. Then we can't access the site from that system. The main advantage of our system is that we can't access the blocked site not only from our browser (we created) but also from other browsers too. Here simple demo browser program is developed using Java and send the URL value to python which is entered by the user.

2. RELATED WORKS

There have been different experiment made to discover a robust and dynamic solution to this problem which can authenticate whether a site is phishing or not placed on its current attributes rather than previously defined rules.

Ankit Kumar Jain et al. [3] introduced a complete analysis of all the Phishing attacks familiar, along with their appearing consequences. Moreover, it also contribute a very helpful vision over the various machine learning based methods for phishing detection with the aid of a comparative study. This creates various viewpoints in terms of discovering further efficient solutions with help of machine learning in near future. A detection technique for phishing websites is suggested by Abdulghani Ali Ahmed et al. [1] which scan Uniform Resources Locators (URLs) of suspected web pages as per five extracted features. Phishtank and Yahoo directory datasets are used to determine the accuracy of the results given by proposed solution. The final report thus establishes that the detection mechanism can reveal various types of phishing attacks without fail. However, there are still chances

of accepting false alarms. In dynamically heuristic anti-fraudulence system [12], Shree Jaswal et al. proposed a system wherein first the demonstrate records of phishing websites are assigned to verify if the given website is one of them. If not, then four heuristics are extracted from URL and are used to compute performance measure. Additionally, when certain conditions are met, images or logos are distinguished to assure more efficiency. Alexa Ranking and no of URL- based features, for detecting phishing URLs is proposed by Varsharani Ramdas Hawanna et al. [2]. It shows an alert message if the URL is classified as phishing; otherwise it displays a safe message. This algorithm added to the performance when dealing with known/old phishing URLs. Priyanka Singh et al. [5] has performed two algorithms named Backpropagation network with Support Vector Machine (SVM) and Adaline network to increase the detection rate and classification using datasets of Phishtank and Alexa. Training time, testing time, mean square error and prediction accuracy are the parameters used to assess the performance of both algorithms. Adaline network gives 99.14% accuracy. Training time used by the Adaline network is very less when distinguished with the Backpropagation network with SVM. A Hybrid Model which uses 30 features to clarify the phishing websites problem is granted by Sohail Asghar et al. [4]. A single model cannot efficiently detect the phishing websites, therefore to enlarge the accuracy, efficiency and performance rate, two or more models are joined together to form a more robust classifier. Firstly, the individual performance of a classifier is analyzed and then the best classifier in terms of high accuracy and less error rate is calculated. The best classifier model is combined with other classifiers one by one and finally, a better hybrid classification model is accomplished.

3. METHODOLOGY

Detection and Prevention of Phishing Websites using Machine Learning Approach is a mechanism that is proposed in order to ensure high security. In the proposed system, a cloud-based model is used for phishing website detection. The model will be based on classification algorithm and will be trained using a training dataset. This model will be deployed in the cloud, which will directly communicate with the chrome extension. The detection of the phishing website will be based on URL and website attributes. This system will be an integration of all the functions carried out on the client and server side. On the server side, there will be a classifier model trained using the random forest algorithm; whereas on the client side, a chrome extension will be built and added to the chrome web browser. Here we deals with the URLS and the URL check with machine learning technique and predict whether it is phishing website or not. The system architecture is represented in Fig -1 .Here we create a web browser for browsing. Each time we browse a site, the corresponding URL of site will be checked with machine learning technique. A voice alert will be played based on the predicted result. If the predicted result will negative, then that site will be blocked. Then we can't access the site from that system. The main advantage of our system is that we can't access the blocked site not only from our browser (we created) but also from other browsers too.

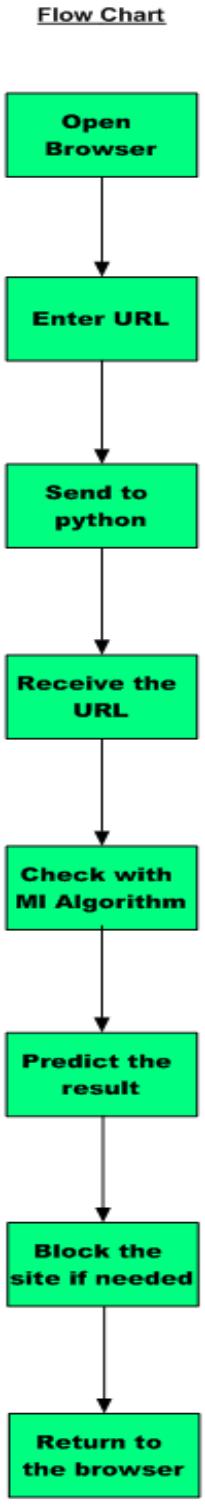

Fig -1: System Architecture

3.1 Modules

- **User**

In this module user enter a URL for browse a site, that URL be the input for our machine learning algorithm (RANDOM FOREST). Based on the predicted result of the algorithm a user can access the site. If the entered site is phishing site then the user can't access it.

- **System (Process modules)**

System Module involves five steps. They are:

1. Dataset collection
   Text dataset of Phishing site details.
2. Pre-processing
   In pre-processing step, separating data and label.
3. Feature Extraction
   In feature extraction stage, the text features are extracted. The system cannot understand the text data so we convert the text data into numerical values.
4. Training
   Here we train the machine learning Classifier. We use the random forest classifier to classify the entered site into phishing site or not.
5. Testing
   In testing phase the classifier predict a new URL of website is phishing site or not.

3.2 DATASET

For the proposed model, a publicly accessible dataset offered by UCI repository [9] will be used for training. It consist of 11055 records, out of which 4,898 are phishing websites while 6,157 are legitimate websites.

3.3 ALGORITHM

We are using random forest classification. A Random Forest[8] is an ensemble technique accomplished of performing both regression and classification tasks with the help of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to integrate multiple decision trees in deciding the final output rather than relying on individual decision trees.

4. CONCLUSION

Through this system, the goal is to implement the detection of the phishing websites using data mining. This work will be done by extracting the features of the website via URL when the user go through it. The collected features will act as test data for the model. In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm. We implemented classification algorithm and techniques to extract the phishing data sets criteria to analyze their legitimacy. Here, Random Forest Algorithm can be used to train the proposed model. Thus the system detect the phishing website and alert the user beforehand so as to prohibit the users from getting their credentials misused.

REFERENCES

[1] Ahmed, Abdulghani Ali, and Nurul Amirah Abdullah. "Real time detection of phishing websites." Information Technology, Electronics and Mobile Communication Conference (IEMCON), 7th Annual. IEEE, 2016.

[2] Hawanna, Varsharani Ramdas, V. Y. Kulkarni, and R. A. Rane. "A novel algorithm to detect phishing URLs." Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on. IEEE, 2016, pp. 548-552.

[3] Jain, Ankit Kumar, and B. B. Gupta. "Comparative analysis of features based machine learning approaches for phishing detection." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016, pp. 2125-2130.

[4] Tahir, M. Amaad Ul Haq, et al. "A Hybrid Model to Detect Phishing Sites Using Supervised Learning Algorithms." Computational Science and Computational Intelligence (CSCI), 2016 International Conference on. IEEE, 2016, pp. 1126-1133.

[5] Singh, Priyanka, Yogendra PS Maravi, and Sanjeev Sharma. "Phishing websites detection through supervised learning networks." Computing and Communications Technologies (ICCCT), 2015 International Conference on. IEEE, 2015, pp. 61-65.

[6] J. Crowe, 'Phishing by the Numbers: Must-Know Phishing Statistics 2016', 2016. [Online]. Available: https://blog.barkly.com/phishingstatistics-2016.

[7] Phishing Websites Detection Using Machine Learning Based Classification Techniques Mazharul Islam,Nihad Karim Chowdhury Department of Computer Science & Engineering University of Chittagong.

[8] Random Forest Regressor Explained In Depth [Online].Available :https://gdcoder.com/random-forest-regressor-explained-in-depth/amp/

[9] RM. Mohammad, 'Index of /ml/machine-learning-databases/00327', 2015. [Online]. Available: https://archive.ics.uci.edu/ml/machinelearning-databases/00327

[10] Phishtank, "Out of the Net, into the Tank." [Online]. Available: https://www.phishtank.com/

[11] Detection Of Phishing Websites Using Data Mining [Online].Available : https://www.ijert.org/detection-of-phishing-websites-using-data-mining-techniques

[12] Shree Jaswal, Siddhesh Shirke, et al. "Dynamically heuristic antifraudulence system.", Global Technology Initiatives, 2015 4th International Conference on. IJGTI, Volume 4, Issue 1, pp. 44-51.