International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Phishing Site Detection Using Machine Learning

T. Sharathkumar[1*], Pranith R. Shetty[2], D. Prakyath[3], A. V. Supriya[4]
[1,2,3]*B.E. Student, Department of Information Science and Engineering, Srinivas Institute of Technology, Mangaluru, India*
[4]*Assistant Professor, Department of Information Science and Engineering, Srinivas Institute of Technology, Mangaluru, India*
*Corresponding author: sharaththadegallu@gmail.com

***Abstract*: Phishing is one of the major threats in this internet era. Using Phishing, targets are contacted by email, telephone or text message. Sensitive data such as personal information, banking and credit card details and passwords can be stolen. This information is used to access important accounts and can result in identify theft and financial loss. Using Machine Learning techniques like feature extraction, SVM (Support Vector Machine) and Random forest algorithm over other traditional detecting methods, this problem of detecting a phishing website can be resolved. This system uses the user inputted URL and on submitting, extracts features of the URL, and classifies them as a phishing site or a non-phished site using Random Forest algorithm. Another module in the system is a page extension which determines whether the site opened by the user is phished or not and also the percentage of how much a site is phished or not phished. A dataset containing URL features is used to train the Random Forest algorithm to do so. Random Forest is the best algorithm in classification, which gives reliable results. An accuracy of 98.68% is obtained in using Random Forest algorithm.**

***Keywords*: Phishing detection system, Phisher, Random Forest algorithm, RDF, Keras, OpenCV.**

1. Introduction

These days, as many people are aware of using internet to perform various activities like online shopping, online bill payment, online mobile recharge and banking transaction. Due to the widespread use of these technologies the client faces various security threats, such as cybercrime. Many cyber crimes are widely executed, such as spam, fraud, cyber terrorism and phishing. Among this phishing is new cybercrime and very popular nowadays.

Phishing is a fraudulent method designed to obtain users' sensitive information. Phishing costs Internet users billions of dollars per year. Phishing refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting internet users. The attacker use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords.

The above problem shows that in addition to user training, computer-based solutions are needed to prevent phishing attacks. Such a solution would enable the computer to identify malicious websites to prevent users from interacting with it.

A common method for identifying illegal phishing websites is to rely on their uniform resource locator (URL). A URL is a global address of a document in the World Wide Web, and it serves as the primary means to locate a document on the Internet. Even in cases where the content of websites is duplicated, the URLs could still be used to distinguish real sites from imposters. Because of the adaptability of phishers tactics, detecting and identifying phishing websites is really a complex and dynamic problem. In this paper, a novel approach is proposed for detecting phishing websites.

2. Literature Survey

Related works which were referred in developing the system are given. One of the proposed methods is fast and reliable because it does not rely on third parties, but only extracts functionality from the URL and source code. Using this method, they have reached 99.09% of the overall detection accuracy of phishing websites. This paper [1] has concluded that this approach has limitation as it can detect webpage written in HTML. A non-HTML webpage cannot detect by this approach.

In another method, extracted features are classified into three categories such as URL Obfuscation features, Third-Party-based features and Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services. Also, this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has a limitation of more execution time for the websites with more number of links [2].

The next method presented a hybrid model for classification of phishing website [3]. This, proposed model carried out in two phases. In stage 1, they separately perform classification techniques and select the best three models based on high accuracy and other performance criteria. While in stage 2, they further combined each individual model with best three models and makes hybrid model that gives better accuracy than

individual model. They achieved 97.75% accuracy on testing dataset. There is limitation of this model that it requires more time to build hybrid model.

Another method, which uses RDF (Resource Description Framework) and Random Forest Algorithm phishing, is identified in two phases. First phase explores the strength of RDF in identifying phishing web pages based on their metadata. Second stage, explores a machine learning technique that can help the system in taking a decision given a feature set as input. The entire system is viewed as three-step process. First, given a suspicious webpage and the system extracts the metadata from the source code and content of the page and then constructs an RDF structure for the page. Second, keywords are extracted from the given suspicious page and this keyword vector is fed to a search engine. Top "n" results are collected in the same order as appeared. RDFs are created for all the web pages that have come as top results in the search. Third, a comparison is made between the RDF (Resource Description Framework) of suspicious page and RDFs of the search results to come to a conclusion. If this previous stage fails without giving a conclusion, then employ the second stage, where a feature vector is given as an input to Random forest classifier to take a decision [4]. The drawback being that that accuracy obtained is just 91% and it is not integrated into web browsers.

3. Methodology

The architecture as shown below uses a database collection which includes dataset of phished and non-phished URL's. The user input is entered into the system and the features of the URL are extracted along with WHOIS features. The page rank is also checked and then sent to the Random Forest Algorithm for classification. Based on all the data extracted from the URL in the Feature Extraction stage, the site is classified as phished site or not and displayed to the user along with the percentage of how much a site is phished or not phished.

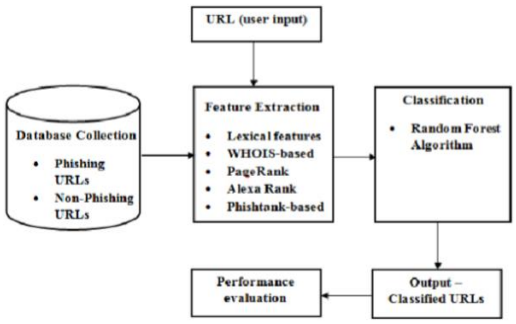

Fig. 1. System architecture of phishing site detection using machine learning

A. Phishing URL Feature Checking

Few selected features can be used to differentiate between legitimate and spoofed web pages.

- Feature of IP address like Network ID and Host ID is checked to verify if the IP address exists in the URL. For instance, a URL as "http://192.100.3.124//fake.html" indicates that someone is trying to steal some information from the user.
- Long URLs usually used by the phisher to hide the suspicious part. There is no exact length to indicate the phishing site; however, the normal length of URL does not exceed 54 characters.
- Phisher tend to add prefixes or suffixes separated by the mark (-) so that the user will trust the URLs as a legitimate web page URL.
- Some URLs of phishing web pages have an additional item in front of the real URL.
- This function checks the location of the symbol "//" in the URL.
- If the URL starts with "HTTP", it means that the symbol "//" should appear in the sixth position. However, if the URL uses "HTTPS", the symbol "//" should appear in the seventh position.
- Using the "@" symbol causes the browser to ignore everything before the "@" symbol, and the actual address is usually located after the "@" symbol.

B. Phishing Visual Content Checking

Keras is used for image pre-processing to and image is processed using OpenCV, to compare the similarity of the webpage corresponding to the given URL with the actual legitimate webpage. Based on it, the user will receive a percentage probability indicating a webpage is phished or not.

C. Implementation

To determine a classifier with the best performance for using URL based features we use Random Forests algorithm. Random Forest or Random Decision Forest is a holistic learning method for classification, and regression. This method operates by constructing a large number of decision trees during training and output the class predicted as the pattern of the class or the mean of a single tree. Random Forest Algorithm creates a forest of trees and the output is generated.

For URL features, in the training phase, a classifier is generated using URLs of phishing sites and legitimate sites collected in advance. These are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features. In the detection phase, the classifier determines whether a requested site is a phishing site.

When a page request occurs, the URL of the requested site is transmitted to the feature extractor, which extracts the feature values through the predefined URL-based features. Those feature values are inputted to the classifier. The classifier determines whether a new site is phishing site based on learned information. It then alerts the page-requesting user about the classification result.

Another feature of the system, which is the use of a page extension which can be used in browser. When a website is entered on the browser, the user can click on the page extension and scan the website. The extension then uses Keras for pre-processing the image and then OpenCV to process the image

and then returns whether the site is phished or not. It also displays how much a site is phished or not phished in terms of percentage.

4. Results

Fig. 2. Front page

Fig. 2 shows us the front page where the user's URL to be tested is entered. If the entered URL is in the correct format, then on clicking submit the system displays whether the entered URL is phished or not.

Fig. 3. Page showing URL entered is legitimate

Fig. 3 shows the result obtained when entered URL is detected as legitimate URL.

Fig. 4. Page showing URL is phishing site

The above image Fig. 4 shows the result obtained when entered URL is detected as phishing URL.

Fig. 5. Page showing URL is suspicious

The above image fig. 5 shows the result obtained when entered URL is detected as a suspicious URL.

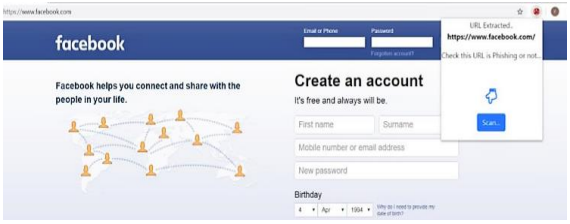

Fig. 6. Page showing URL extension

The above image fig. 6 shows the result obtained when the page extension is used to detected whether the opened site in a browser is phished or not.

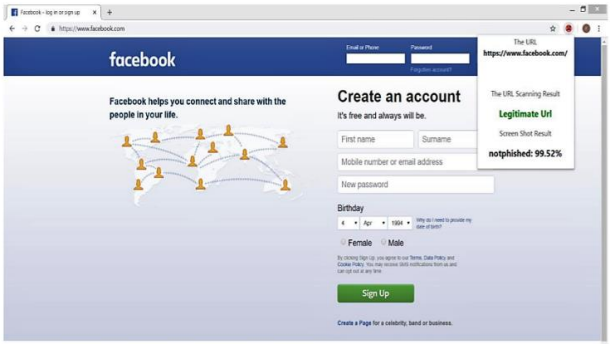

Fig. 7. Page showing URL extension detecting site as legitimate along with the percentage the site is not phished

The above image fig. 7 shows the result obtained when a legitimate site is opened and checked using the page extension. It also shows the percentage of accuracy.

5. Conclusion and Future Work

The phishing detection system has been successfully implemented. It uses has 3 features like detecting whether the URL entered by the user is phished or not and automatically detect whether the opened website is phished or not using a page extension and also using image processing to detect whether the website is phished or not phished. All aspects of the system are functional. The system has been tested with several test cases and has produced results as expected. This project has led to a better understanding of methods and techniques of predicting phishing website. This work has given a platform for improving cyber security. It has assisted in bringing critical information regarding threats to sensitive data to those that require it in a timely and effective manner. In this paper, the structure of URL is identified and features are extracted using Random Forest Algorithm techniques and system is trained to perform classification to detect phishing. The experimental results show the solution is effective to detect URL phishing and can be used as plug-in in browsers to filter the phishing sites.

The system has been designed and implemented with a view to provide detection of phishing websites and notify the users about such suspicious website. Future enhancements for the same may be performed in the following ways.

- Improve the accuracy of the existing prediction by means of either improving the existing algorithms or by replacing it with one better suited for it. This can reduce the chances of a wrong prediction and unnecessary remedial actions.
- The system can be installed at the state or central levels to provide a central monitoring and management system for large platform. This can lead to centralized monitoring and control.

- The system can be optimized to consume fewer resources and take lesser time for analysis of a website. The hardware can be upgraded to handle large volume of data and intensive analysis and therefore react accurately and quickly to the phishing websites.

References

[1] Zhang, Y, Hong, J I, & Cranor, L. F. (2007, May). Cantina: A content-based approach to detecting phishing web sites. In proceedings of the 16th international conference on World Wide Web (pp. 639-648). ACM.

[2] Dunlop, M., Groat, S & Shelly D. (2010, May). Goldphish: Using images for content-based phishing analysis. In Internet Monitoring and Protection (ICIMP), 2010 Fifth International Conference on (pp. 123-128). IEEE.

[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013.

[4] G. Liu, B. Qiu, L. Wenyin, "Automatic detection of phishing target from phishing webpage", Pattern Recognition (ICPR) 2010 20th International Conference on, pp. 4153-4156, Aug. 2010.

[5] P. Prakash, M. Kumar, R. R. Kompella, M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks", INFO COM'10: Proceedings of the 29th conference on Information communications. Piscataway NJ USA: IEEE Press, pp. 346-350, 2010.

[6] http://resources.infosecinstitute.com/category/enterprise/phishing/phishing-countermeasures/anti-phishingthe-importance-of-phishing-awareness-training/ (as on 19-06-2018)