IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10736

To Detecting Phishing Attacks Using Natural Language Processing and Machine Learning

Mr. D. Anand Joseph Daniel, M. E, (Ph.D.)[1], G. Reshma[2], C. Selvarani[3]

Assistant Professor, Computer Science and Engineering, Anand Institute of Higher Technology,

Kazhipattur, Chennai-603103[1]

Student, Computer Science and Engineering, Anand Institute of Higher Technology,

Kazhipattur, Chennai-603103[2,3]

Abstract: Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. Deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Random forest and Support vector machine algorithms are used to detect phishing websites. The propose a learning-based approach to classifying Web sites into 3 classes: Benign, Spam and Malicious. Benign are the safe websites with normal services. Spam is the Website performs the act of attempting to flood the user with advertising or sites such as fake surveys and online dating etc. Malware are the Website created by attackers to disrupt computer operation, gather sensitive information, or gain access to private computer systems. Thus, it eliminates the run-time latency and the possibility of exposing users to the browser-based vulnerabilities.

Keywords: phishing, website, security, vulnerabilities. cyber security, information

I. INTRODUCTION

While the Internet has brought unprecedented convenience to many people for managing their finances and investments, it also provides opportunities for conducting fraud on a massive scale with little cost to the fraudsters. Fraudsters can manipulate users instead of hardware/software systems, where barriers to technological compromise have increased significantly. Phishing is one of the most widely practiced Internet frauds. It focuses on the theft of sensitive personal information such as passwords and credit card details. Phishing attacks take two forms are the attempts to deceive victims to cause them to reveal their secrets by pretending to be trustworthy entities with a real need for such information and attempts to obtain secrets by planting malware onto victims' machines. The specific malware used in phishing attacks is subject of research by the virus and malware community and is not addressed in this thesis. Phishing attacks that proceed by deceiving users are the research focus of this thesis and the term 'phishing attack' will be used to refer to this type of attack.

II. RELATED WORKS

[1] Phishing websites, fraudulent sites that impersonate a trusted third party to gain access to private data, continue to cost Internet users over a billion dollars each year. In this paper, we describe the design and performance characteristics of a scalable machine learning classifier we developed to detect phishing websites. We use this classifier to maintain Google's phishing blacklist automatically. Our classifier analyses millions of pages a day, examining the URL and the contents of a page to determine whether a page is phishing.

[2] Constructing classification models using skewed training data can be a challenging task. We present RUS Boost, a new algorithm for alleviating the problem of class imbalance. RUS Boost combines data sampling and boosting providing a simple and efficient method for improving classification performance when training data is imbalanced. In addition to performing favourable when compared to SMOTE Boost (another hybrid sampling/boosting algorithm), RUS Boost is computationally less expensive than SMOTE Boost and results in Significantly shorter model training times.

[3] A small subset of machine learning algorithms, mostly inductive learning based applied KDD 1999 Cup intrusion detection dataset resulted in dismal performance for user-to-root and remote-to-local attack categories as reported in the

recent literature. The uncertainty to explore if other machine learning algorithms can demonstrate better performance compared to the ones already employed constitutes the motivation for the study reported herein.

[4] Phishing causes billions of dollars in damage every year and poses a serious threat to the Internet economy. Email 1s still the most used medium to launch phishing attacks. In this paper, we are sent a comprehensive natural language-based scheme to detect phishing emails using features that fundamentally characterize phishing. Our scheme utilizes all the information present in an email, namely, the header, the links, and the text in the body. Although it is obvious that a phishing email is designed to elicit an action from the intended victim, none of the existing detection schemes use this fact to identify phishing emails.

[5] Phishing has become an increasing threat in online space, largely driven by the evolving web, mobile, and social networking technologies. Previous phishing taxonomies have mainly focused on the underlying mechanisms of phishing but ignored the emerging environments, and countermeasures for mitigating new phishing types. This survey investigates phishing attacks and anti-phishing techniques developed not only in traditional environments such as e-mails and websites, but also in new environments such as mobile and social networking sites.

III. PROBLEM DEFINITION

Before being trapped into phishing attack It can work on its avoidance. After study lots of details about phishing can avoid such conditions because of which user get into such crime. Before responding user gets very careful to respond on such e-mails who demand for personal information or offer some money. Typing of URL never ever click on the URL given in the e-mails. Go to the URL by typing them into browser window. If there is any chance of difference in URL then it gets reduced by typing it. Suspicious Website: if user find any suspicious about the web site, then user can check for its authenticity. By checking its https in the beginning of URL, padlock icon in the browser any sign which makes it different from original site. Use of secure browser: user must use the browser with latest security against phishing attack Use latest versions of browser with updated phishing filter. Fantastic offer: don't believe such offers that are not easy to believe check for the all-necessary details of the web site and ask too many questions before sharing any personal detail over the internet.

IV. EXISTING SYSTEM

A poorly structured NN model may cause the model to under fit the training dataset. On the other hand, exaggeration in restructuring the system to suit every single item in the training dataset may cause the system to be over 10 fitted. One possible solution to avoid the Over fitting problem is by restructuring the NN model in terms of tuning some parameters, adding new neurons to the hidden layer, or sometimes adding a new layer to the network. A NN with a small number of hidden neurons may not have a satisfactory representational power to model the complexity and diversity inherent in the data. On the other hand, networks with too many hidden neurons could over fit the data. However, at a certain stage the model can no longer be improved, therefore, the structuring process should be terminated. Hence, an acceptable error rate should be specified when creating any NN model, which itself is considered a problem since it is difficult to determine the acceptable error rate a priori. For instance, the model designer may set the acceptable error rate to a value that is unreachable which causes the model to stick in local minima or sometimes the model designer may set the acceptable error rate to a value that can further be improved. Disadvantage: • It will take time to load all the dataset. • Process is not accuracy. • It will analyse slowly.

V. PROPOSED SYSTEM

Lexical features are based on the observation that the URLs of many illegal sites look different, compared with legitimate sites. Analysing lexical features enables us to capture the property for classification purposes. It first distinguishes the two parts of a URL: the host name and the path, from which will extract bag of-words (strings delimited by '/', '?', '.', '=', '-' and ''). It finds that phishing website prefers to have longer URL, more levels (delimited by dot), more tokens in domain and path, longer token. Besides, phishing and malware websites could pretend to be a benign one by containing 11 popular brand names as tokens other than those in second-level domain. Considering phishing websites and malware websites may use IP address directly to cover the suspicious URL, which is very rare in benign case. Also, phishing URLs are found to contain several suggestive word tokens It check the presence of these security sensitive words and include the binary value in our features. Intuitively, malicious sites are always less popular than benign ones. For this reason, site popularity can be considered as an important feature. Traffic rank feature is acquired from Alexa.com. Host-based features are based on the observation that malicious sites are always registered in less reputable hosting centres or regions.

Advantage:

- All of URLs in the dataset are labelled.

DOI 10.17148/IJARCCE.2021.10736

• It is used two supervised learning algorithms random forest and support vector machine to train using sci kit-learn library.

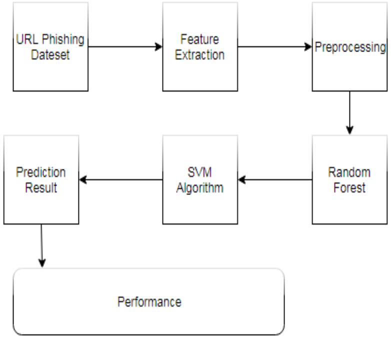

Figure 1: Architecture Diagram

VI. EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated based on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

VII. TESTING

Testing is performed to identify errors. Testing is used for equality assurance. Testing is an integrated part of the entire development and maintenance process. The goal or the entire during phase is to verify that the specification has been accurately and completely incorporated into the design, as well as to ensure the correctness or the design itself. Testing is one of the important steps in the software development phase.

TABLE 6.1 TEST CASE DESIGN

Test Case ID	**Test Objectives**	**Test Procedure**	**Test Input**	**Expected Result**	**Actual Result**
T01	Check the URL	Open the gui interface	Enter the URL link	The given link should match with the dataset	Matched successful
T02	Feature Extraction	Extraction each features of the given URL links	Separated Features of the link	The features data should match with the dataset	Features extracted successfully
T03	Check “benign”	To check the link is benign	Enter the good URL	Should display “benign”	Displayed Successfully
T04	Check “malware”	To check the link is malware	Enter the fake URL	Should display “malware”	Displayed successfully

T05	Check “Malicious”	To check the link is malicious	Enter the fake URL	Should display “malicious”	Displayed successfully

TABLE 6.2 TEST CASE LOG DESIGN

S. No	Test Case ID	Test Objectives	Test Status
1.	T01	Check the URL link	PASS
2.	T02	Feature Extraction	PASS
3.	T03	Check Benign or not	PASS
4.	T04	Check Malware or not	PASS
5.	T05	Check Malicious or not	PASS

VIII. CONCLUSION

Finally, phishing attacks are a major problem. It is important that they are countered. The work reported in this thesis indicates how understanding of the nature of phishing may be increased and provides a method to identify phishing problems in systems. It also contains a prototype of a system that catches those phishing attacks that evaded other defences, i.e., those attacks that have “slipped through the net”. An original contribution has been made in this important field, and the work reported here has the potential to make the internet world a safer place for a significant number of people. In the future it provides some technical solution by improve the efficiency of spam filters. By which too many mails are classified correctly and properly. By this legitimate user can surf internet with less fear. The user-phishing interaction model was derived from application of cognitive walkthroughs. A large-scale controlled user study and follow-on interviews could be carried out to provide a more rigorous conclusion. The current model does not describe irrational decision making nor address influence by other external factors such as emotion, pressure, and other human factors. It would be very useful to expand the model to accommodate these factors. It has theoretically and experimentally evaluated of Phish Limiter. It has evaluated the trustworthiness of each SDN flow to identify any potential hazards based on each deep packet inspection. Likewise, it has observed how the proposed inspection approach of two SF and FI modes within Phish Limiter detects and mitigates phishing attacks before reaching end users if the flow has been determined untrustworthy. Using our real-world experimental evaluation on GENI and phishing dataset, it has demonstrated that Phish Limiter is an effective and efficient solution to detect and mitigate phishing attacks with its accuracy of 90.39%.

IX. REFERENCES

1.Gunter Ollmann., (2007) “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBM Internet Security Systems.

2.Mahmoud Khonji, Youssef Iraqi., (2013) "Phishing Detection: A Literature Survey IEEE, and Andrew Jones.

3.Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset.

4.G. Liu, B. Qiu, L. Wenyin, (2010)"Automatic detection of phishing target from phishing webpage", Pattern Recognition (ICPR).

5.Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi, (2017) “Malicious URL Detection using Machine Learning.

6.A. Y. Fu, L. Wenyin, X. Deng, (2006) "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)", IEEE.

7.P. Prakash, M. Kumar, R. R. Kompella, M. Gupta, (2010)"Phishnet: predictive blacklisting to detect phishing attacks".

8.S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, C. Zhang., (2009) "An empirical analysis of phishing blacklists".

9.S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair., (2007)"A comparison of machine learning techniques for phishing detection".