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

AN AUTOMATED SYSTEM FOR DETECTION OF SOCIAL ENGINEERING PHISHING ATTACKS USING MACHINE LEARNING

Devika C.J Nair[1], Teslin Jacob[2]

[1]Student, Computer Science Department, Goa College of Engineering, Goa, India
[2]Assistant Professor, Computer Science Department, Goa College of Engineering, Goa, India

---

Abstract - *Social engineering is the act of breaking security by manipulating the users into divulging confidential information. Social engineering uses psychological tricks to gain trust of the humans to achieve sensitive information for various purposes. Phishing is a method of computer based social engineering attack. Phishing is a criminal act of acquiring personal information by sending out forged emails with fake websites and fraudulent weblinks in web pages. The aim of this research is to develop a system that will detect phishing URLs from the webpage and classify the URL whether it is legitimate or illegitimate URL using machine learning algorithm.*

Key Words: Phishing URL, Legitimate URL, Machine Learning, Logistic Regression, Prediction.

1. INTRODUCTION

Social engineering is a way of exploiting people into performing actions like getting into their confidential information. A social engineer uses human psychology to exploit people for his or her own use. The term social engineering applies to deception for the purpose of information gathering, fraud, identity theft or computer system access. Social engineering attacks are more challenging to manage since they depend on human behaviour and involve taking advantage of vulnerable employees.

Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. A phishing attack is when a criminal sends an email or the URL pretending to be someone or something he's not, in order to get sensitive information from the victim. The victim in regard to his/her curiosity may enter the details like username, password or credit card number and they are likely to get cheated.

Phishing becomes a threat to many individuals, particularly those who are not aware of the threats in the internet. Commonly, users do not observe the URL of a website. Sometimes, phishing scams engaged through phishing websites can be easily identified by observing whether a URL belongs to a phishing or legitimate website.

The problems of phishing implies that computer-based solutions are needed for guarding against these attacks along with user education. Having such a solution might enable the computer to have the ability to identify malicious websites in order to prevent users from interacting with them. A general approach to recognize illegitimate phishing websites is through the Uniform Resource Locator (URL). Even there are cases where the contents of websites are duplicated but still using URLs can distinguish real sites from phished websites. A common solution is to have a blacklist of malicious URLs developed by anti-virus groups. The drawback of this approach is that the blacklist cannot be exhaustive because new malicious URLs keep cropping up continuously. Therefore, approaches are needed that can automatically classify a new, previously unseen URL as either a phishing site or a legitimate one. For such type of solutions, machine learning based approaches are used where a system can categorize new phishing sites through a model developed using training sets of known attacks. One of the main problems with developing machine learning based approaches is that very few training data sets containing phishing URLs are available in the public domain. As a result, studies are needed that evaluate the effectiveness of machine learning approaches based on data sets that exists. The main goal of this research is to use machine learning algorithm on a large dataset where features from the data URLs have already been extracted and the class labels are available. Machine learning algorithm used in this research is Logistic Regression. The accuracy is also calculated with respect to the dataset and machine learning algorithm.

The remainder of this paper as follows: Section II describes the related work in classifying phishing websites Section III gives the details of methodology, Section IV describes the results of the tests and Section V gives the conclusion, limitations and directions for future work.

2. RELATED WORK

In survey paper [1], the authors in their survey study defines social engineering and explains how an attacker can read human mind to capture useful information. The author also provides recommendations on how to protect system against attackers using social engineering techniques. After conducting a survey, the authors have concluded that even after using the best and even the most expensive security technologies, an organization or a company or an individual is completely vulnerable. The authors also say that a key mechanism for combating social engineering must be their

education of potential victim in order to raise their awareness of the techniques and how to spot them.

In the paper "Phishing Website Detection using Machine Learning: Review" [2], the authors performed a detailed literature survey and proposed a new approach to detect phishing website by features extraction and machine learning algorithm. Phishing is a way to obtain user's private information via email or website. The study done by authors conclude that as there is lot of research work done, there is not any single technique which is enough to detect all types of phishing attacks. As technology increases, phishing attackers uses new methods day by day which enables to find a effective classifier to detect phishing attack. After a detailed research the authors have concluded that tree-based classifiers in machine learning is best suitable than any other methods.

The paper [3] for detection of phishing websites using machine learning proposed two methods for detection that is classification and association which will optimize the system. The proposed model focuses on identifying the phishing attack based on checking phishing website features, blacklist and WHOIS database. Selected features can be used to differentiate between legitimate and spoofed web pages. Features of URLs and domain names are checked using several criteria. These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites.

In the paper "Phishing Websites detection using Machine Learning" [4], the authors have carried out a method to develop methods for defense utilizing various approaches to categorize websites. The paper uses four classifiers: Decision Trees, Naïve Bayesian classifier, Support Vector Machine (SVM) and Neural Network. The features are extracted from the dataset and run on the four classifiers. Using decision trees for classification, the problem of overfitting occurred which was not feasible for the project. Neural Network did not perform well for the selected dataset because of less units in the hidden layer and feature values were discrete. SVM performed the best with respect to feature selection and dataset.

The author Waleed Ali [5], in his research has suggested methods to cope with the problem of growing web phishing attacks. This paper presents a methodology for phishing detection based on machine learning classifiers with a wrapper features selection method. The author has proposed some common supervised machine learning techniques with significant features selected using the wrapper features selection approach to accurately detect phishing websites. In the wrapper based evaluation, a search algorithm is used to search through the space of possible features and evaluate each subset by running a model on the subset. Supervised machine learning algorithms such as back propagation neural network (BPNN), radial basis function network (RBFN), support vector machines (SVM), naïve bayes (NB), decision tress (C4.5), random forest (RF) and k-nearest neighbor (kNN) were implemented. The experimental results showed that BPNN, KNN and RF achieved the best CCR while RBFN and NB achieved the worst CCR for detecting phishing websites. The machine learning classifiers based on the wrapper-based features selection accomplished the best performance in terms of CCR, TPR, TNR and GM.

The paper on "Detection and Prevention of Phishing websites using Machine Learning approach" [6], speaks about how the phishing problem is huge and having more than one solution to minimize all the problems effectively. The authors have come up with three approaches for phishing detection. First by analyzing various features of URL, second by checking legitimacy of website and third approach uses visual appearance -based analysis for checking genuineness of website. The machine learning algorithms used in this paper are logistic regression, decision trees and random forest. The linear regression plot of expected output verses predicted output was observed for random forest algorithm. The plot had a slight deviation from expected output. The authors concluded by saying the efficiency can be achieved by using hybrid solution of heuristic patterns, visual features and blacklist and whitelist approach to feed them to machine learning algorithms. The new system can be designed in this way to avail more accuracy.

The authors Meenu and Sunil Godara [7] uses several machine learning methods for predicting phishing emails. Their study mainly compares the predictive accuracy f1 score, precision and recall of machine learning algorithms like logistic regression, support vector machine, decision tree and neural networks. The paper also shows the improvement in logistic regression by using additional methods to logistic regression like using feature selection methods. The authors concluded by finally having a comparison with the four classifiers. Among all, the improved logistic regression gave best results with respect to evaluation metrics.

3. PROPOSED METHOD

The proposed model improves the accuracy by employing a pre-processing technique along with logistic regression algorithm. The system focuses on classifying the URL as phishing URL or legitimate URL. This classification is carried out using logistic regression function.

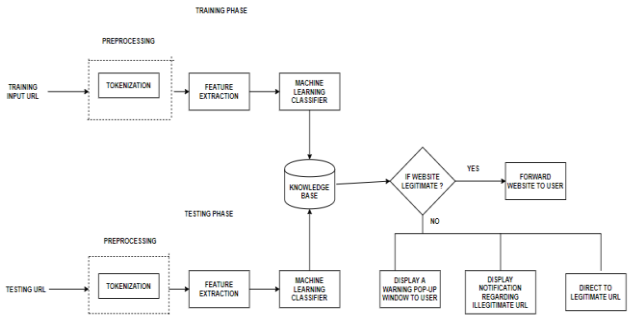
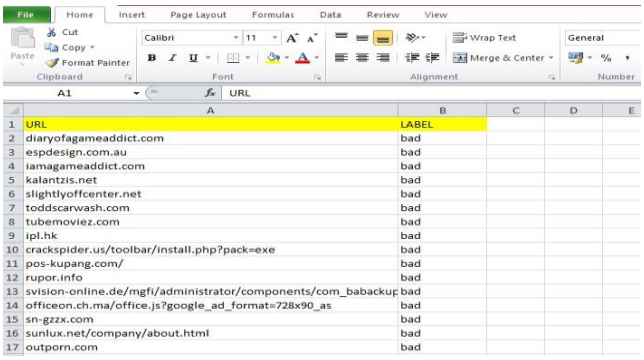


Fig. 1 Block Diagram

3.1 DATASET

The dataset is extracted from Github and PhishTank. Collected URL data is verified using websites with multiple website reputation engines and domain blacklisting services, to facilitate the detection of dangerous websites related to malware, phishing, scam and fraudulent activities. (eg : URLVOID). The dataset is then converted to comma-separated values file (CSV) using Microsoft Excel, Although supporting and encouraging the use of new XML-based formats as replacements, Excel 2007 remained backwards-compatible with the traditional, binary formats.



	A	B
1	URL	LABEL
2	diaryofagameaddict.com	bad
3	espdesign.com.au	bad
4	iamagameaddict.com	bad
5	kalantzis.net	bad
6	slightlyoffcenter.net	bad
7	toddscarwash.com	bad
8	tubemoviez.com	bad
9	ipl.hk	bad
10	crackspider.us/toolbar/install.php?pack=exe	bad
11	pos-kupang.com/	bad
12	rupor.info	bad
13	svision-online.de/mgfi/administrator/components/com_babackup	bad
14	officeon.ch.ma/office.js?google_ad_format=728x90_as	bad
15	sn-gzzx.com	bad
16	sunlux.net/company/about.html	bad
17	outporn.com	bad

Fig. 2 Dataset

3.2 PREPROCESSING (TOKENIZATION)

Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security.

3.3 LOGISTIC REGRESSION

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model. In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, in our algorithm its good or bad.

The logistic regression hypothesis is defined as :

$$h_\theta(x) = g(\theta^T x) \tag{1}$$

where the function g is a sigmoid function defined as :

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Therefore, the hypothesis can also be represented as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{3}$$

Hypothesis is interpreted as:

$$h_\theta(x) = p(y = 1 \vee x, \theta) \tag{4}$$

Since probabilities should sum up to 1, hypothesis is defined as

$$p(y = 0|x, \theta) = 1 - p(y = 1 \vee x, \theta) \tag{5}$$

3.3 KNOWLEDGE BASE

The classified URL will be stored in knowledge base and a new test URL will be tested based on the features of the already classified URL.

4. RESULTS

The predicted output using Logistic Regression is shown below:

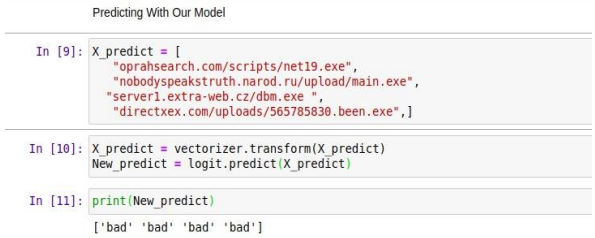


Fig. 3 URL prediction for phished URL

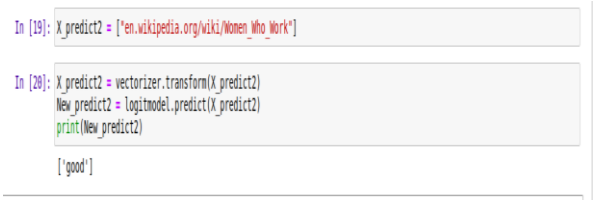


Fig. 4 URL prediction for legitimate URL

Using the fit() method on training dataset, calculates only the value and keeps it internally in the Imputer. Then, the transform() method is called on the test dataset with the same Imputer object. This way the value calculated for training set is saved internally in the object. To train the section of regression, vectorizer.transform() is used to check the output and accuracy.
The URL(Uniform Resource Locator) classification accuracy of classifying the URLs into phishing URLs (bad) or legitimate URLs(good) using regression gives an accuracy of .96 on the base of 1.

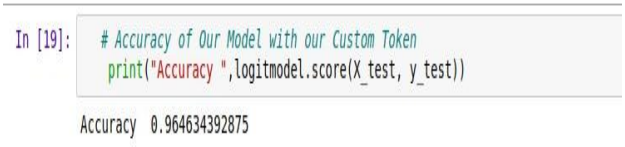


Fig. 5 Accuracy of the model

5. CONCLUSION

This research focuses on an efficient URL phishing detector by using Logistic Regression to classify the URLs. The automatic detection of phishing URLs is a method that has been implemented to gather a range of phishing website patterns. The proposed method has been implemented on a dataset of 4,20,467 phishing and legitimate URLs. The dataset URLs are analysed using features of the URL internally in the program code and the experiment furnish a classification of phishing and legitimate URL with 96% accuracy. This is an automated machine learning approach that rely on characteristics of phishing URL properties to detect and prevent phishing websites and to ensure high level security. The classification is done in Jupyter notebook (web based interactive environment).

6. FUTURE SCOPE

The same proposed technique can be used to develop a tool based on a web-browser add-on component like an extension to the webpage which can detect and prevent phishing websites on real time system. Different machine learning algorithms can be used to classify the URLs with respect to their features. Dataset size can be increased more by extracting more URLs from the webpage to obtain higher accuracy.

REFERENCES

[1] Anshul Kumar, Mansi Chaudhary and Nagresh Kumar, "Social engineering threats and awareness: A survey", European Journal of Advances in Engineering and Technology, pp-2(11): 15-19, ISSN:2394-658X,2015

[2] J Purvi Pujara and M.B Chaudhary, "Phishing website detection using Machine Learning: A review", International Journal of Scientific Research in Computer Science Engineering and Information Technology, Volume 3, Issue 7, ISSN: 2456-3307, 2018.

[3] Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, "Detection of Phishing Website using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume 5, Issue 3, ISSN:2395-0072, March, 2018.

[4] Arun Kulkarni, Leonard L Brown, "Phishing Websites detection using Machine learning", International Journal of Advanced Computer Science and Applications, Volume 10, Issue 7, 2019.

[5] Waleed Ali, "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection", International Journal of Advanced Computer Science and Applications, Volume 8, Issue 9, 2017.

[6] Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, Prof. S.P Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach", IEEE, 978-1-5386-5257-2/18, 2018.

[7] Meenu, Sunil Godara, "Phishing detection using Machine Learning techniques", International Journal of Engineering and Advanced Technology, Volume 9, Issue 2, ISSN: 2249-8958, December, 2019.

[8] Hanan Sandouka, Dr. Andrea Cullen, Ian Mann, "Social Engineering Detection using Neural Networks", IEEE International Conference on Cyber Worlds, 978-0-7695-3791-7/09, 2009.

[9] Weina Niu, Xiasong Zhang, Guowu Yang, Zhiyuan Ma, Zhongliu Zhuo, "Phishing Emails Detection using CS-SVM", International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC) , 0-7695-6329-5/17/31, 2017

[10] Yuki Sawa, Ram Bhakta, Ian G Harris, Christopher Hadnagy, "Detection of Social Engineering Attacks through Natural Language Processing Conversations", 2016 IEEE Tenth International Conference on Semantic Computing, 978-1-5090-0662-5/16, 2016.

[11] Abid Jamil, Syed Mudassar Alam, Muhammad Kashif Nazir, Zikha Ghulam, "MPMPA: A Mitigation and Prevention Model for Social Engineering Based Phishing Attacks on Facebook", 2017 IEEE International Conference on Big Data, 978-1-5386-5035-6/18, 2018.

[12] Sadia Afroz, Rachel Greenstadt, "PhishZoo: Detecting Phishing Websites by Looking at them", International Research Journal of Engineering and Technology (IRJET), Volume 6, Issue 4, ISSN:2396-1072, 2016