IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⪾ Impact Factor 7.39 ⪾ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11760

CLASSIFICATION OF A PHISHING WEBSITE

Emmanuel prasad S[1], Sheikh Abubakkar Siddiq[2], Srinidhi H R[3]

Student, NIE Institute of Technology, Mysore, Karnataka[1-2]

Assistant Professor, NIE Institute of Technology, Mysore, Karnataka[3]

Abstract: Phishing is a considerable problem differs from the other security threats such as intrusions and Malware which are based on the technical security holes of the network systems. The weakness point of any network system is its Users. Phishing attacks are targeting these users depending on the trikes of social engineering. Despite there are several ways to carry out these attacks, unfortunately the current phishing detection techniques cover some attack vectors like email and fake websites. Therefore, building a specific limited scope detection system will not provide complete protection from the wide phishing attack vectors. This paper develops detection system with a wide protection scope using URL features only which is relying on the fact that users directly deal with URLs to surf the internet and provides a good approach to detect malicious URLs as proved by previous studies. Additionally, Anti-phishing solutions can be positioned at different levels of attack flow where most researchers are focusing on client-side solutions which turn to add more processing overhead at the client side and lead to losing the trust and satisfaction of the users. Nowadays many organizations make centralized protection of spam filtering. This paper proposes a system which can be integrated into such process in order to increase the detection performance in a real time.

Keywords: Phishing, security threats, URLs, websites

I. INTRODUCTION

Phishing attack classically starts by sending an email that appears to come from an enterprise to victims asking them to update or confirm their personal information by visiting a link within the email. Although, phishers are now using several techniques in creating phishing sites, they all use a set of mutual features to create phishing websites since without those features, they lose the advantage of deception. This helps us to differentiate between honest and phishy websites based on the features extracted from the visited website. Overall, two approaches are used in identifying phishing sites. The first one is based on a blacklist, in which the requested URL is compared with those in that list. The downside of this approach is that the blacklist usually cannot cover all phishing websites since, within seconds, a new fraudulent website is launched. The second approach is known as heuristic-based methods, where several features are collected from the website to categorize it as either phishy or legitimate.

II. LITERATURE SURVEY

A literature survey or review which combines both summary and synthesis of specific conceptual categories. Literature survey gives conclusion about how one can analyses and understand gaps exist in how a problem has been researched to date. A literature review surveys, scholarly articles or any other resources which are relevant to our area of interest in the research provides a brief description and critical evaluation of works which are related to our research problem. Literature survey provides an overview of sources that we have explored or referred during our research, whereas in this project we have referred IEEE papers for survey review.

In [1], a system is proposed using Fuzzy Logic as classifier, to overcome the problem of phishing they have designed a framework to detect it using the fuzzy logic as a classifier. They have used only URL based features for the extraction process [10]. For that they have collected dataset from the Phish tank site, Open phish site and the URL of the website which can live on the web. they collect the 1000 URLs for our dataset purpose. In [2], a system is proposed A Hybrid Model based approach has been proposed target to solve the phishing web sites problem. A single model cannot efficiently detect the phishing websites because there is needed to enhance and tuning the single model or second approach is to combine any two or three models for improving the accuracy for detecting the phishing-sites attack.
In [3], a system, study considers the phishing detection problem as an AI-based classification problem wherein the result of the decision-making phase leads to detecting if a given website is either a legitimate or a phishing website. Thus, consideration of the AI meta-learner algorithms as the basis for developing a credible and viable phishing website detection model to combat phishing threats and its evolving nature was made. In [4], system is proposed to derive

DOI: 10.17148/IJARCCE.2022.11760

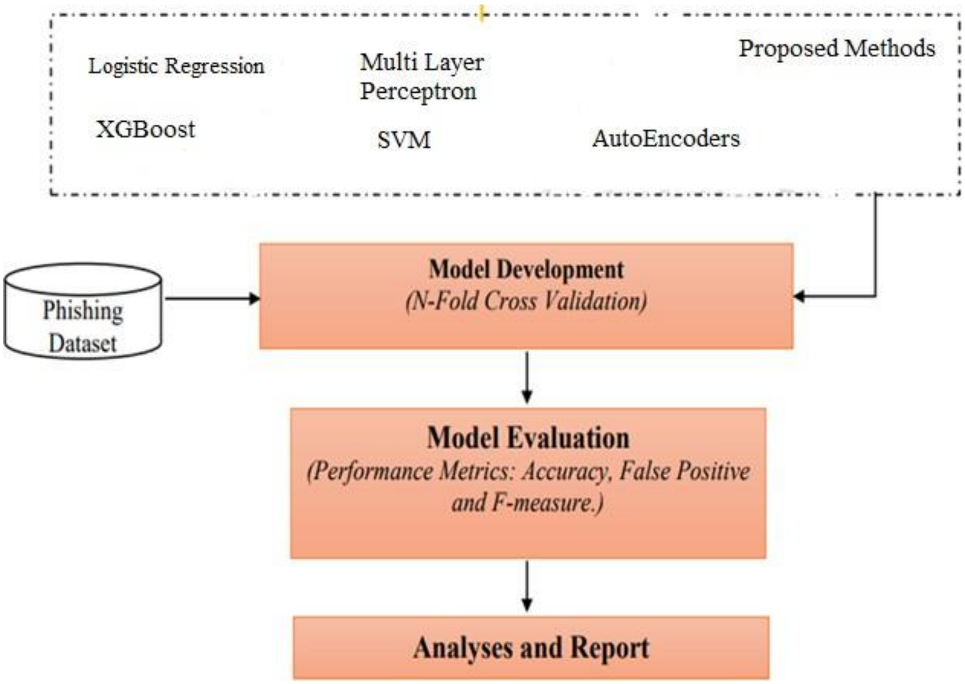
classification models that detect phishing websites by analysis of the lexical and host-based features of URLs. We analyze different classifying algorithms in Waikato Environment for Knowledge Analysis (WEKA) workbench and MATLAB. The work consists of host based, page based and lexical feature extraction of collected URLs and analysis. The first step is the collection of phishing and benign URLs. In [5], this paper they surveyed currently used features in automated spam / phishing email detection systems and extracted 40 features from a body of over 10,000 emails, which were divided amongst three classes ham, spam and phishing. We then calculated the information gain of all of these features. From this we created C5.0 classifiers using three groups of features, those with the best IG values, the median IG values, and finally the worst IG values. As expected, in each case, the classifier trained on the best features outperformed all the others. In [6], The primary focus of this study is covering the software detection approaches of phishing attack detection. It is explained blacklist-based phishing detection approaches, it is analyzed phishing domains and extracted its distinguishing features from legitimate domains, and how they can be detected using machine learning and natural language processing. In [7], proposed system Fuzzy Rough Set is used theory as a tool to select most effective features from three benchmarked data sets. The selected features are fed into three often used classifiers for phishing detection. To evaluate the FRS feature selection in developing a generalizable phishing detection, the classifiers are trained by a separate out-of-sample data set of 14,000 website samples. In [8], a proposed system a new end-host based anti-phishing algorithm is used, which is called as LinkGuard, by utilizing the generic characteristics of the hyperlinks in phishing attacks. These characteristics are derived by analyzing the phishing data archive provided by the Anti- Phishing Working Group (APWG). Because it is based on the generic characteristics of phishing attacks. LinkGuard is implemented in windows XP. Experiments verified that LinkGuard is effective to detect and prevent both known and unknown phishing attacks with minimal false negatives.

In [9], the paper surveys the current achievements, studies their limitations, restates what induction factors need to boost for a successful real-time application. Consequently, future outlooks are recommended on how to devote well-performed anti-phishing scheme.

III. METHODOLOGY

Proposed System:

- Classification of hyperlinks based on the attributes provided by Anti Phishing Working Group (APWG).
- This project aims to implement an anti-phishing technique using few algorithms namely Logistic Regression, SVM, XGBoost, Multilayer Perceptron and AutoEncoders.
- As the existing systems only aims at tree-based algorithms (such as Decision Trees, Random Forest), our project also aims to detect new phishing sites which are not yet blacklisted and targeted attack against small brokerages and corporate intranets.



DOI: 10.17148/IJARCCE.2022.11760

METHODOLOGY:

Machine learning is one of the most exciting recent technologies. Machine learning had been positioned to address the shortages of human cognition as well as information processing, specifically in handling large data, their relations and the following analysis. In general, machine learning studies the research and algorithms construction that can learn from, and derive predictions about, data. Therefore, the machine learning approach is selected to predict whether a website, according to a dataset with some extracted features, is legitimate or phishing.

Some extracted features acquire the same influence level on classifier accuracy to predict phishing sites and are considered as redundant. Optimization classification performance was conducted in determining the most effective features among all the features extracted. Various feature selection methods were applied to reduce the features that are not relevant and group the reduced features as a new subset. Finally, the experiments required in analyzing the extent to which the established machine learning techniques are effective in determining the most effective subset of features were also carried out.

IV. CLASSIFICATION TECHNIQUES OF FISHING

1. MULTILAYER PERCEPTRON:

Multilayer Perceptron (MLP) is an artificial neural network model which could be employed for data classification. Artificial neural network terminology is the way human brain neuron function and also interact simultaneously for recognition, reasoning, as well as recovery of damage. It is also called a multi-layer feed forward neural network. This algorithm learns by finding the most suitable synaptic weight in classifying patterns in the training dataset. Neurons in the network are being connected with one another through a link called synaptic. Multilayer perceptron is an artificial neural network structure which is also a nonparametric estimator that can be employed for classifying and detecting intrusions.

2. SUPPORT VECTOR MACHINE :

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane. Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them.

3. LOGISTIC REGRESSION :

This classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, Logistic Regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. Logistic regression works well when the relationship in the data is almost linear despite if there are complex nonlinear relationships between variables, it has poor performance. Besides, it requires more statistical assumptions before using other techniques.

4. XGBOOST :

XGBoost is a refined and customized version of a Gradient Boosting to provide better performance and speed. The most important factor behind the success of XGBoost is its scalability in all scenarios. The XGBoost runs more than ten times faster than popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings. The scalability of XGBoost is due to several important algorithmic optimizations. These innovations include a novel tree learning algorithm for handling sparse data; a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning.

5. AUTOENCODERS :

An autoencoder is a type of artificial neural network used to learn efficient codlings of unlabeled data (unsupervised learning). The encoding is validated and refined by attempting to regenerate the input from the encoding.

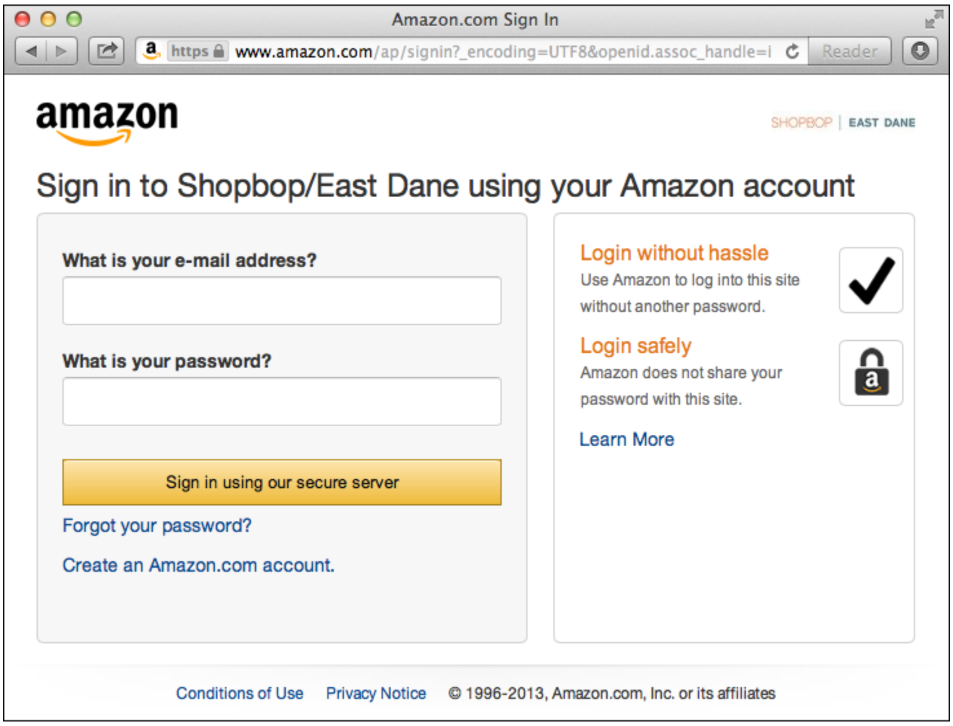
V. IMPLEMENTATION

Technique of phishing :

Phishing is a non-ethical mechanism comprising both social engineering and technical tricks to steal user’s personal data and financial account credentials. Some of the social engineering schemes use spam e-mails, pretending to be from legitimate businesses or agencies, that are specially designed to lead users to knock-off websites that trick recipients to fall into the trap which steal financial data such as usernames and passwords. Technical intrigue schemes install malicious software onto the systems (computers), to steal credentials directly, often using systems to intercept users online account user names and passwords.

DOI: 10.17148/IJARCCE.2022.11760

Overall, two approaches are used in identifying phishing sites. The first one is based on a blacklist, in which the requested URL is compared with those in that list. The downside of this approach is that the blacklist usually cannot cover all phishing websites since, within seconds, a new fraudulent website is launched. The second approach is known as heuristic-based methods, where several features are collected from the website to categorize it as either phishy or legitimate. In contrast to the blacklist method, a heuristic-based solution can recognize freshly created phishing websites. The accuracy of the heuristic-based methods depends on picking a set of discriminative features that could help in distinguishing the type of website. Data mining is one of the research fields that can make use of the feature knowledge that ensures correctness, reliability and completeness, as well as reduce the time of knowledge achievement. Several studies have been conducted about phishing detection based on website features but these researches were unable to identify precise rules to classify the type of website.



Amazon.com Sign In

https www.amazon.com/ap/signin?_encoding=UTF8&openid.assoc_handle=i

amazon

SHOPBOP | EAST DANE

Sign in to Shopbop/East Dane using your Amazon account

What is your e-mail address?

What is your password?

Login without hassle

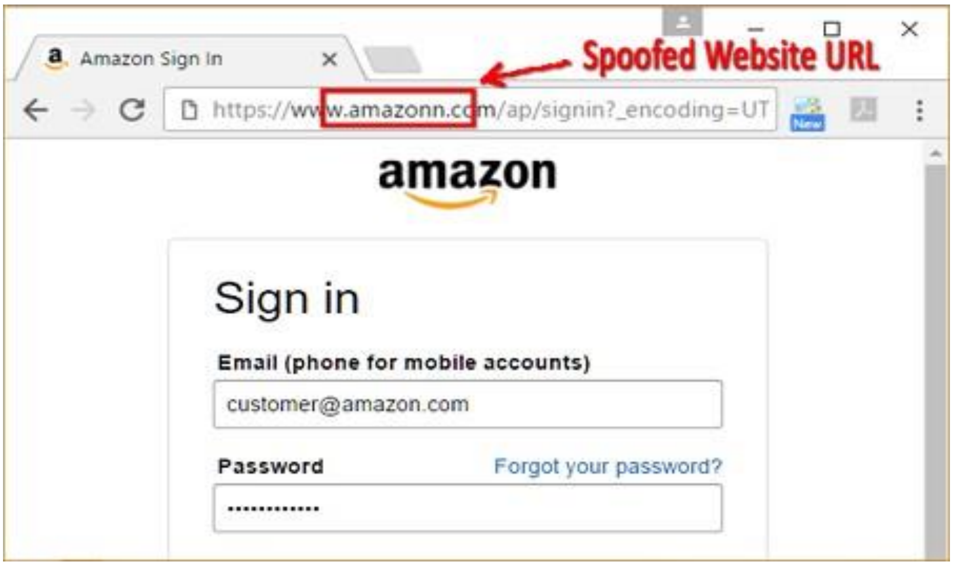
Use Amazon to log into this site without another password.

Login safely

Amazon does not share your password with this site.

Reader
Sign in using our secure server
Forgot your password?
Create an Amazon.com account.
Learn More
Conditions of Use
Privacy Notice
© 1996-2013, Amazon.com, Inc. or its affiliates

Figure 1: Original Amazon sign in page



Amazon Sign In

https://www.amazonn.com/ap/signin?_encoding=UT

amazon

Sign in

Email (phone for mobile accounts)

customer@amazon.com

Password

••••••••••••

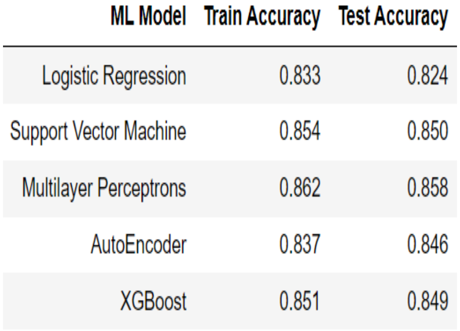
Figure 2: Phishing Webpage

RESULTS

The main findings of our preliminary work include:

- Phishing URLs and domains exhibit characteristics that are different from other URLs and domains.
- Phishing URLs and domain names have very different lengths compared to other URLs and domain names in the Internet.

Many of the phishing URLs contained the name of the brand they targeted.



ML Model	Train Accuracy	Test Accuracy
Logistic Regression	0.833	0.824
Support Vector Machine	0.854	0.850
Multilayer Perceptrons	0.862	0.858
AutoEncoder	0.837	0.846
XGBoost	0.851	0.849

Table 1: Classifier Performance

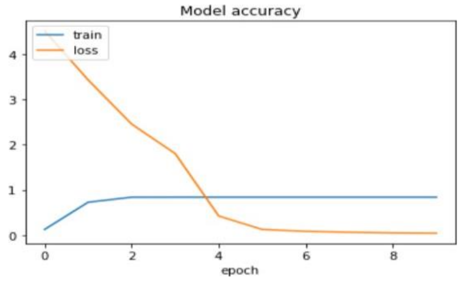


Figure 2: Loss function for Auto Encoders

We have found that our system provides us with 85.1 % of accuracy for XG Boost Classifier, 85.4% accuracy for SVM Classifier, 83.3 % accuracy for Logistic Regression Classifier, 83.7% accuracy for AutoEncoders and finally 86.2 percentage of accuracy when using Multi-Layer Perceptron Classifier. Hence, we found that the best among all the above classifiers is MLP and SVM Classifier which shows maximum accuracy. The proposed technique is much more secured as it detects new and previous phishing sites.

VI. CONCLUSION

Phishing is a major problem, which uses both social engineering and technical deception to get users' important information such as financial data, emails, and other private information. Phishing exploits human vulnerabilities; therefore, most protection protocols cannot prevent the whole phishing attacks. Many of them use the blacklist/whitelist approach, however, this cannot detect zero-hour phishing attacks, and they are not able to detect new types of phishing attacks.

DOI: 10.17148/IJARCCE.2022.11760

This project aims to enhance detection method to detect phishing website using machine learning technology. We achieved 86.2% detection accuracy using multilayer perceptron algorithm. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing website more accurately, for which multilayer perceptron algorithm of machine learning technology and blacklist method will be used.

VII. REFERENCE

[1]. Microsoft,“MicrosoftSecurityIndex Report”https://news.microsoft.com/en- sg/2014/02/11/microsoftconsumer-safety-index-reveals-impact-of-poor-online- safetybehaviours-in-Singapore/ #sm.0000c8bivc14h3dyfvxak6545kbcz #4FXDf2H3VbYmD1b1.97 , accessed May 2017.

[2]. W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web-based systems." in Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech, Morocco: IEEE, July 2008, pp. 326- 331.

[3]. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, “Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions,” in Proceedings of the 28th international conference on Human factors in computing systems, ser. CHI ’10. New York, NY, USA: ACM, 2010, pp. 373–382.

[4]. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” in Proceedings of the 6th Conference in Email and Anti-Spam, ser. CEAS’09, Mountain view, CA, July 2009.

[5]. Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), pp.2091-2121.

[6]. Google, “Google safe browsing API,” http://code.google.com/apis/safebrowsing/ , accessed May 2017.

[7]. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “Phishnet: predictive blacklisting to detect phishing attacks,” in INFOCOM’10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.

[8]. Y. Cao, W. Han, and Y. Le, “Anti-phishing based on automated individual white-list,” in DIM ’08: Proceedings of the 4th ACM workshop on Digital identity management. New York, NY, USA: ACM, 2008, pp. 51–60.