Evaluation of Phishing Techniques Based on Machine Learning

Merlin .V.Kunju
CSE Department
KITS
Coimbatore, India
merlin.v.k1996@gmail.com

Mrs Esther Dainel
Assistant Prof- CSE
KITS
Coimbatore, India
estherdainell@gmail.com

Heron Celestie Anthony
CSE Department
KITS
Coimbatore, India
indu.celes@gmail.com

Sonali Bhelwa
CSE Department
KITS
Coimbatore, India
sonalibhelwa97@gmail.com

Abstract.

Phishing sites is the major problems for online security challenges because of large number of online transactions is done every day. The objective of the paper is to do survey about the phishing: A social attack and its detection and to make aware of the users who doesn't know about this major attack as many of them are still falling in the trap. Most of the users are unaware about this problem; they unknowingly fill many forms that belong to phishing website which are hidden. This leads to the leaking of sensitive information of the victim. This study also gives brief knowledge about several machine learning techniques such as kNN Algorithm, Naïve Bayes, Decision Tree, Support Vector Machines, Neural Network and Random Forest algorithm for predicting phishing sites.

Keywords:, Phishing, detection, attacks, machine learning.

I. INTRODUCTION

Most of the experienced users know to detect the phishing site but due to the speed of life, user ignores the uncertainty in the URL's and hence leading to leaking the sensitive information of the particular user in a phishing site. On the other side most of the users are not aware of the detail knowledge about the URL's and they tend to fall in the trap of phishers site. Phishing attack is majorly increased while the technology is increasing. The first phishing attack in the world is AOL (American Online) in early 1990's. [1]

II. TYPES OF PHISHING

Spear Phising: Spear phishing could be a direct attack on the corporate or a selected website. This principally steals the fundamental information of the actual company's users. It is customized attack on specific employee or a group or a company.[2]

Clone Phıshıng: The phisher as to make a nearly identical replication of the legitimate website or message which can easily trick the victim and which eventually falls into the trap lead by phishers.

Whalıng: Whaling could be a common cyber-attack that happens once associate degree aggressor utilizes spear phishing ways to travel when an oversized, high-profile target.

III. METHODS FOR PHISHING ATTACKS

Browser Vulnerabilities: It's alerted that hackers use vulnerability within the internet browsers to point out the unofficial domains as official websites like Google, Yahoo, Amazon, Apple.[3]

Click Jacking: Click jacking is also called as user interface (UI) redressing attack that manipulates the User interface of a websites which leads the innocent people playing associate work accidentally. Any Click by the user on the website permits to unseaworthy of their sensitive information. [4]

Cross- website Scripting: It's a vulnerability exploit of a site that provide access to the phishers to feature a harmful code and use the custom URL to a site. JavaScript is that the most used language for XSS attack. [3]

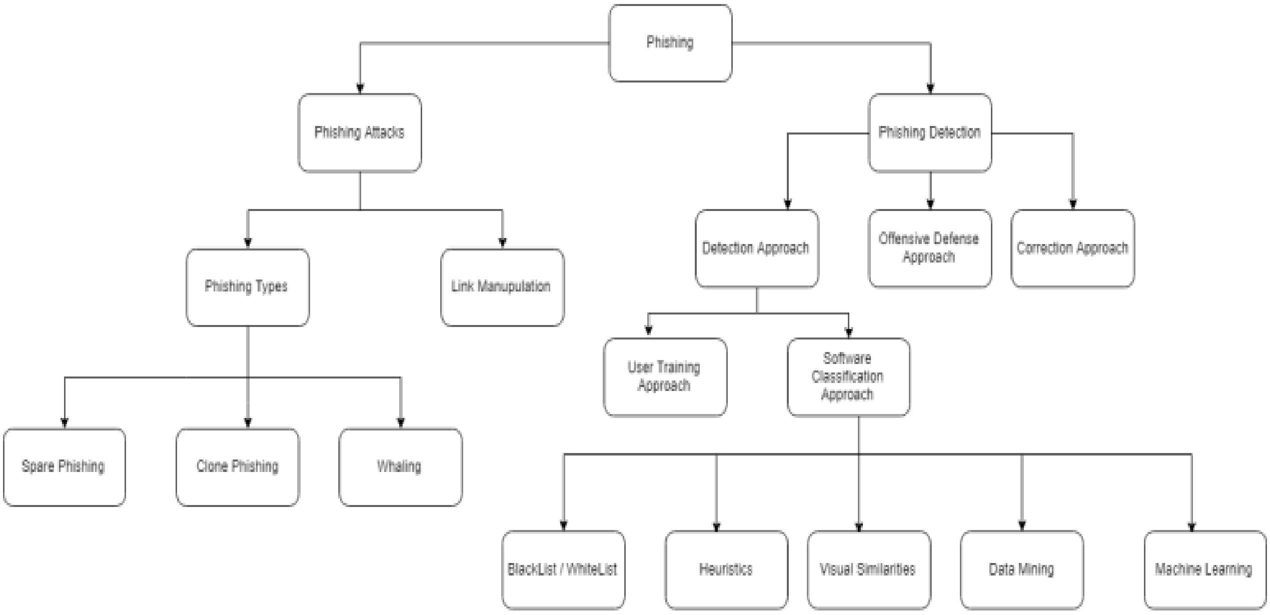

Figure 1 Classification Diagram of Phishing and its Detecting Techniques

Drive-by-Download: It's a delivery technique that inserts a machine smidgeon some harmful virus or shell code by simply visiting the web site or seeing associate markup language[5].

A. Detection Approaches

There are two kinds of Detection approaches which include:

User preparing approaches- Users must be taught for improving comprehension of malicious assault that accurately driven it to distinguish phishing and authentic sites.

Software grouping approaches-Aim of this arrangement is to recognize phishing and authentic sites for the benefit of client the engineers create programming apparatuses that secure and distinguish phishing assaults as the client disregards. Phishing programming order methodologies, for example, Blacklist, White-list, Heuristics, Visual similitude, Machine learning. [6]

*Blacklist – White-list***:** It is a series of different Internet Protocol (IP) addresses, URL's, or keywords that are very harmful. It has been collected using methods in which user interest is prior. Whereas, Whitelist are the opposite, when compared to blacklist and can be used to reduce False Positive rates. Some of the best examples are as follows: (i) Google Safe Browsing API (ii) DNS based API**.**

Heuristics: A heuristic based methodology uses standard highlights of spam site, for example, the enlistment structure, login site, URL and Web traffic to take certain decision. Babagali et al has used nonlinear algorithm with two feature such as decision tree and wrapper class.[7]

- *Phish-Guard: A Browser Plug-in***:** This injects numerous amounts of fake credentials into a suspicious domain and identifies phishing on the basis of response.As the user logs in the username it fires random passwords for every username

- *Phish-Wish:* Phish-Wish is a stateless phishing channel utilizing negligible principles. This arrangement points towards giving better insurance against party time assault than boycott, Minimal false positive, a arrangement that requires generally negligible asset.

Visual Similarity: Visual Comparability is the fundamental thought of any phishing assaults. It is a based phishing identification that checks for the likeness with the content pieces, picture inserted in sites by large aspects of the malicious and legitimate site.

- *CANTINA:* CANTINA is a task that finds out malicious website by understanding its data. Xiang Hong used a feature wealthy machine leaning frame work and come up with 2 vital options of the phishing scams (i) an internet site that's a alikeness of a establishment site. (ii) A faux login page inquiring for sensitive information [8]. GUANG XIANG used feature-rich machine learning approach which is used to detect phishing web. the number of features from their previous work was increased when compared to this one.[9]

Data Mining: Information emulating is a procedure of looking through a lot of information and selecting important data. It is portrayed as the nontrivial extraction of already obscure and possibly helpful data from vast datasets.

Machine Learning: There are four sorts of machine learning techniques that is supervised, unsupervised, Semi-supervised and reinforced.[10] Michal Trnka has worked on Authentication based on user and domain level which eventually Increases the security of communication.[11] Classifiers Dependent on the techniques of machine learning for phishing detection are isolated into:

Bag-of-word model-based methods

It is a rearranged portrayal strategy for common language preparing. It is for the most part utilized in techniques for record arrangement.

Support Vector System: It is the most well-known classifier nowadays. This technique finds the ideal isolating hyper plane between two classes by augmenting the edge between the classes nearest point [12] M.Chandrasekaran,et al, has worked on the version between MTA and MUA, in which he used NLP to analyse each sentence.[13]

K-Nearest neighbour : It is a technique used to grouping and relapse. Linping Ma has used 13 orthographic features which gave highly effective and achieve reliable results.[14]

*Bart(Biayasian Additive regression Tree)***:** It is a learner to predict contitentive outcome for obeservation via regression.AbuNimeh et al has Compared six classifier relating to machine learning and proved that there is no standard classifier for phishing email prediction.[16]

Neural Networks: A neural system is structure as a lot of interconnected neurons. It comprises of three layers input layer, concealed layer and yield layer. Interconnections are utilized to signals from one neuron to the next. Feng et al has worked on Neural network based classification with a simple and stable Monte Carlo algorithm, which does not depend on third parties and has stability of detection and the accuracy rate is increased.[17]

AdaBoost: It very well may be utilized the execution of any AI calculation. It is best utilized with powerless student. This calculation learns a solid calculation by joining a lot of powerless calculation and set of weight. The weights are found out to manage preparing disconnected.

*Decision tree***:** A choice tree is a tree like diagram or a model. It resembles altered tree since it has its root at the best and develops downwards. This grouping model is utilized to foresee the estimation of an objective property dependent on a few info quality.[18]

Random forest algorithm: It's Associate in Nursing approach for classification and regression methodology appropriate for handing issues involving grouping of knowledge into categories.Ozgur Koray Sahingoz has proposed NPL based features which is used to focus on the usage of the word in the url without performing any other operations.[19] This approach has 2 toolbars that helped the attackers to seek out the important information:-

Table 1:Comparision of the Phishing Methodology

#	Title	Algorithm	Methodology	Advantage	Disadvantage
1.	Phishing E-mail Detection based on Structural properties[13]	Support Vectore Machine (SVM)	The prototype implementation between MTA(mail trasfer agent) and MUA(mail user agant)	Uses NLP to detect appropriateness of each sentence.	Small size dataset only can be used. It is time consuming.
2.	Machine learning based phishing detection from URL [20]	Random Forest Algorithm, SVM, adaboost, KNN, Naïve bayes,	A anti-phishing detection method using 7 different machine learning	NPL based features is used in which it focus on the use of the words present in url .	Machine learning will not be more effective to handle with the usage of huge dataset .
3.	A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites [9]	CANTINA+	Feature-rich machine learning technique is used to indentify phishing site(which implements a content-based system)	Understand the evolving phishing attacks. İt increase the number of specific values from the work they have done preciously	They used onlya limited number of dataset (8118 phishing And 4883 Orginal web pages) Third-party services is used
4.	Heuristic nonlinear regression strategy for detecting phishing websites. [7]	Valued heuristic based nonlinear regression algo and the feature selection approach	A heuristi based nonlinear regression algorithm by using two feature selection methods: decision tree and wrapper.	Original dataset was reduced and this feature set will get a best result using decision trees.	They used less dataset (11055 phishing and real web pages).
5.	The application of the novel neural network in the detection of phishing sites.[17]	neural network classification method	Monte Carlo algorithm.was used with Neural network based classification	Stability of detection and the accuracy rate is increased.	need to download the whole page.
6.	Detecting Phishing Attacks Using Natural Language Processing and Machine LearningTianrui Peng [15]	Naïve bayes classifier	NLP techniques and machine learning using naïve bayes classifier was used to identify phishing email	Use NLP to detect appropriateness of each words.	Rely on text analysis of the emails. Machine Learning Is used to eastablish the blacklist of virus pairs
7.	Survey of Authentication	Machine learning features	Authentication based on user and	Increases the security of	Receiver and sender side should have same technology

	and Authorization for the Internet of Things [11]		domain level	communication	
8.	A Comparison of Machine Learning Techniques for Phishing Detection [16]	Logistic Regression, CART, Support Vector Machine, NNET, BART, Randome Forest	Compared six classifier machine learning and came with the conclusion that there is no standard algorithms to detect phishing email	43 features are used for train and test the classifier.	Time and memory was used up if more features were used.
9.	Stablishing phishing provenance using orthographic features [14]	K-means clustering algorithm	13orthographic features, created the targeted functions, clustering email	Highly effective and achieve reliable results	Offline technique only can be used in K-mean algorithm

Offensive defence approaches

Naïve Bayes Classifiers: Gullible bayes is a standout amongst the most proficient and best inductive learning calculations for AI and information mining.Tianrui Peng has used NLP to detect appropriateness of each sentence by using naïve bayes classifier.the disadvantage was that ML was used to construct the blacklist of malicious pairs.[15]

Boosting: It is an AI troupe meta-calculation for decreasing inclination and furthermore change. Its additionally changes over powerless student to solid ones. *Multi-Classifier Algorithm based methods.* This methodology is correlation between the arrangements of the classifiers.

Logistic Regression: It is the most broadly utilized measurable model for parallel information forecast. It is outstanding for its effortlessness and extraordinary interpretability. This performs well when the connections in the information are straight.

Bogus Biter: this is often a novel consumer side anti-phishing tool that's complementary to existing bar based mostly mechanisms. this is often clear to users. This submits faux information into hypertext markup language forms, rather than showing alert messages to the user, whenever the phishing web site is visited.

ML based enemy of phishing classifiers has demonstrated that it is conceivable to accomplish genuine positive rate as over 99% and false positive rate as under 1%.

IV. CONCLUSION

Phishing has become the major issue faced by the internet security and economy all over the world. The rate of phishing attack is highly increasing day by day through many ways. User education or training aims to develop the technical awareness level of the end-user to lower their vulnerability to phishing attack. As a solution for it, many phishing techniques are developed and many attacks are detected. Some software solutions are Blacklists, Visual Similarity and many more. Phishing attacks have severe negative influence on web and user's trust. The major issues implementing these algorithms are finding the right feature set for a particular phishing attack. Therefore, we presented a paper which shows the survey of different ways of detecting phishing techniques and algorithms. This survey provides better understanding about the phishing detection techniques, many problem solving solutions. It also briefs that some approaches have limitations like its accuracy and failing to detect the phishing attack. All phishing attacks cannot be solved with these algorithms. On one side users are ignorant about the attacks and on the other hand the phishers easily attack the victims by many malicious ways. A Comparison table is been prepared to tally the advantages, drawbacks, methodologies of the various approaches. By using single techniques we cannot adopt ourselves for phishing detecting purposes. Detecting of phishing with high reliability is still a bigger challenge for further development and research

V. REFERENCES

[1]https://www.uniassignment.com/essay-samples/information-technology/the- social-impact-of-phishing-scams-information-technology-essay.php

[2] https://www.techopedia.com/definition/4121/spear-phishing

[3] Chiew, Kang Leng, Kelvin Sheng Chek Yong, and Choon Lin Tan. "A survey of phishing attacks: their types, vectors and technical approaches." Expert Systemswith Applications (2018)

[4] https://en.wikipedia.org/wiki/Clickjacking

[5] https://en.wikipedia.org/wiki/Drive-by_download

[6] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," IEEE Commun. Surv. Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013

[7] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," Soft Comput., pp. 1–13, 2018.

[8] Y.Zhang, J.I.Hong, andL.F.Cranor,"Cantina:a contentbased approach to detecting phishing websites,"in Proceedings of the 16th International WorldWide Web Conference (WWW'07),pp. 639–648,Banff,Canda,May2007.

[9] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, p. 21, 2011.

[10] https://www.lifewire.com/what-is-whaling-2483605

[11] H. Kim and E. A. Lee, "Authentication and Authorization for the Internet of Things," IT Prof., vol. 19, no. 5, pp. 27–33, 2017.

[12] S. Abu-nimeh, D. Nappa, X. Wang, and S. Nair, "<P60-Abu-Nimeh.Pdf>," pp. 60–69, 2007.

[13] J. Drew and T. Moore, "Automatic identification of replicated criminal websites using combined clustering," Proc. - IEEE Symp. Secur. Priv., vol. 2014–Janua, pp. 116–123, 2014

[14] L. Ma, J. Yearwood, and P. Watters, "Establishing phishing provenance using orthographic features," 2009 eCrime Res. Summit, eCRIME '09, no. November, 2009.

[15] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[16] S. Abu-nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," pp. 60–69, 2007.

[17] F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Q. Wang, "The application of a novel neural network in the detection of phishing websites," J. Ambient Intell. Humaniz. Comput., vol. 0, no. 0, pp. 1–15, 2018.

[18] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert Syst. Appl., vol. 37, no. 12, pp. 7913–7921, 2010.

[19] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Syst. Appl., vol. 117, pp. 345–357, 2019.