Phishing Detection Using Machine Learning Technique

Conference Paper · December 2020

DOI: 10.1109/SMART-TECH49988.2020.00026

4 authors, including:

Toqeer Mahmood
National Textile University
75 PUBLICATIONS **1,902** CITATIONS
SEE PROFILE

Muhammad Wasif Nisar
COMSATS University Islamabad
78 PUBLICATIONS **1,448** CITATIONS
SEE PROFILE

Tahira Nazir
Riphah International University
54 PUBLICATIONS **1,103** CITATIONS
SEE PROFILE

Phishing Detection Using Machine Learning Technique

Junaid Rashid
Department of Computer Science,
Air University Islamabad, Kamra Campus, Pakistan
junaidrashid062@gmail.com

Toqeer Mahmood
(Corresponding Author)
Department of Computer Science,
National Textile University Faisalabad, Pakistan
toqeer.mahmood@yahoo.com

Muhammad Wasif Nisar
Department of Computer Science,
COMSATS University Islamabad, Wah Campus, Pakistan
wasifnisar@gmail.com

Tahira Nazir
Department of Computer Science,
University of Engineering amd Technology,Taxila, Pakistan
tahira.nazir77@gmail.com

***Abstract*— Today, everyone is highly dependent on the internet. Everyone performed online shopping and online activities such as online Bank, online booking, online recharge and more on internet. Phishing is a type of website threat and phishing is Illegally on the original website Information such as login id, password and information of credit card. This paper proposed an efficient machine learning based phishing detection technique. Overall, experimental results show that the proposed technique, when integrated with the Support vector machine classifier, has the best performance of accurately distinguishing 95.66% of phishing and appropriate websites using only 22.5% of the innovative functionality. The proposed technique exhibits optimistic results when benchmarking with a range of standard phishing datasets of the "University of California Irvine (UCI)" archive. Therefore, proposed technique is preferred and used for phishing detection based on machine learning.**

Keywords—phishing, web, machine learning, principal component analysis, support vector machine

I. INTRODUCTION

The web has become a platform to help various criminal companies like spam, financial fraud and operators spread malware. The correct commercial reasons for this plan may be different, but a common thread is a requirement that users don't have to visit their site. This visit should be possible using email, web query items, or connections from other site pages, however, the client must snap to make a move, for example, indicating the ideal URL (Uniform Resource Locator) and get significant data. To overcome this, the security community responded by developing a blacklist service packaged in toolbars, devices and search engines, providing warnings or alerts with accurate feedback. The site is too new, unclassified, or misclassified, so many harmful sites are not blacklisted.

Phishing is a kind of cyber-attack that utilizes sites to take solid buyer insights like store card numbers, accounts, login qualifications, and that's only the tip of the iceberg. In June 2018, "APWG (Anti-Phishing Working Group)" explicitly proposes 51,401 phishing sites [1]. According to another RSA report, phishing incidents cost around $ 9 billion worldwide. According to statistics in 2016 [2], traditional anti-phishing options and efforts have been ineffective.

The anti-phishing solution is most extensively deployed on a blacklist warning system, which is existing in common web browsers like Chrome, Internet Explorer and Mozilla Firefox. The blacklisting interrogative gadget has a central database of regarded phishing URLs, and consequently can't discover newly launched phishing web sites [3, 4].

Machine learning based phishes detection gadget relies upon efficiently on the aspects of accuracy. The most of anti-phishers researchers center of attention on optimizing new feature proposals or classification algorithms, where developing proper features analysis and selection techniques is not the important plan [5, 6]. In [5], The 12 features of this site are legitimate, phishing-enabled, reaching an effective positive rate of 97% and a false positive rate of 4%. The features are obtained by META tagging, web pages content, URLs, hyperlinks, TF-IDF, and more.

Therefore, extraneous aspects might also nonetheless exist, which will increase the price of the technology (i.e. Training time, storage, electricity, etc.), however, it does not affect the average accuracy.

Therefore, identifying a truly effective compact feature set requires an efficient Machine Learning based technique for Phishing detection. Figure 1 shows the phishing detection methods which are awareness of users and software detection. In software detection, there are many methods like blacklist, machine learning, and hybrid approaches.

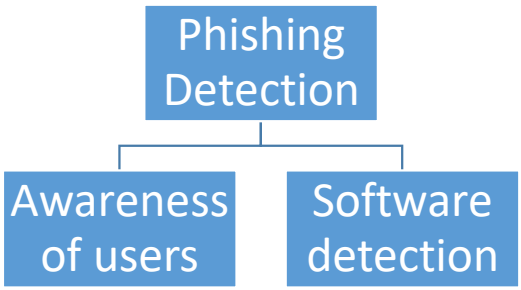

Fig. 1. Phishing detection methods

Generally, for selecting features the two main techniques were used: filter size and wrapper. Filter measurements, on the

978-1-7281-7407-5/20/$31.00 ©2020 IEEE
DOI 10.1109/SMART-TECH49988.2020.00026

different hand, are metrics. It is calculated from statistical and informative theories, which can mirror the deserves of every character function besides calling a precise classifier. Wrapper technique executes repeatedly, every execution entails producing a subset of elements and evaluating them via classification. When evaluating a set of characteristics, various iterations of engineering exponentially scale wrapper and will be of practical computing for real-time applications [7, 8].

In [9] described anti-phishing technology that removes 19 features on the buyer's side to determine phishing websites from approved sites using machine learning. They used 2,141 phishing pages from[10] and [11], as well as the famous Alexa website, some online debit gateways, and some great banking websites.

In [12] proposed to use a probability minimization standard and Monte Carlo algorithm using a new neural network-based classification technique for detecting phishing net pages. The thirty points were used to categorize the four main areas, especially around the bar-based, anomaly-based, HTML and JavaScript.

Some studies, such as [11], the authors the usage of a hybrid technique for photograph inspection as nicely as a laptop mastering approach. The most necessary disadvantage of photo / visual-based phishing detection is the want for prior information (web history) of the preliminary photograph database or net web page [13]. However, the proposed method no longer has this dependency. So far, the use of natural language processing (NLP) has been found in the literature.

Hybrids [14] extracts features towards URL, text, and web content and utilize Extreme Machine Learning (ELM) technology. First step in this method is to write a text content of the classifier to determine content of the label text through ELM. In this case, OCR software is used to retrieve the text from the image. It is a second stage-based hybrid that combines text and other function classifiers.

In [15] proposed an approach to constructing a stochastic neural network (PNN). The benefits of fast train PNN time, numbness to outliers, and generalization are optimal. However, PNNs can increase data significantly, requiring high space and time. As a result, the author's group K-medoids with PNNs to minimize training cases. In [16], proposed an anti-phishing method for the Iranian e-banking system. The author identifies 28 characteristics that attackers use to deceive Iranian banking sites. In the Iranian banking system, the detection accuracy was 88%. This method is specifically developed to discover Iranian bank sites and can only filter every types of phished and legitimated websites.

Machine learning based techniques, usually depend on comparisons between suspect site functions and a predetermined set of functions [17]. Consequently, the accuracy of the system depends on a set of functions and the accuracy with which the defender chooses the function [18, 19].

In current studies [20], NLP was implemented for phishing email detection. Detect malicious intent by performing semantic analysis on email content (plain text).

In this paper machine learning algorithms is used to detect URLs of web pages to detect phishing web pages. When implementing a machine learning algorithm, it is also important to extract features from the dataset. As a result, Google collects a lot of legitimate and fraudulent web page URLs from existing datasets. The efficiency of the proposed system is measured by the function defined by the word vector.

II. Proposed Technique

Figure 2 shows the overall steps for the proposed technique for phishing website detection using the machine learning technique.

A. Step 1

Automatically collect web pages using GNU Wget and Python scripts. In addition to the entire HTML document, we also download related resources (e.g. images, CSS, JavaScript) so that we can provide a browser to all the web pages that are downloaded. Also, screenshots of all web pages are saved for further inspection and filtering.

Download datasets that are processed further to load phishing and legitimate data sets or to eliminate web page failures that have failed on the "Error 404" page. A duplicate instance of the web page is also removed. After filtering the sample, the function is extracted. There are two main types of features used in traditional phishing site detection studies: a) internal functions; (b) External characteristics. The internal functions are derived from the URL of the webpage and the HTML source code, both of which can be accessed directly from the web page itself. On the other hand, external functions derived from requests from third-party services, such as domain registers, search engines, WHOIS records, etc., focus on benchmarking.

Registration at this point defines the URL provided in the proposed technology. To collect basic functionality, the website has been stacked into two separate classes between January and May 2015 and between May and June 2017. In particular, we selected 5000 phishing web pages, and all web pages are more stable, especially based on URLs. The fish tank is based entirely on the Alexa URL and Common Crawl archives.

B. Step 2

Vocabulary, host, and word used to extract feature vectors from the input URL. The vocabulary feature is a feature of text URLs such as hostname size, URL length, tokens observed in the URL, etc. A simple calculation, security, and precision for excessive classification of machine learning vocabulary features.

The vocabulary function is the URL text of the property itself, not the page itself. These properties include the length of the hostname, the length of the full URL, the number of dots in the URL, the hostname (separated by ") and the binary function (separated string) for each symbol In the URL path (. '/' ,,, '=', '?' " '-'East' _ '). This is also called "pocket".

Host-based features can describe "where malicious sites are hosted", "have" and "manage". The following are host attributes identified by the hostname as part of the URL.

Words in vectors are particularly desirable to perform certain important functions the URL of the web page. It mainly consists of a text containing many words. Instead of changing the text in this manual, the automatic vectorization process is better. Each URL is translated into carrier-specific words using a Weka function called "StringtoWordVector".

After obtaining the relevant vectors, you can easily use them in your chosen machine learning algorithm.

Features results extracted using various rules and used in the next step.

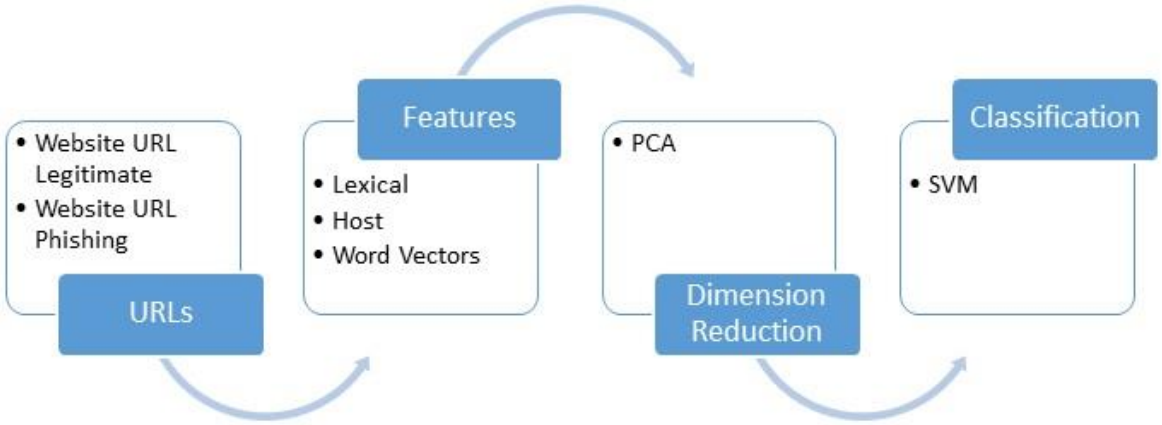

Fig. 2. Proposed technique

C. Step 3

To avoid high dimensionality, the Principal Component Analysis (PCA) [21] utilized for features. PCA purpose is to reduce the large variable set to a small set of variables whose information is maintained. It is a famous statistical method that tries to explain the covariance shape of facts utilizing a small range of components. These components are linear mixtures of the authentic variables and frequently permit interpretation and a higher grasp of the different sources of version.

D. Classification

In this phase, we use a classifier to get the final result. The classifier is just a machine learning algorithm trained to predict results and perform classification. Because there is no complete and precise single classifier. Classifiers are primarily chosen because they have been used for Google-like issues such as spam detection, phishing emails, phishing websites, and malicious URLs. The system simply tries to use this system for final prediction and classification activities. Support vector machines are used for classification.

A support vector machine (SVM) is often utilized in a phishing attack detection classifier. SVM works by examples of training and changes that have been set that makes maps of feature set to produce a feature room changed, save a sample of URLs from two classes with a hyperplane in the feature space changes.

III. Results and Discussion

The proposed machine learning based technique is compared with the previous technique. In our experiment, comparable classification method is used to train the split test. Each partition, 70% of the statistics used for training remain for testing purposes. Accuracy is calculated using Equation 1.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (1)$$

TP where positive means right, TN represents a true negative, implying FN false negative and false positive FP capabilities. According to this equation, we calculate the accuracy and the results for machine learning algorithms. The overall performance of the proposed technique is superior to other previous techniques. Table 1 shows the overall performance of the proposed technique along with other techniques. Therefore, the proposed machine learning technique reduces very effective factors. It is said that the overall performance of the proposed technique provides higher accuracy for classification algorithms. Also, the promising results shows that the proposed technique is effective and can be flexibly applied to different datasets.

TABLE I. Summary of Results

Classifier	**Feature set**	**No of Features**	**Accuracy**
FACA [2]	Full	30	90.44
Random Forest [22]	Full	30	94.27
Random Forest [22]	HEFS	5	93.22
SVM	**Proposed**	**5**	**95.66**

IV. CONCLUSION AND FUTURE WORK

This paper provides phishing detection based on machine learning technology. Also, classifiers generated by machine learning algorithms identify legitimate phishing websites. The proposed technique used SVM with an accuracy of 95.66% and a very low false-positive rate. The proposed technique can detect new temporary phishing sites and reduce the damage caused by phishing attacks. The performance of the proposed technique based on machine learning is more effective that previous phishing detection technologies.

In the future, it will be useful to investigate the impact of feature selection using various classification algorithms.

REFERENCES

[1] Higashino, M., et al. An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage. in 2019 5th International Conference on Information Management (ICIM). 2019.

[2] H. Bleau, Global Fraud and Cybercrime Forecast,. 2017.

[3] Michel Lange, V., et al., Planning and production of grammatical and lexical verbs in multi-word messages. PloS one, 2017. 12(11): p. e0186685-e0186685.

[4] Rahman, S.S.M.M., et al. Performance Assessment of Multiple Machine Learning Classifiers for Detecting the Phishing URLs. 2020. Singapore: Springer Singapore.

[5] He, M., et al., An efficient phishing webpage detector. Expert Systems with Applications, 2011. 38(10): p. 12018-12027.

[6] Mohammad, R.M., F. Thabtah, and L. McCluskey. An assessment of features related to phishing websites using an automated technique. in 2012 International Conference for Internet Technology and Secured Transactions. 2012.

[7] Abdelhamid, N., A. Ayesh, and F. Thabtah, Phishing detection based Associative Classification data mining. Expert Systems with Applications, 2014. 41(13): p. 5948-5959.

[8] Toolan, F. and J. Carthy. Feature selection for Spam and Phishing detection. in 2010 eCrime Researchers Summit. 2010.

[9] Jain, A.K. and B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems, 2018. 68(4): p. 687-700.

[10] 1PhishTank, Phishing dataset. 2018, Verified phishing URL.

[11] 1Openfish, Phishing dataset. 2018.

[12] 1Chiew, K.L., et al., Utilisation of website logo for phishing detection. Computers & Security, 2015. 54: p. 16-26.

[13] Benavides, E., et al. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. 2020. Singapore: Springer Singapore.

[14] Zhang, W., et al., Two-stage ELM for phishing Web pages detection using hybrid features. World Wide Web, 2017. 20(4): p. 797-813.

[15] El-Alfy, E.-S.M., Detection of phishing websites based on probabilistic neural networks and K-medoids clustering. The Computer Journal, 2017. 60(12): p. 1745-1759.

[16] Montazer, G.A. and S. ArabYarmohammadi, Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system. Applied Soft Computing, 2015. 35: p. 482-492.

[17] Wang, Y.-G., G. Zhu, and Y.-Q. Shi, Transportation spherical watermarking. IEEE Transactions on Image Processing, 2018. 27(4): p. 2063-2077.

[18] De Maio, C., et al., Time-aware adaptive tweets ranking through deep learning. Future Generation Computer Systems, 2019. 93: p. 924-932.

[19] De Maio, C., et al., Social media marketing through time ‐ aware collaborative filtering. Concurrency and Computation: Practice and Experience, 2018. 30(1): p. e4098.

[20] Peng, T., I. Harris, and Y. Sawa. Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. in 2018 IEEE 12th International Conference on Semantic Computing (ICSC). 2018.

[21] Abdi, H. and L.J. Williams, Principal component analysis. WIREs Computational Statistics, 2010. 2(4): p. 433-459.

[22] Sahingoz, O.K., et al., Machine learning based phishing detection from URLs. Expert Systems with Applications, 2019. 117: p. 345-357.