Peertechz

ENGINEERING GROUP

TRENDS IN
Computer Science and Information Technology
OPEN ACCESS JOURNAL

ISSN: 2641-3086 DOI: https://dx.doi.org/10.17352/tcsit

Research Article

Detection and classification of phishing websites

Manoj P*, Bhuvan Kumar Y, Rakshitha D and Megha G

Department of CSE, NIE Institute of Technology, Mysore, India

Received: 23 July, 2021
Accepted: 10 August, 2021
Published: 11 August, 2021

***Corresponding author:** Manoj P, Department of CSE, NIE Institute of Technology, Mysore, India,
E-mail: manojp6268@gmail.com

ORCID: https://orcid.org/0000-0001-7618-8757

Keywords: Phishing; Malicious sites; Machine learning; Semantics-based attack

https://www.peertechzpublications.com

Check for updates

Abstract

'Phishing sites' are some type of the internet security issues that mainly targets the human vulnerabilities compared to software vulnerabilities. Phishing sites are malicious websites that imitate as legitimate websites or web pages and aim to steal user's personal credentials like user id, password, and financial information. Spotting these phishing websites is typically a challenging task because phishing is mainly a semantics-based attack, that mainly focus on human vulnerabilities, not the network or software vulnerabilities. Phishing can be elaborated as the process of charming users in order to gain their personal credentials like user-id's and passwords. In this paper, we come up with an intelligent system that can spot the phishing sites. This intelligent system is based on a machine learning model. Our aim through this paper is to stalk a better performance classifier by examining the features of the phishing site and choose appropriate combination of systems for the training of the classifier.

Abbreviations

MLP: Multilayer Perceptron; ML: Machine Learning; SVM: Support Vector Machine; WEKA: Waikato Environment for Knowledge Analysis; ANN: Artificial Neural Networks; URL: Uniform Resource Locator

Introduction

Phishing is a non-ethical method comprising both social engineering and technical tricks to capture user's information and sensitive credentials like financial credentials. Some of the social engineering techniques use spam mails, pretending as a legitimate company or organization, that are specially designed to forefront users to knock-off websites that manoeuvre recipients to fall into the trap which steal financial credentials like user-ids and passwords. Technical intrigue methods install malicious software onto the systems, to capture the data directly, often using systems to intercept users online account user names and passwords [1].

A. Technique of phishing

Generally, there are two approaches that are typically used in detecting phishing websites. The first approach is typically based on a blacklist, where in the given URL is compared with the URLs present in the blacklist. The other part of this approach is that the blacklist usually cannot identify all phishing sites, hence a new fraudulent website is launched. The alternate or the second approach is referred to as heuristicbased methods, where few of the features are collected from the sites to distinguish it as either phishing or legitimate.

Compared to the blacklist approach, a heuristic-based solution can detect recently created phishing sites. The accuracy of the heuristic-based approach rely on selecting a set of selective features that might help in classifying the type of given website. Data mining techniques are some of the research fields which can utilize the features knowledge that promises the nature, reliability and completeness, also reduce the time of knowledge achievement. Basically, there are two types of rules-induction techniques in data-mining: associative technique and classification-rule technique. The usage of classification rules is of the concern in this paper. The classification task's aim is to assign every test data to one of the predefined classes in the test dataset. Various studies have been conducted regarding phishing website detection depending on the website features but these researches were unable to detect the exact or precise rules to classify the nature of website Table 1, Figures 1,2.

B. Phishing attack figures

Phishing pursues to be the fast-growing zones of identity thefts on the internet which cause both short-term and long-term economic rupture. There was nearly 33,000 phishing

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040

Table 1: Most Infected countries.

Ranking	Country	Infection Rate (%)
1	China	47.09
2	Turkey	42.88
3	Taiwan	38.98
4	Guatemala	38.56
5	Ecuador	36.54
6	Russia	36.02
7	Peru	35.75
8	Mexico	35.13
9	Venezuela	34.77
10	Brazil	33.13

attacks that took place globally every month in the year 2012, mount up to a loss of $687 million [1]. The example of phishing took place in June 2004. The Royal Bank of Canada alerted customers about those spam e-mails claiming to originate from the Royal Bank itself that were sent to the customers asking them to verify account numbers and personal identification numbers (PINs) via link that was added in the e-mail. The spam e-mail exclaimed that if the user did not click on the link and spill in his client card number and pass code, access to his/her account would be banned. These spam e-mails were sent to customers within a week of a computer go wrong that stopped customer accounts from being updated [2]. Financial Services remains to be the most targeted industry sector by Phishers [1].

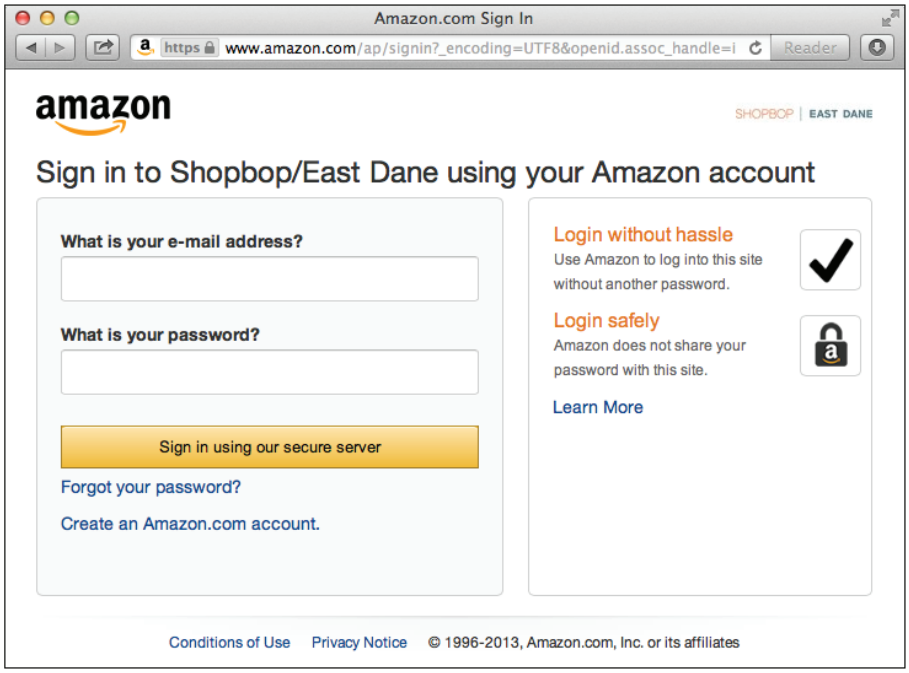

Figure 1: Original Amazon sign in page.

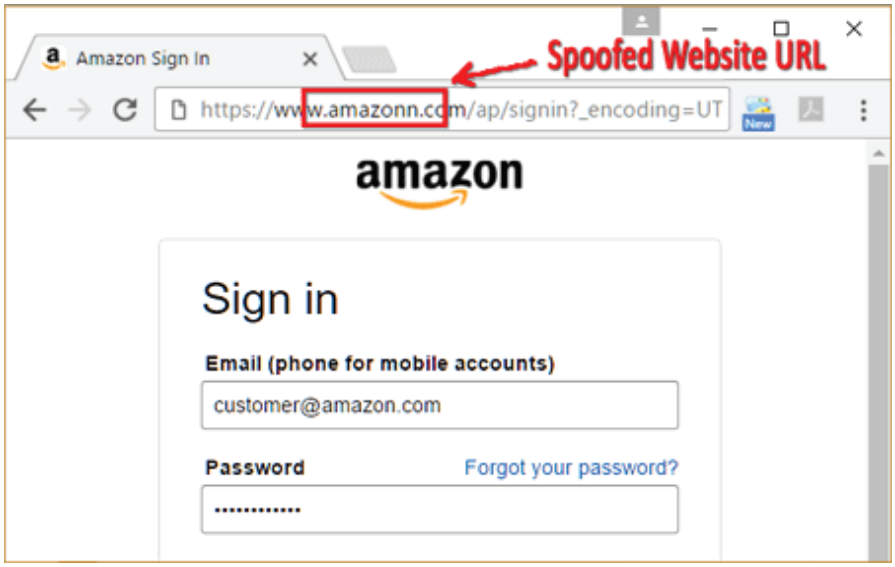

Figure 2: Phishing webpage.

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040

Related work

Couple of researchers have analysed the stats of malicious sites in some way. Our method picks up some of the important ideas from previous case studies. Ma, et al. [3,4] compared various batch-based learning algorithms used in classifying phishing sites and stated that a combination of host based and lexical-based features outcome in the highest accuracy in classification. Besides, they are also compared with the performance of batch-based algorithms with the online-based algorithms which when utilizes complete features and noticed that online-based algorithms, especially Confidence-Weighted (CW), stand out performing batch-based algorithms. The attributes include the existence of the red flag keywords present in the website, attributes that are based on Google's Page Rank and Google's Web page quality guidelines. One cannot compare directly without access to the same websites and attributes.

Problem statement

Internet has dominated the world by dragging half of the world's population exponentially into the cyber world. With the booming of internet transactions, cybercrimes rapidly increased and with anonymity presented by the internet, Hackers attempt to trap the end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among all these attacks, phishing reports to be the most deceiving attack. Our main aim of this paper is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.

Motivation

Detection and prevention of phishing websites endure measure continuously a major space for analysis. There are different types of phishing techniques that offer torrential and essential ways that offer attackers to penetrate the data of people and organizations. Uniform resource locator URLs sometimes are also referred to as "Weblinks" play a vital role in a phishing attack. Uniform resource locator has a vulnerability of redirecting the pages i.e., through the hyperlink; which could redirect to the legitimate website or the phishing site. Different techniques in making phishing sites are emerging day by day. This actually motivated several researchers to put up their concentrate on finding the phishing sites.

Data flow

The technique comprises of host based, page based and lexical feature extraction of collected websites. The primary step is the collection of phishing and benign websites. In the host-based approach, admiration based and lexical based attributes extractions are performed to form a database of attribute value. This database consists of knowledge mined that uses different machine learning techniques. On evaluating the algorithms, a selective classifier is opted and is implemented in Python. The data flow is shown in Figure 3.

A. URL collection

We collected URLs of benign websites from www.alexa.com [5-9] www.dmoz.org [7] and personal web browser history. The phishing URLs were collected from www.phishtak.com [8]. The data set consists of 17000 phishing URLs and 20000 benign URLs. We obtained PageRank [10] of 240 benign websites and 240 phishing websites by checking PageRank individually at PR Checker [11]. We collected WHOIS [12] information of 240 benign websites and 240 phishing websites.

B. Host-based analysis

Host-based features explain "where" phishing sites are hosted, "who" they are managed by, and "how" they are administered. We use these features because phishing Web sites may be hosted in less reputable hosting centres, on machines that are not usual Web hosts, or through not so reputable registrars.

Below are the characteristics of the host-based that are notified.

i. **WHOIS properties:** WHOIS [12] properties give information regarding the registrations, updates and expiry, differentiating the admin and the user. Phishing URLs are taken down repeatedly, the date of registration will be recent compared to legitimate sites. Majority of phishing URLs contain IP address in their hostname [5].

ii. **Geographic properties:** Geographic properties provides the information regarding the continent/state/country to which the corresponding IP address belongs to. Analyse attributes using machine learning techniques. Selection of a Classifier Implement the classifier stores phishing & Benign websites host-based and page-based attribute Lexical feature extraction.

C. Lexical feature

Lexical features are the textual characteristics of the URLs themselves and are not the contents of the page to which it points. Uniform Resource Locators are humanreadable character strings that are tokenised in a standard manner by client programs. Via multistep resolution process, the browsers translate every URL into set of instructions which point to the server that is hosting the website and specify where the website is present in that host. To make easier this translation process, Uniform Resource Locators consists the standard syntax. :// one such example of Uniform Resource Locators resolution is shown: The module of the Uniform Resource Locators signifies which network protocol to be used in order to fetch the requested resource. The common protocols that are used are Hypertext Transport Protocol or HTTP (http). HTTP with Transport Layer Security (https), and File Transfer Protocol (ftp). Hackers sometimes hide path tokens to avoid inspection, or they may intentionally create tokens to imitate the appearance of a legitimate website. The flowchart of feature extraction is shown in Figure 4.

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040

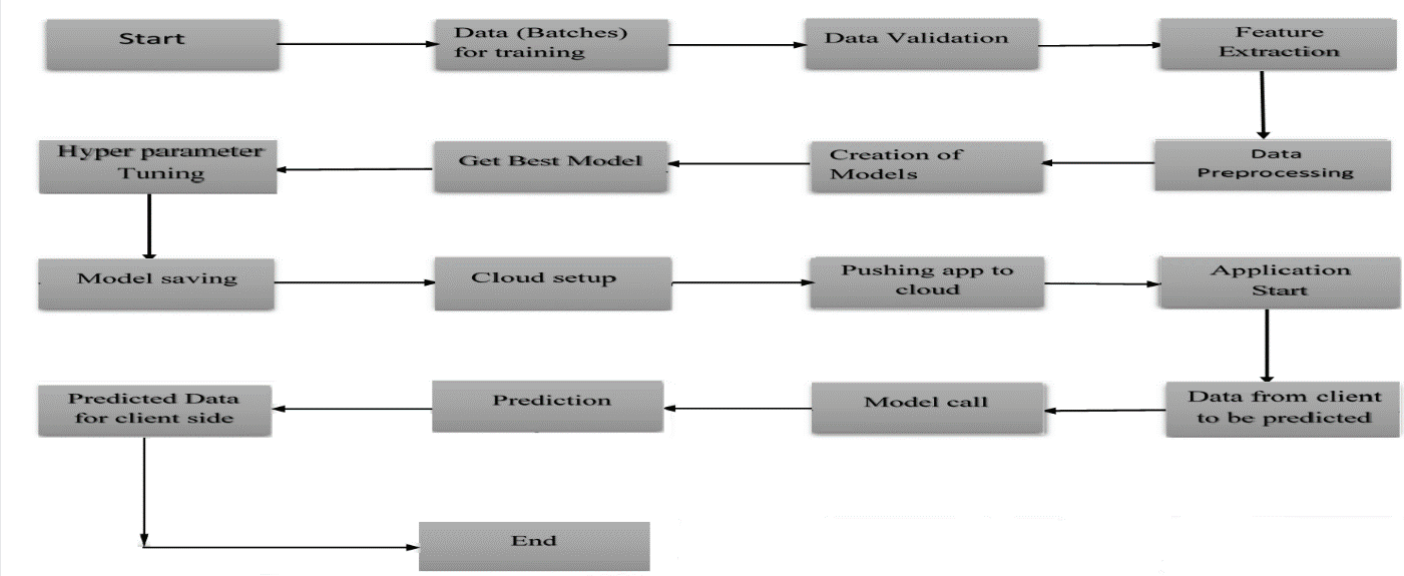

Figure 3: Data Flow Diagram.

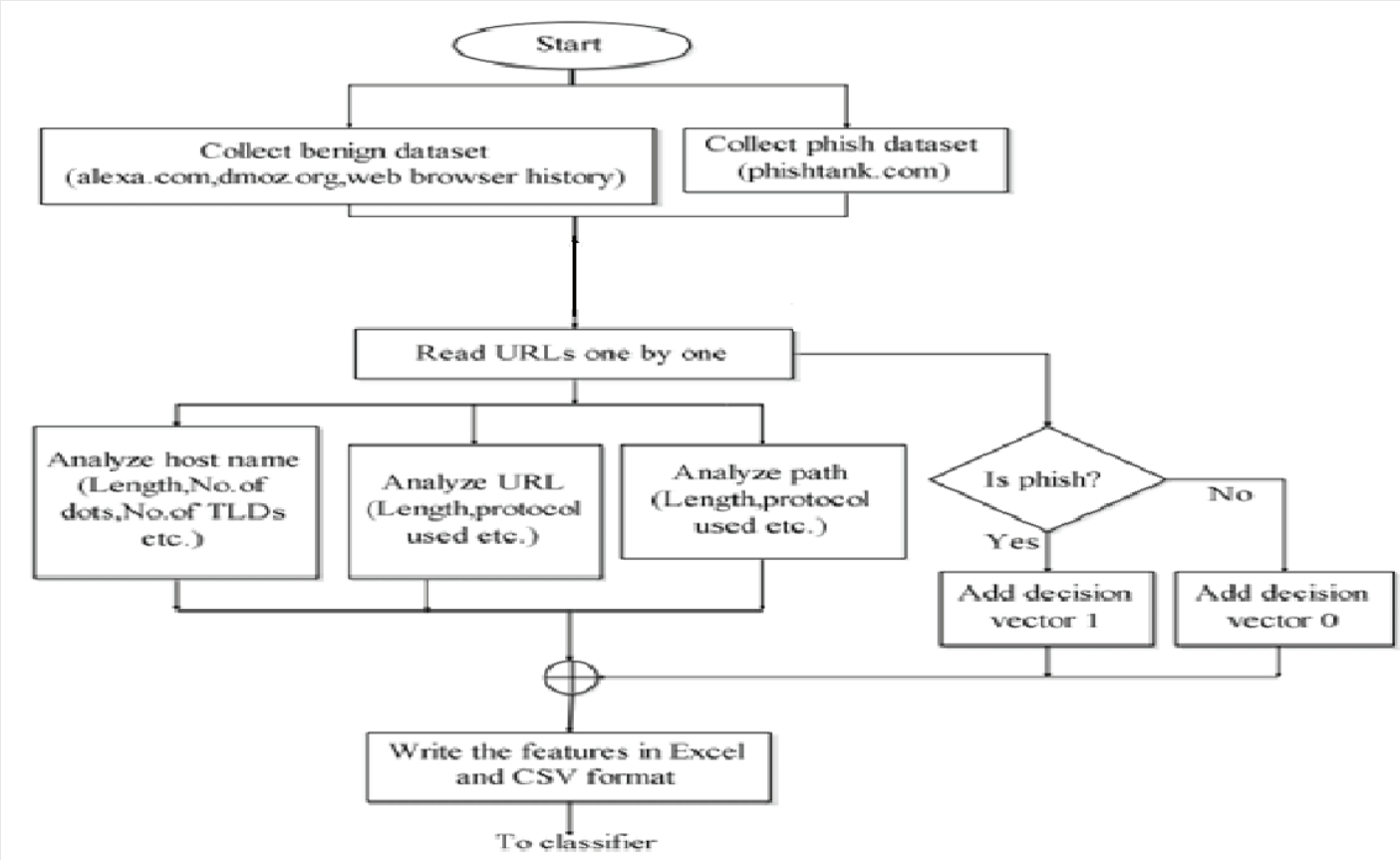

Figure 4: Feature Extraction Process.

D. Machine learning algorithms

The evaluation of the various classifying algorithm is done by using the workbench for data mining, Waikato Environment for Knowledge Analysis (WEKA) [13-16]. Four types of input data files i.e., Attribute Relation File Format (.arff), Comma Separated Values (.csv). In our experiment .csv file format was used. The input file to the Waikato Environment for Knowledge Analysis was obtained by program by appending 'YES' in place of decision vector '1' (phish) and 'NO' in place of decision vector '0' (benign) of the dataset generated from input URL list. The dataset was made split into 70% for training and remaining 30% for testing purpose.

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040

The five machine learning algorithms considered for processing the feature set are:

1) **Logistic regression:** It is a statistical model that uses a logistic function to build a dependent variable, which can also have many more complex extensions.

2) **SVM:** The Support Vector Machine performs a classification task by finding the 'hyper plane' which maximizes the margin between two groups of classes. The vectors that signify the hyper plane are known as the support vectors.

3) **XGBoost:** Boosting is a machine learning technique used in regression, classification and other tasks, that predicts a model in the form of an ensemble prediction models, favourably decision trees.

4) **MLP:** A Multilayer Perceptron (MLP) is a class neural network typically Artificial Neural Networks (ANN). Intuitively known as Perceptron of multiple layers.

5) **AutoEncoders:** An autoencoder is a type of neural network typically Artificial Neural Networks (ANN) that is used to learn the patterns of the unlabelled data (unsupervised learning). The Figure 5 shows the loss function of the autoencoders model. Binary crossentropy is the loss function that is defined on the training data and in Figure 5, the blue line represents the training loss and the orange line represents the accuracy of AutoEncoders model Figure 6.

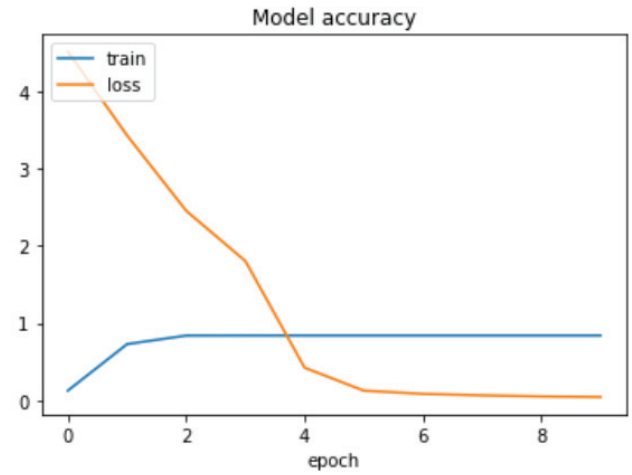

Figure 5: Loss function for AutoEncoders.

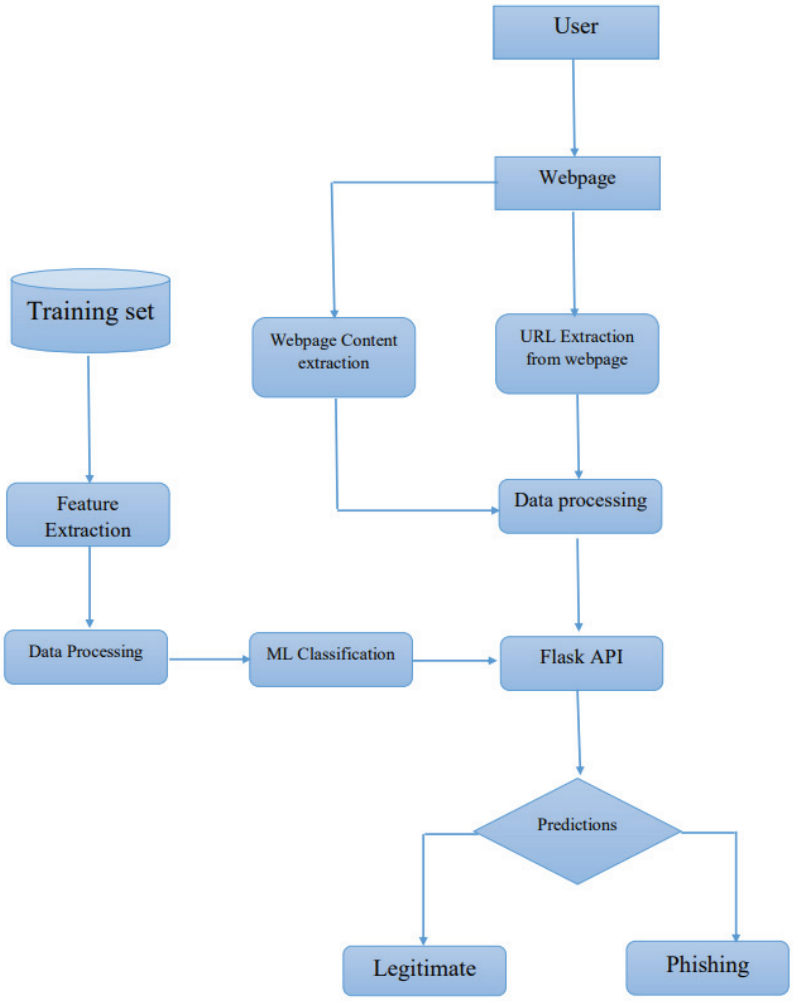

Figure 6: Flow of the classifiers program.

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059.
DOI: https://dx.doi.org/10.17352/tcsit.000040

Results

The key notable points of our initial work embed:

Phishing sites and their domains reveal the features that are different from other sites and domains. (For example, Google; www.google.com and some random phishing website be like; www.googlee.com).

Phishing Uniform Resource Locators and 'domain names' typically have a different length when compared to other websites and domain names.

The above Table 2 provides the training accuracy and testing accuracy of all the models. The difference between the values of train and test accuracy shows that the models are not overfitting over large dataset Figures 5-9.

Table 2: Classifier performance.

ML Model	Train Accuracy	Test Accuracy
Logistic Regression	0.833a	0.824
Support vector machine	0.854	0.85
Multilayer perceptrons	0.862	0.858
AutoEncoder	0.837	0.846
XGBoost	0.851	0.849

Figure 7: A Simple UI to fill the URL of the site.

Figure 8: Output of a Legitimate URL.

Figure 9: Output of a Phishing URL.

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040

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.

References

1. Microsoft. Microsoft Security Index Report.
2. Yu WD, Nargundkar S, Tiruthani N (2008) A phishing vulnerability analysis of web-based systems. Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008). Marrakech, Morocco: IEEE 326- 331. Link: https://bit.ly/2VJhDer
3. Sheng S, Holbrook M, Kumaraguru P, Cranor LF, Downs J (2010) 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 373–382. Link: https://bit.ly/2VL0NeA
4. Sheng S, Wardman B, Warner G, Cranor LF, Hong J, et al. (2009) An empirical analysis of phishing blacklists. In Proceedings of the 6th Conference in Email and Anti-Spam, ser. CEAS'09, Mountain view, CA. Link: https://bit.ly/3Az9TdT
5. Khonji M, Iraqi Y, Jones A (2013) Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials 15: 2091-2121. Link: https://bit.ly/3CCcfKz
6. Google (2017) Google safe browsing API.
7. Prakash P, Kumar M, Kompella RR, Gupta M (2010) Phishnet: predictive blacklisting to detect phishing attacks. In INFOCOM'10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press 346–350. Link: https://bit.ly/3jJkrA2
8. Cao Y, Han W, Le Y (2008) Anti-phishing based on automated individual whitelist. In DIM '08: Proceedings of the 4th ACM workshop on Digital identity management. New York, NY, USA: ACM 51–60. Link: https://bit.ly/3fPGnsi
9. Rbldnsd. Link: https://bit.ly/3s5UFtI
10. PhishTank. Link: https://bit.ly/37vGa8Z
11. TechHelpList. Link: https://bit.ly/3jz4Cfa
12. Alexa. Link: https://bit.ly/3jIDn1U
13. Cymon. Link: https://bit.ly/3jI2eDa
14. All Cybercrime IP Feeds by Firehol. Link: https://bit.ly/3Cz3TTX
15. Volkamer M, Renaud K, Reinheimer B, Kunz A (2017) User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. Computers & Security. Link: https://bit.ly/37x2mQ4
16. Anti-Phishing Working Group (APWG) (2016) Phishing activity trends report — last quarter 2016. Link: https://bit.ly/3lNULVH

Citation: Manoj P, Bhuvan Kumar Y, Rakshitha D, Megha G (2021) Detection and classification of phishing websites. Trends Comput Sci Inf Technol 6(2): 053-059. DOI: https://dx.doi.org/10.17352/tcsit.000040