International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT19522

# Detecting URL Phishing Attacks Using Machine Learning & NLP Techniques

**Anitha R[1], Swathi S[2], Vasuhi R[2], Thenmozhi P[2]**

[1]Assistant Professor, Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore Tamil Nadu, India

[2]B.E Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore Tamil Nadu, India

## ABSTRACT

The internet plays a huge role in day to day life of people.URL Phishing is done by hackers who aim to obtain information from the internet user by making the url seem like, it is from a trustworthy organization. It is necessary to be aware of such phishing attack that takes place online in order to safeguard sensitive information from being stolen. This paper focuses on detecting URL phishing. URL features are extracted from the url that is to be tested using various NLP (Natural Language Processing) techniques such as tokenizing, finding popularity, checking the presence of IP (Internet Protocol) address, etc. This system uses machine learning algorithm (Support Vector Machine) which can be used for classification challenges. SVM (Support Vector Machine) is used to identify the phishing or safe status of the given URL. A dataset containing url features is used to train the SVM algorithm to do so. SVM is the best algorithm in classification (based on the features of given data) which gives reliable results.

**Keywords :** Phishing, Phishing detection system, Hacking, SVM, URL, NLP

## I. INTRODUCTION

Due to the rapid increase of internet users the attackers focus on new technology such as url phishing instead of bank/shop robbery. Victims who use the false website are made to disclose personal information like account IDs, usernames, passwords, etc. Criminals demand and receive money through online banking, online payment systems-commerce (electronic commerce) websites and m-commerce (mobile commerce) applications. However, detection of phishing attack is much more complex so we have to go for computerized detection system like machine learning approach.

Machine learning algorithms build a mathematical model by training it with appropriate data in order to make predictions without being explicitly programmed. Machine learning algorithms are used in various applications like email filtering, detection of network intruders and computer vision. Machine learning is related to computational statistics, which focuses on making predictions The study of mathematical optimization delivers theory, methods and application domains to the field of machine learning.

Attackers create and use websites that seem like original websites in order to get the personal information of the user without their knowledge and the attackers can view, update or change any information from the user's account.

## II. LITERATURE SURVEY

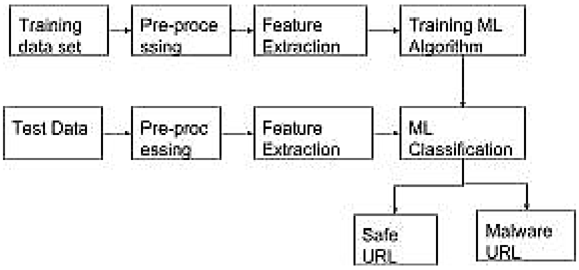
In the literature there are three main ways for handling phishing. One of the approaches to encounter phishing is to blacklist, that contains known phishing websites acquired by techniques such as user votes, those blacklists are included as plug-in code in search engines in order to check whether the url is present in the blacklist in order to find if the url is phished and it prevents the user whenever he attempts a connection to one of these malicious websites. Some examples are internet explorer phishing filter [1], google safe browsing for Firefox [2]. This approach has an issue since it offers no protection against the new phishing websites that are not included in the blacklist and the processing of the blacklist is slow and consumes a lot of time. Researchers have started using artificial intelligence and data mining to detect the phishing websites. This is the path that gives reliable results far more and wherein this project work falls. Researches such as CANTINA [3], the work of Xiang et al detects url phishing based on features and query results through search engines in addition to some elements of HTML (hypertext markup language) pages and it has a recognition rate of 92%. Moreover Fu et al. [4] has proposed a detection system based on the visual similarity of web pages calculated by earth mover's distance. [5] Hybrid system that used the image is PSO-SVM (particle swarm optimization support vector machine) to achieve a recognition rate of 99%. This system uses the two different DNS (domain name system) server and compares them, but an attacker can tamper with the results of the two DNS servers using a man in the middle attack and which ruins the recognition of the system. Thomas et al. [6] developed a real-time spam and phishing detection system, their system uses several criteria such as the characteristics of the URL, the number of redirects, web pages HTML elements and JavaScript, geo-location data, and DNS (domain name system) data and this technique will be impossible if the attacker blocks the IP (internet protocol) of their crawler from collecting the needed data. The work of Jeeva et al. [7] uses a white list called repository to check if the URL exists in the repository. If it is not present then, it is recognized using association rule mining algorithm. Finally, the research of Ramesh et al. [8] which reached an impressive recognition rate of 99.62% use suspicious web page keywords as an input to a search engine to get links, and then compare them with the links within the suspicious web page to keep only the existing links as an input to TID algorithm (target identification algorithm) and finally a DNS lookup is performed to check the domain name of the website with and checks if it lies in the weak link which makes it prone to the man in the middle attack.

The other solutions proposed to address phishing issue in the works do not identify legitimate and phishing websites. Instead, they aim to consolidate user authentication in order to overcome this problem. The study by Huáng et al. [9] proposes to replace the use of a permanent password by a (one-time password) that should be provided to the user by a third party under a message form. It completely depends on the third party which serves as a threat to security.

## III. PROPOSED SYSTEM

NLP (natural language processing) is a route for computers to investigate, comprehend, and get importance from human language in a shrewd and helpful way.NLP techniques are applied to parse each sentence and identify the semantic roles of important words in the sentence in relation to the predicate. The proposed system uses URL features to perform the recognition. The features are URL length, largest token, largest path, domain token count, number of dots, number of tokens, average token length, presence of IP address, etc. Various NLP (Natural Language Processing) techniques such as tokenizing, finding popularity, checking ip address ,etc. help in extracting the features. The SVM (support vector machine) algorithm is trained with the features extracted using NLP techniques and it classifies the safe and phished URL. Support vector machine is implemented with the help of Scikit-learn python library and it helps in classifying the url by finding a hyperplane that divides the two categories.

## IV. WORKING METHODOLOGY



The process is divided into training and testing. The training data contains a list of URLs in form of csv file with a value at end (either zero or one) to denote either it is malware or benign. Dataset containing features like URL length, largest token, largest path, domain token count, number of dots, number of tokens, average token length, presence of IP address etc is obtained by extracting these features using NLP techniques on csv file. The url that is to be tested is preprocessed and features are extracted similar to that of the training data. Both the training and testing data features are given as input to the SVM algorithm is used to train the SVM model which performs classification.

Support Vector Machine (SVM) is a discriminative classifier that creates a separating hyperplane. In different words, given labeled work information (supervised learning), the rule outputs the best hyperplane that categorizes new data based on trained patterns. Hyperplane might be a line dividing a plane in a pair of elements wherein each class lay in either facet.

SVM has several kernels to perform classification. Linear Kernel is used for classification in the proposed system because it involves classification between two categories of urls.

The training data set is used by svm model to learn and classify other data based on the knowledge acquired from training data patterns. When there is an instance of an SVM classifier, a training dataset and a test dataset, the model is ready to be trained using the fit() function in SVC(support vector classification) class present in SVM library contained in Skicit-Learn.

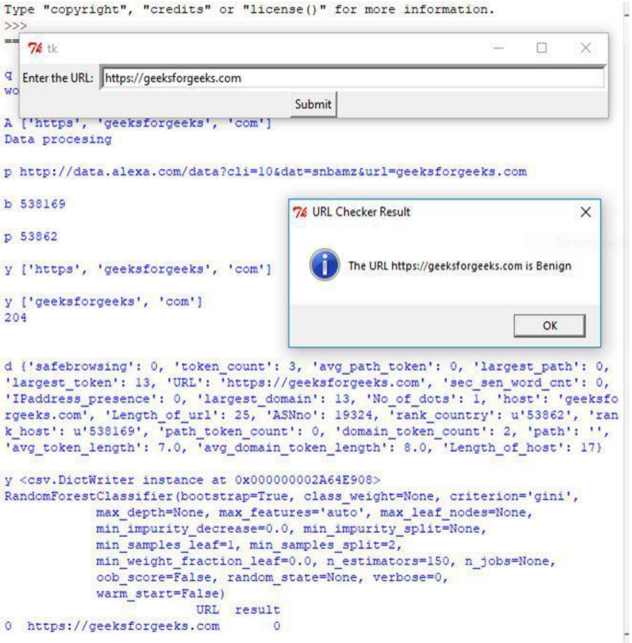
```
>>>svmClassifier.fit(x_train,y_train)
```

The function creates a working model to detect whether a url is safe or phished based on the features extracted from the train data and test url.

Packages to be imported from the python library:

Pandas , Trainer, Tkinter, tkMessageBox, urllib2, re, Pygeoip, minidom matplotlib, numpy

## V. EXPERIMENTAL RESULT



```
Type "copyright", "credits" or "license()" for more information.
>>>
==
q
wo
A ['https', 'geeksforgeeks', 'com']
Data procesing

p http://data.alexa.com/data?cli=10&dat=snbamz&url=geeksforgeeks.com

b 538169

p 53862

y ['https', 'geeksforgeeks', 'com']

y ['geeksforgeeks', 'com']
204

d {'safebrowsing': 0, 'token_count': 3, 'avg_path_token': 0, 'largest_path': 0,
'largest_token': 13, 'URL': 'https://geeksforgeeks.com', 'sec_sen_word_cnt': 0,
'IPaddress_presence': 0, 'largest_domain': 13, 'No_of_dots': 1, 'host': 'geeksfo
rgeeks.com', 'Length_of_url': 25, 'ASNno': 19324, 'rank_country': u'53862', 'ran
k_host': u'538169', 'path_token_count': 0, 'domain_token_count': 2, 'path': '',
'avg_token_length': 7.0, 'avg_domain_token_length': 8.0, 'Length_of_host': 17}

y <csv.DictWriter instance at 0x000000002A64E908>
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=150, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
                        URL  result
0  https://geeksforgeeks.com      0
```

tk

Enter the URL: https://geeksforgeeks.com

URL Checker Result

The URL https://geeksforgeeks.com is Benign

The above image shows the result of test url preprocessing, features extraction of the test url entered in the text box and the pop up window displays if the url is malware or safe.

## VI. CONCLUSION

In modern internet world, security of private information is nightmare to every person. There are so many illegal ways to steal internet user's information. One such threats is phishing. Phishing attacks are one of the most common and least defended security threats. It can be very dangerous and can cause a huge amount of loss. In some cases, the phisher may resell the illicitly obtained sensitive information to a secondary market. To make the detection more accurate Natural Language Processing & Machine Learning Technique (SVM algorithm) is used. In this paper, the structure of URL is identified and features are extracted using NLP techniques and svm is trained to perform classification to detect phishing. The experimental results show the solution is effective to detect URL phishing and can be used as plug-in in browsers to filter the phishing sites.

## VII. REFERENCES

[1]. "Microsoft (2005) Anti-phishing white paper" http://www-pc.uni-regensburg.de/systemsw/ie70/Anti-phishing_White_Paper.doc.

[2]. Schneider F, Provos N, Moll R, Chew M, Rakowski B (2007) "Phishing protection design documentation". https://wiki.mozilla.org/Phishing_Protection:_Design_Docu mentation.

[3]. Xiang G, Hong J, Rose CP, Cranor L (2011) CANTINA+: "A feature-rich machine learning framework for detecting phishing web sites". ACM Trans Inf Syst Secur 14(2):21. doi:10.1145/2019599.2019606

[4]. Fu AY, Wenyin L, Deng X (2006) "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)". IEEE Trans Dependable Secure Computing 3(4):301–311. doi:10.1109/TDSC.2006.50

[5]. Li Y, Chu S, Xiao R (2015) "A pharming attack hybrid detection model based on IP addresses and web content". Optik- Int J Light Electron Optics 126(2):234–239. doi:10.1016/j.ijleo.2014.10.001

[6]. Thomas K, Grier C, Ma J, Paxson V, Song D (2011) "Design and evaluation of a real-time URL spam filtering service". In: proceedings of the thirty second IEEE conference on security & privacy, California, 22–25 May 2011, p. 447–462

[7]. Jeeva SC, Rajsingh EB (2016) "Intelligent phishing universal resource locator detection victimisation association rule mining". Human-centric Comput Inf Sci 6:10. doi:10.1186/s13673-016-0064-3.

[8]. Ramesh G, Krishnamurthi I, Kumar KSS (2014) "An efficacious technique for police investigation phishing webpages through target domain identification". Decis Support Syst 61:12–22. doi:10.1016/j.dss.2014.01.002.

[9]. Huang C-Y, Ma S-P, Chen K-T (2011) "Using one-time passwords to prevent password phishing attacks". J Netw Comput Appl 34(4):1292–1301.

**Cite this article as :**

Anitha R, Swathi S, Vasuhi R, Thenmozhi P, "Detecting URL Phishing Attacks Using Machine Learning & NLP Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 53-56, March-April 2019.
Journal URL : http://ijsrcseit.com/CSEIT19522