Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

Identifying Phishing Websites Through URL Parsing

Anitha A, Kavya Sri Gudivada,
Rakshitha Lakshmi M.K, Shikha Kumari
Dept. of CSE, BNMIT
Bengaluru, India

Usha C R C
Assistant Professor, Dept. of CSE, BNMIT
Bengaluru, India

Abstract*—Identifying the acts of phishing in the websites is mostly done using the classification techniques based on machine learning or data mining. Many algorithms have been used for this purpose. The algorithms can be used to recognize the phishing websites. The performance factors like accuracy and speed can be used to compare different algorithms. Algorithms related to classification were used in the proposed model. There can be no perfect system to detect all the phishing websites, but using these methods it will create a better and more effective way for phishing website identification. The aim is to detect and classify the websites into phishing or legitimate. Thus it improves security for the users in online platforms. Four algorithms were compared for their performance with the same data set. The URLs are scanned for various features to know if they are phishing URLs or not. *(Abstract)

Keywords— phishing, legitimate, feature extraction, machine learning algorithms.

I. INTRODUCTION

The societal online attack is a very quotidian threat that is used to reveal personal and sensitive information of the users by deceiving them. The ultimate motive of the social attack is to acquire subtle information such as credit card details, user names, and passwords by masquerading as a virtuous entity in an automatic technology. Phishing appears in many forms such as messaging SMS, Voice over Internet Protocol (VoIP), fraudster email, etc. Communications from social and auction sites, banking, online transaction processing are frequently used to entice the unsuspecting public. Phishing fraudster emails contain links to sites that are contaminated with the virus. It is commonly carried out by email spoofing or instant messaging and often directs users to enter all their sensitive and personal details in a sham website whose appearance is nearly identical to the legitimate site. Usually, users have multiple accounts on different websites such as email, banking, social networking. Hence, the upright user is frequently vulnerable target for such an attack. Most of the innocent users do not have an idea about how to secure their sensitive information which helps to make the attack victorious.

Phishing attack utilizes the social engineering techniques to entice the victim by sending a spoofed link by taking the innocent user to a fake web page. The spoofed link is present on the frequently accessed web page or sent through email to the victim. The fake web page appears exactly similar to the legitimate web page which the victim cannot recognize. Therefore when the users clicks on the respective required link, automatically they will be redirected to the fake web page which the users are not aware of. Hence the victim will be directed to an attacker server rather than to the real web server.

The existing solutions of antivirus, firewall the implementation of Secure Socket Layer (SSL) and the digital certificate does not provide full protection or security against the phishing attack and also does not prevent the web spoofing attack. The user is diverted to a fake web server during a web spoofing attack and in fact, a few types of SSL and CA can be completely forged while the appearance looks the same as the legitimate one.

The study develops an anti-web spoofing solution based on inspecting the Uniform Resource Locators (URLs) of fake web pages. The solution developed a series of steps to check the particular characteristics of various websites URLs. Phishing web pages URLs look different from the URLs of legitimate web pages due to particular unique characteristics. Therefore, the URL is used to determine resource location in data communication.

II. LITERATURE SURVEY

This section reviews the various approaches in the field of web content filtering focusing on security against the phishing attacks that occur on websites. Although there are many varied implementations each of them have their own merits and demerits.

According to Muhammet Baykara et.al the approach of using blacklisting technique for URL addresses and spam keywords along with the “anti phishing simulator” is used to ensure the detection of phishing in email platform. The methodology used is mainly Bayesian network classifiers. They have a large database containing blacklisted URLs and spam keywords. Avid programmers and use the “URL add feature” to add some phishing URLs and also similarly there is an “add spam feature” that involves addition of spam words to database, through these techniques the model’s accuracy can be improved overtime [2].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

Jian Mao et.al mainly deals with identification of phishing on the web through analysis of the CSS properties of the web pages. As the visual characteristics of both the legitimate and phishing websites need to be the same in order to deceive the end users, the visual aspects of web page such as HTML structure of web page, CSS properties and the overall visual appearance is considered. The presence of a great resemblance between any two pages is used to identify possible phishing attacks. CSS Extractor, DOM Extractor and Visual Characteristics Filter are used to accurately quantify the visual similarity of each page element in the web page [3].

Jakov Andric et.al mainly researches about the people's capacity to identify the various phishing attacks in the internet. The students of Croatia are taken as test subjects and each have been presented with phishing attacks of various forms side by side with the legitimate web content, the students were asked to identify the phishing attempt. Spam content in web pages, email spam, phishing websites, malicious content and as on were used to test the students in their ability protect themselves from various attacks in the internet. They had examined the students familiarity with threats in the form of phishing attacks conducted via the Internet and concluded that 59 % successfully were able to identify spam content were as 21% were incapable of doing so and remaining were undecided [4].

Tajinder Singh et.al focuses on the feature oriented approach to detect spam content using fuzzy classifier. Their aim was to develop a hybrid spam detection based on fuzzy logic. This approach was advantageous as Fuzzy logic provides a spam classifier to adapt to variety of context without compromising much on accuracy. The authors had presented a model for spam detection based on fuzzy classification using three main features they are -syntactic analysis (keywords), semantic analysis of the text and behavioral feature of spammed link (Link Redirection). Results indicates its advantages over two popular classifiers such as Naive Bayes and JRIP rule based approaches [5].

Tiago A. Almeida et.al compares the current popular algorithms used for web spam detection and concludes the MDL class algorithm is the most appropriate approach to detect spam content in real time scenarios. They explain that although the results obtained from other algorithms such as KNN, SVM may be superior to MDL, MDL has several benefits to offer such as being faster to train the model, being lightweight and inherently offer multi - class learning. The computational complexity is far better than other popularly used algorithms for this application [6].

Priyanka Salunkhe et.al explores the online social networking sites where people can post offensive messages on the walls and create problems. It Focuses on analyzing the methodology of Information Filtering on social networks. They have ventured into the different forms of filtration techniques such as content based filtration technique where in the textual content is parsed through and determines the presence of unwanted and unpleasant content. Policy based filtration where the model takes the users opinion on what to view and what not to view based on questionnaire. Also there is collaborative filtration where the model learns based on user's activities and behavior on the web and modifies the viewable content suitably [7].

Shuhua Liu et.al mainly concentrated on web content classification. They have explored the textual information of a web page and applied various techniques such as word weighting, text summarization and sentiment analysis techniques to extract topic features, content similarity features and sentiment indicators of web pages to build classifiers. They have incorporate semantics (in the form of concept terms) in determining the topic representation of the web categories Violence and Hate. In addition, they take into account the outgoing link information in their approach [8].

III. EXISTING SYSTEM

Currently websites on the internet contain all kinds of spam content such as malicious content, advertisements, phishing attacks through redirection and links. There is an impending need for a phishing detection mechanism that works efficiently for real time scenarios. Existing system for the identifying phishing includes various approaches such as content based filtering method where in the textual content of the web pages are parsed through to identify certain blacklisted keywords and meta tag content. Semantic analysis of the web pages is also performed by grouping of the related words in equivalent classes and performing analysis on it. Using a feature extraction based model to detect anomalies in different segments of the web page. There are also approaches where in the visual aspects i.e. CSS properties of the web pages are analyzed for similarity in order to detect phishing.

According to Abdulghani Ali Ahmed[1] the content based filtering method was implemented to combat against phishing attempts. There were several features that were extracted from the URLs and domain names are checked using several criteria such as IP Address, long URL addresses, adding a prefix or suffix i.e presence of "-" in domain name, redirecting to another site using the symbol "//", and URLs having the presence of symbol "@". The above features are considered and the outcome of the rules are given to a PhishChecker algorithm that will conclude the presence of phishing attacks. Although the accuracy was high, as only a few features are considered it is not effective in identifying the phishing attacks in current real world scenarios where websites are implemented using latest technologies and new approaches.

IV. METHODOLOGY

The raw data set consists URLs (Uniform Resource Locators) collected from a open source repository. This input consists of legitimate and the phishing URLs. These are passed to the feature extraction step where the URLs are tokenized to get the protocol used, the domain names and the path. A set of 12 features are extracted based on the tokenized information. These features along with the tokens are stored in the CSV format. Then the input to the machine learning algorithms is given from the CSV files both for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

training and testing with training having 90% and testing having 10% of the total URLs in the data set. The algorithms used are decision tree classifier, random forest tree classifier, Naive Bayes classifier and logistic regression. Then the model is tested with the 10% URLs from the same original set. The final prediction tells whether the URL is phishing or legitimate.

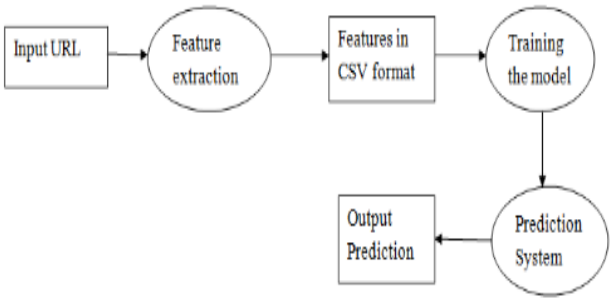


Figure 1: SYSTEM FLOW DIAGRAM

Some of the features considered are :

1. Length of the URL: Long URL can be used by the attackers to hide the doubtful part in the address bar.

2. Presence of symbols like @ domain: The browser will ignore the text preceding the '@' symbol in the URL. This can also be used to hide the suspicious part.

3. Sub-Domain and Multiple Sub-Domains: Most of the legitimate URLs don't contain more than 3 dots (.) in their domain name.

4. Adding Prefix or Suffix Separated by (-) to the Domain: Legitimate URLs rarely have the dash (-) symbol in them. Prefixes or suffixes separated by (-) can be added to make a phishing website to look like similar to a legitimate website.

5. Using the IP Address instead of the domain name in the URL: An IP address can be added instead of the domain name to fool users and hence can steal sensitive information.

6. Redirection using '//': By examining the location of '//' in the URL, decision can be made if there is any unnecessary redirection to other websites.

7. Tiny URL: URL shortening is a method on the "World Wide Web" in which a URL can be made considerably smaller in length and still lead to the required web page.

8. Existence of the 'https' token in the domain part of the URL: The attackers can use this token to trick the users.

9. Web Traffic: The ranking or the popularity of the website determines whether that website is a phishing website or not.

10. Domain Registration Length: Phishing websites tend live only for a short period of time. Some of the legitimate websites live for at least 6 months and most of them at least for 1 year.

11. Statistical reports: Verify whether the URL is present in the records of the phishing URL lists provided by PhishTank etc.

After feature extraction, the processed data set obtained is provided as input for the four algorithms and their performances are compared.

V. IMPLEMENTATION

The features were extracted from the URLs and stored in a CSV file for giving as input to algorithms. The performance evaluation of the proposed system is done in this section. Set of nearly 1000 URLs including legitimate and phishing included for training. The split ratio of 90 percent is used for training and 10 percent is used for testing. Decision tree, Random forest tree, Logistic regression and Naive Bayes algorithms were taken into consideration for comparison of performances.

Decision tree is a classification algorithm which presents the attributes to be classified in the form of tree structure and each leaf node specifies a class label. Top-down approach is followed for construction of decision tree.

Random forest tree is the algorithm used mainly for classification and regression problems by constructing multiple decision trees during training of data. The final class value can be obtained by performing either mean or mode on individual decision tree output class labels.

Logistic regression algorithm is used for binary classification problems. This algorithm is used for regression analysis to examine the relationship between binary dependent variable and one or more independent variables.

Naive Bayes algorithm is used for predictive modeling and uses Bayes theorem for making output prediction. The principle of this algorithm is that each classified feature is independent of each other and has equal weightage for output prediction.

Among the four algorithms chosen Random forest tree and decision tree algorithm almost performed the same with accuracy between 83-86 percent.

The confusion matrix shown in the figure 3 is presenting the test set results of decision tree classification. Similarly, the confusion matrix of the testing results for the other three algorithms is obtained as well.

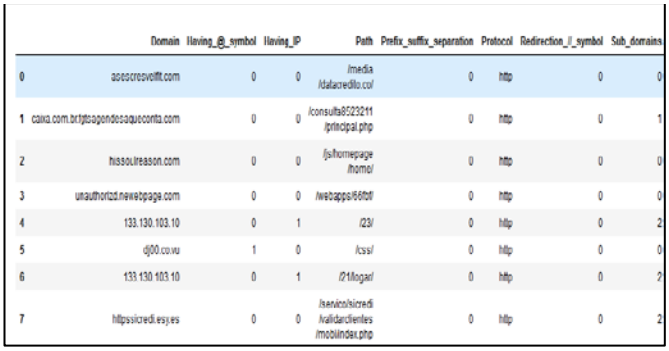
The parameters which are used for performance evaluation are accuracy, precision, recall, f1 score.

The value of these performance parameters are calculated using the equations below:-

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019



	Domain	Having_@_symbol	Having_IP	Path	Prefix_suffix_separation	Protocol	Redirection_//_symbol	Sub_domains
0	aoeocreovelfit.com	0	0	/media /dalacredito.co/	0	http	0	0
1	caixa.com.br.tgtoagondosaqueconta.com	0	0	/consulta8523211 /principal.php	0	http	0	1
2	hissolireason.com	0	0	/js/homepage /home/	0	http	0	0
3	unauthorized.newebpage.com	0	0	/webapps/66fbf/	0	http	0	0
4	133.130.103.10	0	1	/23/	0	http	0	2
5	dj00.co.vu	1	0	/css/	0	http	0	0
6	133.130.103.10	0	1	/21/logar/	0	http	0	2
7	httpssicredi.esy.es	0	0	/servico/sicredi /validarclientes /mobi/index.php	0	http	0	2

Figure 2: EXTRACTED FEATURES FOR THE URLS

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The average value of these performance parameters is calculated and listed down in the table below.

For test data when run for multiple times, Decision tree obtained an average accuracy of 83% while Random forest gave an average accuracy of 85%, whereas Logistic regression and Naive Bayes obtained an accuracy of 76% and 55% respectively on an average.

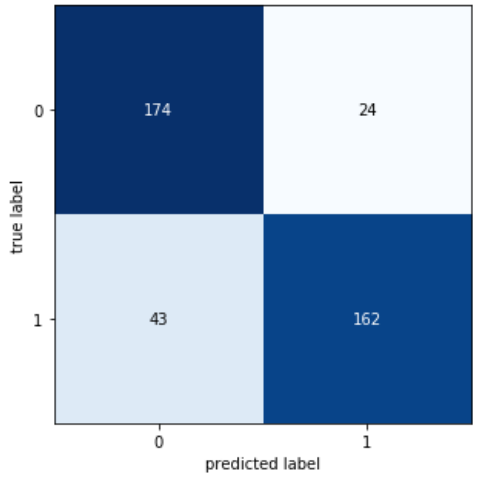


Figure 3: CONFUSION MATRIX FOR TESTING DATASET

Table 1: Average values of performance parameters for the chosen four algorithms

Algorithm	Average Precision	Average Recall	Average F1 score	Average Accuracy
Decision tree	0.80	0.88	0.84	83%
Random forest tree	0.81	0.88	0.86	85%
Logistic regression	0.74	0.80	0.77	76%
Naïve bayes	0.53	1.00	0.69	55%

Below is the comparison graph of testing accuracy for all four algorithms.

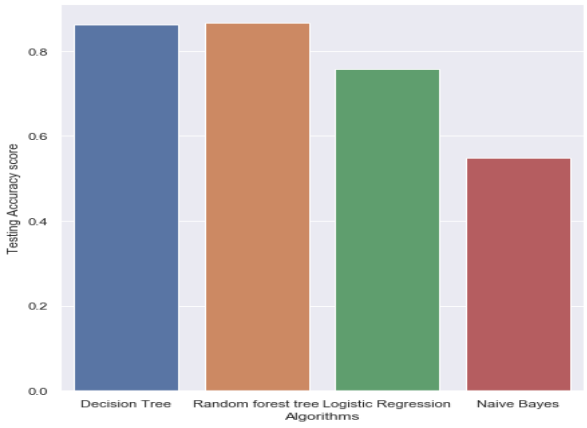


Figure 4: COMPARISON OF ALGORITHMS' ACCURACY

From the graph it can be seen that Random forest tree and Decision tree gives nearly the same testing accuracy ranging between 82 to 86 percent. While Logistic regression gives mediocre testing accuracy but the accuracy of Naive Bayes is the least among the four algorithms chosen.

VI. CONCLUSION

The main objective of this study is to help the users to differentiate between the phishing and legitimate URLs by inspecting the URLs based on particular unique characteristics. This research demonstrates the capability to recognize fake web pages based on their URLs. In order to protect the victim from phishing attacks, educational awareness programs must be conducted. All internet users must follow the security tips given by educational experts during the awareness programs. Users should also be very well trained to recognize a fake website so that they would not enter their personal details believing it is a legitimate website. It is mandatory to inspect the URL link before entering into any website.

In future work, automatic detection of web pages and the web browser extension can be done. Further work can also be done by adding several other features to distinguish the fake web page from a legitimate web page.

REFERENCE

[1] Abdulghani Ali Ahmed and Nurul Amirah Abudullah, " Real Time Detection of Phishing Websites", IEEE University Malaysia Pahang, 2016

[2] Muhammet Baykara and Zahit Ziya Gurel, "Detection of phishing attacks", IEEE, Faculty of Technology Firat University, Elazig, Turkey, 2018

[3] Jian Mao1, (Member, IEEE), Wenqian Tian1,Peti Li1, Tao Wei2, (Member, IEEE), and Zhenkai Liang3, "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", IEEE, 1School of Electronic and Information Engineering, Beihang University, Beijing 100191, China 2Baidu USA LLC, Sunnyvale, CA 94089, USA 3School of Computing, National University of Singapore, Singapore 117417, 2017

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

[4] Jakov Andric, Dijana Oreski and Tonimir Kisasondi, "Analysis of phishing attacks against students", MIPRO NTH, Varazdin, Croatia , Open Systems and Security Laboratory, Faculty of Organization and Informatics, Varazdin, Croatia, 2016

[5] Tajinder Singh, Madhu Kumari and Shweta Mahajan "Feature oriented fuzzy logic based web spam detection", Journal of Information and Optimization Sciences, 38:6, 999-1015, 2017

[6] Renato M. Silva, Akebo Yamakami, Tiago A. Almeida, "Towards Web Spam Filtering using a classifier based on the minimum description length principle", School of Electrical and Computer Engineering, University of Campinas, UNICAMP, Sao Paulo, Brazil, 15th IEEE International conference on machine learning and Applications, pp:470-475, 2016.

[7] Ms. Priyanka Salunkhe (student), Mrs.Smita Bharne, Mrs.Puja Padiya (Assistant Professor), "Filtering Unwanted Messages from OSN Walls", Department of Computer Engineering Ramrao Adik Institute of Technology Nerul, Navi Mumbai, 1st International Conference on Innovation and Challenges in Cyber Security, pp:261-264, 2016.

[8] Shuhua Liu and Thomas Forss, "Text Classification Models for Web Content Filtering and Online Safety", IEEE 15th International Conference on Data Mining Workshops, 2015