International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-6, Issue-6 ,2019 Regular Edition E-ISSN: 2349-7084

Phishing Urls Detection Using Machine Learning Techniques

Sushma Joshi[1*], Dr S.M Joshi[2]

[1,2]*Computer Science and Engineering, SDM College of Engineering and Technology, Visveswaraya Technological University, Dharwad, India*

e-mail: sushmajoshi35@gmail.com, joshshree@gmail.com

**Corresponding Author: sushmajoshi35@gmail.com,*

Available online at: http://www.ijcert.org

Received: 06/June/2019, Revised: 09/June/2019, Accepted: 14/June/2019, Published: 25/June/2019

Abstract:-Phishing is an attempt to get any sensitive information like user identity information, banking details and passwords from target or targets which is considered as fraudulent attack. Phishing causes huge loss to the internet users every year. It is a captivating technique used obtain all the personal and financial information from the pool users of internet. This project deals with the methodologies of identifying the phishing websites with the help of machine leaning algorithms. We have considered the lexical properties, host based and page-based properties of the URLs which are used for identifying the phishing URLs. Various Machine learning algorithms are implemented for feature evaluation of the URLs which have widespread phishing properties. These website properties are refined so that a best suitable classifier tis identified which can distinguish between benign and phishing site.

Keywords: URL, phishing, benign, legitimate, malicious.

---

1. Introduction

Cyber security is the computer system security or information technology security where it deals with the protection of computer systems from any kind of theft or damage to software or hardware.

Phishing is a fraudulent or criminal mechanism to steal the users 'personal information. Here spoofed emails are used claiming to be from legitimate websites which give same look and feel to the internet users making them to give their personal and financial details. Malicious softwares are installed on the systems to steal user information.

Figure 1. and Figure 2. below represents the popular Gmail website. The first figure is the original webpage whereas the second one is the phishing webpage of the site. This phishing webpage of Gmail will always mislead the internet users by which they end up in filling all their financial and personal details in it. Thus, the attacker can use this information for some vicious purposes.

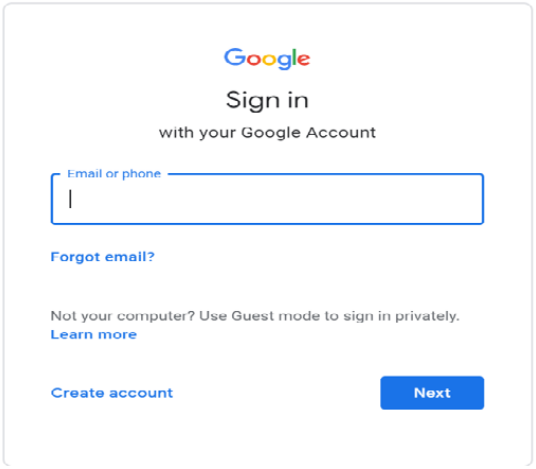

Figure 1. Original Gmail Login Page

 https://doi.org/10.22362/ijcert/2019/v6/i06/v6i0602

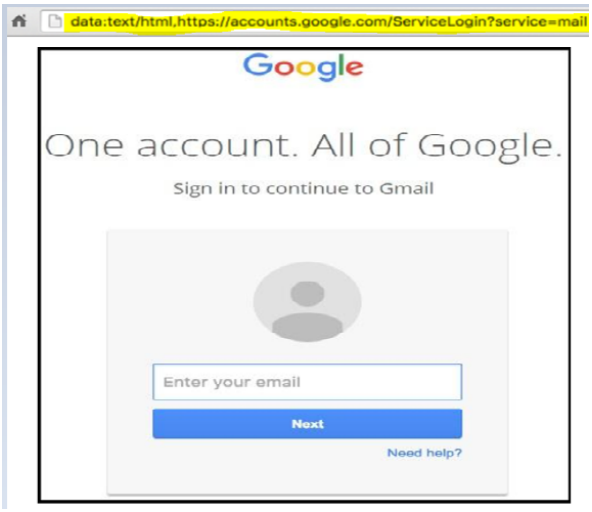

Figure 2. Phishing page

A. ***Phishing Techniques***

The criminals who want to steal the user's information which is a delicate data will first produce the duplicate copies of the original website and email, generally from some financial domain. Email will be created using the same website design that is the logos and the slogans of an original company. The reason for the rapid growth of website creation is the structure and the format of Hypertext Mark-up Language (HTTP). This HTTP structure is so beneficial that it allows users to easily copy the images and also website sometimes. Now, once the email duplicate copy is created the phisher will send the imitated emails to many users possible to make them involve in this fraudulent attempt. When these emails are clicked and opened, users are redirected to the phishing website, which is imitated to be an original one.

B. ***Phishing Attacks statistics***

The phishing attacks are rising up rapidly causing lot of damages to the organization or companies using internet. It has been identified that nearly 1.5 million websites which are phishing sites are generated every month [13].

An example of a phishing attack occurred recently in 2019 at State Bank of India, University Branch Dharwad. Here many customers were sent spoofed emails and instant messages. After clicking the links in the message or email they were asked to fill in all personal and financial details like account number, email -id, mobile number etc. Once the details were submitted the amount from the respective account was stolen which came to the customers' notice after some time.

The United States is the top country for hosting the phishing websites. It is mainly because of the fact that larger percentage of websites and the domain names in the world are hosted by United States.

2. Related Work

Various URLs are analysed by the researches and also phishing websites statistics are measured by them. The previous work of these researchers is reviewed based on which our project is discussed with different ideas.

The work by Joby James, Sandhya L and Ciza Thomas [1] includes the lexical feature, host-based feature and page-based feature analysis of the URLs. Here the dataset is formed using the URLs from browsing history, also from website phishtank.com. The dataset is uploaded into the python program for parsing the URL. Best classifier is chosen. Then the URL is classified as phishing or benign using the chosen classifier. As per this approach the Decision Tree has given the best performance.

Garera et al. [4] have used logistic regression classification model as a key classifier to distinguish between benign and phishing websites. They have identified that logistic regression is very accurate and is applied on several URLs.

McGrath and Gupta [5] have not created any classifier or have used any classification models. Instead they have given a comparative analysis of URLs that is both benign URLs and phishing URLs. Here the benign URLs got from DMOZ Open Directory project are compared with the phishing URLs obtained from the Phishtank website. WHOIS properties, IP addresses, geographic information, registrar provided information and various other features like length, character distribution and the predefined brand names are all analysed in this work for identifying the phishing URLs.

Work by Basnet and Sung [12] have proposed content-based features and they use machine learning algorithms to demonstrate the detection of phishing URLs on real world datasets using the Random Forest classifier.

Bahruddin, Izhar and Shoid [14] have discussed Malicious URL classification which used multilayer perception technique. Here they have used multi-layer perceptron technique as the tool to measure the effectiveness of identifying Malicious URL. Here the dataset was downloaded from Machine learning UCI repository. Data was pre-processed and divided into subsets. Here neural networks get the pre-processed dataset and gives the processed output.

3. Methodology

A. Overview of problem

URLs are also called as "Weblinks" are the most important ones which help us in locating the information on the internet. Our aim is to choose the best classifier which distinguish the URLs into legitimate and phishing site. Classifier is chosen based on the analysis of various properties of the URLs that is lexical, host and page-based properties. We analyze different machine learning algorithms using python language.

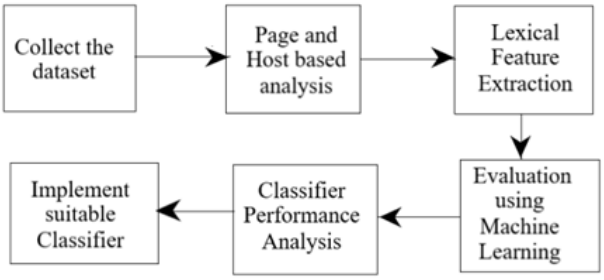

Figure 3. Design Flow

The steps involved in our design flow shown in the Figure 3. are

Step 1: First the dataset is obtained from UCI Machine Learning repository. All the feature fields are vectored with 1 as phishing, -1 as legitimate and 0 as suspicious in our dataset.

Step 2: The page-based, host-based and lexical based feature extractions are performed on the URL which is to be classified into legitimate or phishing URL.

Step 3: The dataset is mined using different machine learning algorithms for feature evaluation.

Step 4: After the feature evaluation is done the performance of the classifier is analyzed and the best classifier is chosen.

Step 5: The suitable classifier is implemented and this helps in distinguishing the URLs into legitimate and the phishing URLs.

B. Host Based Properties

Host based properties will tell "who"," where" and "how" about the phishing sites. That is, where the phishing sites get hosted, who is managing the sites and how the sites are controlled. Phishing websites can be hosted on unusual hosting sites, on machines which are non-reputable or through some unusual registrars.

The properties of the hosts that are identified during host-based analysis are explained below:

1) WHOIS property: WHOIS properties give the details on the registration date, update and expiry information and the information on the registrar and the registrant. If the phishing sites are accessed frequently then they have newer registration dates compared to the legitimate sites. Most of the websites have ip address in their hostname. Below diagram [1] represents host-based analysis.

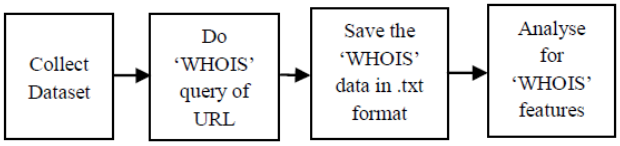

Figure 4. WHOIS property analysis block diagram

2) Age of Domain one of the WHOIS property will check the webpage domain name age. Most of the websites are hosted on domains which are recently registered so they have relatively young age.

3) Geographical Properties: These properties will give information on the ip addresses. It tells about location to which the address belongs to.

4) Blacklist membership: Blacklists are the precompiled lists of malicious URLs which will have malicious sites, ip addresses and different domain names. Blacklist URLs have to be avoided by the users. On similar grounds, white lists will contain all the URL lists which are safe to be used by the customers or users. Black lists can be of following types

- Blacklist based on DNS: A query is submitted by the internet users .Query usually constitutes domain name or ip addresses.This is sent as a question to the DNS server of black list provider, and in reply to that will be an ip address saying whether that is blacklisted or not.
- Browser Toolbars: These act a defence system for the users from the client side. The toolbar sees to it that the URL is intercepted from the address bar much before the user visits a site and the cross refers to check whether URL exists in the blacklist or not. This URL is usually stored locally either at

users' side or the server so that browser can query. If it is a malicious site then the user is given a warning about the site. Some of the examples are McAfee Site Advisor, Google toolbar.

- Network Appliances: This is one more better option for establishing the blacklists. These behave like proxy between the internet and the users. When the users within the organization visits the site the networks appliance will check the outgoing connection and cross references with the precompiled blacklists.

C. *Page Based Properties*

These properties tell us about the popularity of the page that is how much popular the web page is or how users use that web page frequently. Various features are as follows:

1) Page Rank (PR): Google uses this method to determine relevance or importance of a page. Google performs re-indexing frequently during this time the maximum page rank gets changed that too it happens every month. Link analysis algorithm that is page rank algorithm was used by Google initially, where the web numerical weight numbering from 0 to 10 is allocated on each document, where 0 value indicates lesser popularity and 10 the most popular. Suppose the PR value for a particular webpage is unavailable then -1 value is assigned. The sum of all the page ranks is equated to unity forming a probability distribution on all webpages. It is found that the legitimate sites have long life whereas the malicious ones have smaller life. So, the phishing pages have very small value of page rank or sometime their page rank value does not exist.
2) Details of Traffic Rank: Website popularity is identified by this rank. Alexa.com of Amazon lists various websites ranks with respect traffic of internet using previous record. Traffic rank close to value 1 is accurate. Ranks having value greater than hundred thousand (100,000) are not so correct as there could be more chances of error.

D. *Lexical based features*

URL's textual properties are considered as lexical features not the entire content of the webpage. These text strings are parsed in a standard way using client programs. Each URL is translated into instructions by browser, server which hosts the site is located and location of the host on the host site or where the resource is placed is obtained. This is done using multistep resolution process. To understand this process of translation, the following standard syntax of URL is considered.

<protocol>://<hostname><pathname>

Below us the URL resolution example [1]

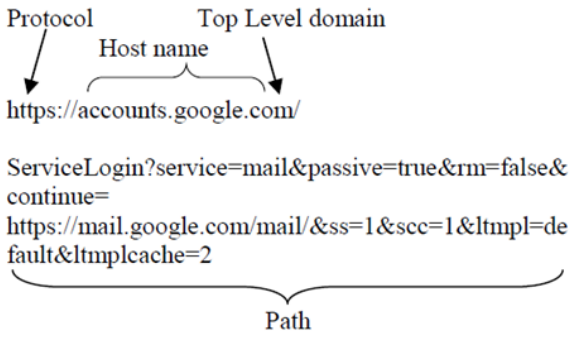

Figure 5. URL Resolution example

The <protocol> part of the URL represents which type of network protocol should be considered to obtain the resource which is requested. The commonly used protocols are HTTP with Transport Layer Security (https), Hypertext Transport Protocol or HTTP (http) and File Transfer protocol (ftp). Web server on the Internet is represented by <hostname> as its identifier. Most of the times it is human-readable name of domain but sometimes it could be machine readable ip. The <path> part of a URL is the path name of a file as shown in the Figure 5. which appears on a local computer. The site organization is shown by the delimited path tokens which are delimited by many punctuations that is slashes, dots, hyphens etc. Sometimes these path tokens are made undefined to avoid scrutiny, or these thieves may purposely create these tokens to disguise the legitimate site.

Lexical properties constitute hostname length, URL length, also the number of dots in the URL. Lexical properties represent a concept called as a "bag-of-words".

Here the multiset of words is considered disregarding the grammar. Lexical properties include below properties:

1) Hexadecimal characters: URL can be typed from the keyboard which is easily understood by the computer. It has numeric decimal value which can be easily translated to hexadecimal base. Web browsers can understand hexadecimal bases more easily."%" character is used as the preceding character in this to represent the typed character from a keyboard. For instance, value %20 represents space character from keyboard.
2) Suspicious character: "@" and "-" symbols are identified as the commonly used suspicious character in phishing

URL to disguise the original one. Here these concatenating characters are used such a way that the left part of the URL is not considered whereas the right part of the symbol is actual URL which is used for obtaining the webpage of the phishing site. Let us consider the URL "http://www.onlinesbi.com@ phishingsite.com". It will navigate to the actual URL which is a "phishingsite.com" but will try to login into www.onlinesbi.com with login details. Thus, the actual URL of the website is hidden. But when its combined with an IP address it appears as the legitimate site even though it is a phishing one.

3) Number of dots in URL: This is another feature to identify the phishing sites. Usually many phishing sites tend to add more dots in their URL which is the key to distinguish them from benign by counting the dots.
4) Redirecting using "//": The presence of "//" symbol in the URL path says that to which site the user will be navigated to. The location of "//" is important. We need to check, if the URL starts with "HTTP" then "//" should appear in the sixth position. But if the URL uses "HTTPS" then the "//" should appear in seventh position.

Some of the other properties of the URL considered in our project are

5) Using the IP Address: Let us consider using IP address rather than using domain name like, "http://123.4.5.6./fake.html". In these cases, user's information can be easily obtained by the phishers.
6) Long URL to Hide the Suspicious Part and URL shortening services: Long URL is another medium to mislead the internet users where this long URL is used to keep the suspicious data such that it goes unidentified to the users.URL shortening is famous method used today where the original URL is made shorter considerably but it meets the requirement of the original webpage. This short URL is another way to mislead the internet users.
7) Sub Domain and Multi Sub Domains: Let us assume we have the following link: "http://www.sdmcet.ac.in/". Here "in" indicates the country code,"ac" indicates the academic and sdmcet indicates actual domain name. ccTLD (country-code Top Level Domain) and SLD (second level domain) are important parts included in domain name. "in" is ccTLD and "ac.in" is the SLD parts of domain name." Now, in order to identify the phishing site, the technique used is, first "www" part is removed, then ccTLD part is removed. Now number of dots in the URL are counted. If the dots are greater than 1 then URL is "suspicious", if dots are greater than 2 then URL is "phishing site" which have multiple domains. If no subdomains are found then the URL is legitimate.
8) HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer): This property is another important one to identify the phishing content in the URL. But this is not enough. We need the authorized certificate associated with HTTPS used in the respective URLs. Various top listed authorized certification providers are Verisign, GoDaddy, GeoTrust, Doster and many more.
9) Domain Registration Length: Phishing URLs will always have shorter lifetime. So, the registration length is important factor which tells for how long the fraudulent domains have been used.
10) Google Index: Whenever the internet user searches in the google search engine that particular site is indexed. So many sites are already indexed. So, this property will help us to check if the website we are searching is already present in the google index or not. Phishing webpages are not indexed because they are accessed for the shorter time period. Thus, this property helps in identifying the phishing site.

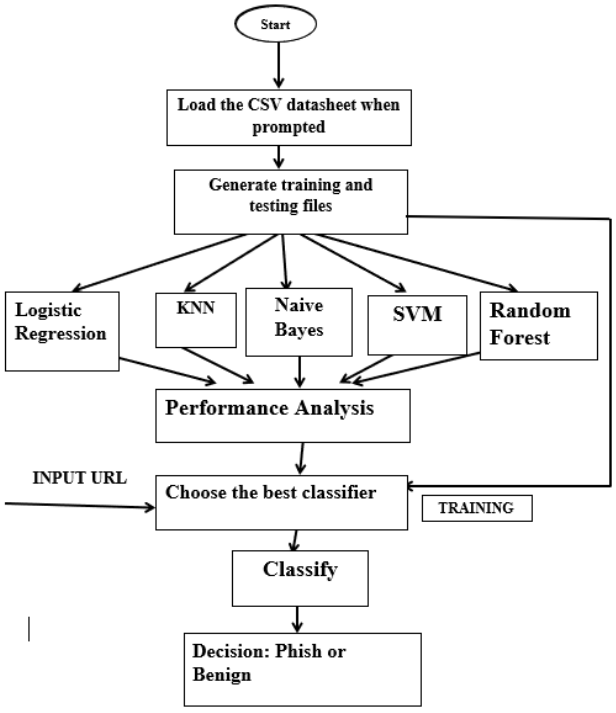

Fig.6. Flow Chart on Design Flow

E. Program Flow

The methodology depicted in the above flowchart Fig.6. used in this work is as follows:

- The dataset with various features is analyzed by machine learning algorithms using Logistic regression, Naive Bayes, K-Nearest Neighbor, Random Forest and Support Vector Machine (SVM).

- Data is split into training data and testing data with the two different percentage splits for analysis. One with slit percentage of 60 where it says 60 percentage is the testing data and remaining 40 percentage is the training data. Similarly, analysis is made for 90 percentage split data. Performance is measured based on the Confusion matrix, accuracy, precision score, recall value, f1-score and support value. Based on these scores the best classifier is chosen.
- Once the analysis is over, a python program is executed which asks for the input URL. This URL is parsed through another python program which does feature extraction of the URL. Feature extraction involves the features of host based, page based and lexical based.
- This parsed URL is then analyzed and evaluated by the best suited classifier. Then the URL is classified as legitimate or phishing one. If the URL is legitimate one then -1 is the output, 1 if it is phishing site and 0 if it is suspicious site.

F. Machine Learning Algorithms

The machine learning algorithms implemented are:

1) Logistic Regression: Logistic regression has a function called logistic function also called sigmoid function. This is a classification algorithm which outputs the discrete value that is 1 or 0. In our work logistic regression is used to predict the URL is benign or phishing one. It uses the sigmoid function to predict.
2) Naive Bayes Classifier: Bayes theorem is the basis for the Naive Bayes Classifier. Here every pair of features that is classified is independent of each other.

In our dataset, it contains all the features related to URL which needs to be identified whether it is phishing or legitimate. Naive Bayes fundamentals is that it makes an equal and independent contribution to the outcome. In our dataset no pair of features are dependent. Example URL length and port. Both does not affect each other. Similarly, all features have been given same importance. Knowing only two or three features we can't predict the output. No attributes are irrelevant and all are contributing equally to the output. All the features should be independent and there should not be any correlation between them. Then only Naive Bayes gives better performance. In our work we can some of the features are correlated so Naive Bayes's performance has gone down compared to other algorithms.

3) K- Nearest Neighbors: This algorithm is based on the samples which are closest in a given feature space. Classification happens based on the majority of the votes of its neighbors.

G. Algorithm

Step 1: Initialize value of K.

Step 2: Iterate from 1 to total number of training data values to get the predicted class

- Calculate the distance between each row of training data and test data. Euclidean distance measure can be used.
- Sort the distances that are calculated in ascending order distance values as the basis.
- From the sorted array obtain the top k rows.
- These rows are used to get the most repeated class.
- Return the class which is predicted.

KNN is one of the most eligible algorithms for classification. It can also be used for the regression problems.

4) Support Vector Machine (SVM): SVM's (Support Vector Machine) hypothesis is not interpreted as the y's (output) probability being 1 or 0 (as it is used for the hypothesis of logistic regression). Instead, it will output either 1 or 0.

Htheta(x) = 1 if theta^T .x >= 0

0 otherwise

H is the hypothesis.

SVM makes use of the term kernel. Kernels allow us to make non-linear complex classifiers using Support Vector Machines. There are two types of SVM defined based on the boundaries that is linear boundaries and nonlinear boundaries. To find the nonlinear boundaries we use a kernel called as Gaussian kernel. For linear decision boundaries SVM works similar to that of logistic regression. In our work SVM is used to get either 1 or -1 if the given URL is phishing or legitimate respectively.

5) Random Forest: Random Forest is a classification model falling under the category of supervised learning algorithms. This algorithm is the collection of decision trees where they are trained by the bagging method. Bagging method says all the learning models are combined to increase the overall result. Random Forest in simple words can be defined as creating the multiple decision trees and combining them together to get well built and accurate prediction. In our work discussed here Random Forest performs the best.

4. Results and Discussion

The main findings of our work are

- Phishing URL's domains and URLs have different characteristics than the legitimate site. Phishing URLs have different length also.
- Most of the URLs have the targeted names of the brands.

The prepared URL was analysed in a python program using logistic regression, knn, naive bayes, svm and random forest classifying algorithms. As explained earlier the data was split into 60 percentage and 90 percentage training data. Performance analysis is done and performance is measured based on Confusion Matrix, Accuracy, Precision, Recall, F1 score and Support. The python program analyses the dataset and gives the classifier performance as tabulated in the below tables.

Table 1. Classifier Performance (60% split)

Test Options	Classifier	Confusion Matrix	Accuracy
Percentage Split – 60%	Logistic Regression	2642 291 186 3514	92.8%
	KNN	2696 237 220 3480	93.1%
	Naive Bayes	2929 4 2638 1062	60.2%
	SVM	2742 191 108 3592	93.9%
	Random Forest	2783 150 66 3634	96.1%

Table 2. Classifier Performance (90% split)

Test Options	Classifier	Confusion Matrix	Accuracy
Percentage Split – 90%	Logistic Regression	4014 388 388 5214	92.7%
	KNN	3935 467 520 5028	90.0%
	Naive Bayes	4395 7 3912 1636	60.6%
	SVM	3736 666 352 5196	92.9%
	Random Forest	4111 291 202 5346	94.4%

Table 3. Confusion Matrix

	Actual Value		
Predicted	4111	491	4402
Value	201	5346	5547
	4312	5637	(4111+5346)/ (4312+5637) = 95%

Let us consider we have 9950 samples in our working dataset displayed in the above confusion matrix. Now if we calculate the accuracy for 90 percentage split data using accuracy formula shown below that is manually, we can see accuracy is 95 percentage which is approximately same as that shown in the Table II for Random Forest. The accuracy formula is

Accuracy = (True Positives + True Negatives)/Total of 4

When we check the performance, we can see that all the classifiers give different accuracies. Random Forest has the best performance compared to all other classifiers. Using the lexical based feature analysis, we were able to get the highest accuracy of 96 percentage with the 60-percentage test split. With the 90-percentage test split, highest accuracy measured was 94.4 percentage. The lowest performance was given by the Naive Bayes classifier.

The URL is loaded to the classifier which finally makes a decision whether URL is 'phish' or 'benign'.

5. Conclusion and Future Scope

Many features of URL were compared using the machine learning algorithm. Results gave accuracy based on the different properties of URL that is page-based property, host-based property and lexical based property. The users of the internet can be given the protection against these malicious sites by identifying the phishing URLs using these features of URL. The major challenge in Cyber Security is that criminals constantly adopt different strategies to overcome our defence measures. We have to use different algorithms which accommodate themselves into these rapidly changing techniques of phishing URLs which is a major challenge in the domain of Cyber Security.

References

[1] Jobv James, Sandhva L, Ciza Thomas: Detection of Phishing URLs Using Machine Learning Techniques. In Proc. Of 2013 International Conference on Control Communication and Computing (ICCC).

[2] J. Ma, L. K. Saul, S. Savage and G. M. Voelker," Beyond Blacklists: Learning to Detect Phishing Web Sites from Suspicious URLs", Proc.of SIGKDD '09

[3] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker," Learning to Detect Phishing URLs", ACM Transactions on intelligent Systems and Technology, Vol. 2, No.3, Article 30, Publication date: April 2011.

[4] Garera S, Provos N, Chew M, Rubin A D, "A Framework for Detection and measurement of phishing

attacks" In Proceedings of the ACM Workshop on Rapid Malloced (WORM), Alexandria, VA.

[5] D K McGrath M Gupta "Behind Phishing: An Examination of Phisher Modi Operandi" In Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET).

[6] C. Whittaker, B. Ryner, and M. Nazif. Large-scale automatic classification of phishing pages. In Proc. Of the 17th Annual Network and Distributed System Security Symposium (NDSS'10), California, USA, February 2010.

[7] Phistank. https://www.phishtank.com

[8] Curlie. https://curlie.org

[9] I. Rogers "Google Page Rank – Whitepaper"

[10] Shraddha Parekh, Dhwanil Parikh, Srusti Kotak, Prof Smita Sankhe, A New Method for Detection of Phishing Websites: URL Detection ", In Proc Of 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)

[11] Ram B Basnet and Andrew H Sung "Learning to Detect Phishing Webpages" In proceedings of Journal of Internet Services and Information Security.

[12] Mohammad Fazli Baharuddin, Tengku Adil Tengku Izhar, Mohd Shamsul Mohd Shoid "Malicious Url Classification System Using Multi-Layer Perceptron Technique", In proceedings of the Journal of Theoretical and Applied Information Technology.

