Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Machine Learning Algorithms for Questing of Phishing URLs

B. Bala Krishnudu* and V. Raghunatha Reddy

Dept. of Computer Science and Technology, Sri Krishna Devaraya University, Anantapuramu, Andhra Pradesh - 515001, India.

ABSTRACT

Phishing is being social engineering attack devastating the web users in terns of loss of information including login credentials and credit card numbers makes them to loose billions of dollars per annum. It is alluring fraudulent technique employed by thieves /intruders /phishers to fish for security or private data in the ocean of unsuspecting web users. Phishers uses spoofed-mails; phishing software is used to steal privacy data and financial details. This paper deals with the identification of phishing web sites by analyzing fraudulent features of benign and phishing URLs assessed by Machine learning techniques through proposed algorithms. The main aim of this (method) algorithm is to detect of phishing websites on the ground of lexical features, host properties and page importance properties. It emphasizes different data mining algorithms for evaluation of the features in order to get a better understanding of the structure of URLs that diffuse phishing. The fine-tuned parameters from URL, page rank and WHOIS query are being selected and used in the appropriate machine learning algorithms for decomposing the phishing sites.

Keywords: Phishing; benign; URL; Page Rank; WHOIS

***Corresponding author**

B. Bala Krishnudu

Dept. of Computer Science and Technology,

Sri Krishna Devaraya University, Anantapuramu,

Andhra Pradesh - 515001, India.

Email: balakrishnudu81@gmail.com, Mob No – 9292250220

INTRODUCTION

Phishing is the fraudulent technicality/crafting using the Social Engg and Technical tactics to steal the customers/consumers Personal Information and Financial Account details. Social Engg system/plot/scheme will use for spoofing emails, and intended to be mislead the legitimate/lawful business and agencies, designed to pioneer the customers to visit dummy websites that deception/trick make the end users or receivers to unveil their personal data that might be financial data or confidential data etc., Technical trickery scheme install malicious software onto Computers, to capture user's personal and privacy data directly, frequently using systems to pick the users online A/C username and passwords.

Figure 1 shows the webpage of very famous website www.linkedin.com. Figure 2 show a webpage akin to that of linkedin; but it is the webpage of a website that scatters phishing functionalities activities. A user may misunderstand the second website as a genuine linkedin website and bestows personal identity information. The phisher can now capture the data and may explore it in on different ways and use it for several notorious reasons which lead to a cruel severity.

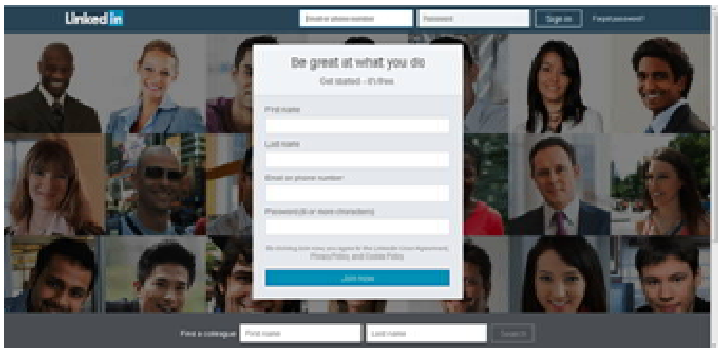

Figure 1. Original Linkedin webpage

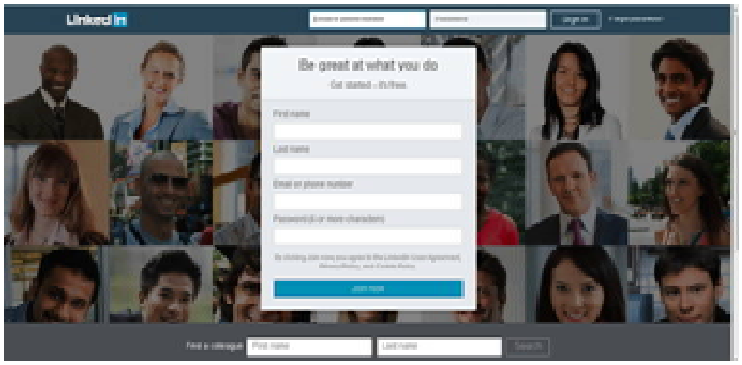

Figure 2. Phishing webpage

The Technique of Phishing

The introducer who wants to attain concern important data, initially he would create a dummy website or email which is a replica for the original One, actually from a financial Organization or some other institution that deals with the financial or confidential data. The main reason of creating a website is that the "internet" was became communication medium to everyone across the globe and scattered drastically, also permits misuse of trademarks and trade names, and other corporate identifiers on which consumers would be dependent as such mechanisms for reputation and authentication.

The intruder can send the 'morphed' or 'spoofed' emails to as many people as possible in one attempt to allure them into the plot. Whenever these emails are opened or a link on the mail is clicked, the consumers are redirected to a 'spoofed' or 'morphed' website that can be appeared as a legitimate entity.

Statistics of Phishing Attacks

Phishing becoming one of the rapidly growing classes of identity theft scams on the internet, which is being caused to both short term and long-term economic damages. Still if continues its legacy. There are some of the statistics about phishing and emails fraud-2018 or phishing and email statistics.

i. The amount of spam emails was increased by four (4) times in 2016, International Business Machines Corporation (IBM).
ii. Phishing account for 90% of data breached/ruined.
iii. 1/131 emails are contained malware in 2016(Symantec).
iv. 15% of people successfully phished would be targeted at least one more time within the annual period/year.
v. BEC (Business Email Compromise) scams accounts around $5 billion in losses from Oct 2013-Dec2016, Federal Bureau of Investigation's (FBI).
vi. 870 organizations received W-2 phishing emails in the first 4-months of 2017, Internal Revenue Service (IRS).
vii. Phishing attempts lave raised 65% in the last year
viii. Around 1.5, new phishing websites are being created in every month (web root)
ix. 70% of organizations reported that victimized to phishing in 2017
x. 76% of business reported that are being victimized to phishing attacks in the last one year.
xi. 30% of phishing messages are opened by targeted users (Verizon)

Actually, an instance of phishing was occurred in June, 2004, the Royal Bank of Canada noted that fraudulent E-mails were intending to generate from the Royal bank sent out asking customers to check account numbers and PIN (Personal Information Number) by a link that included in the Email[2].

The trickery mail/deceptive E-mail stated that if any customer did not click on the link, key on his client card number, pass code, accessing to his account would be blocked. These types of emails sent within a week of a computer malfunction that led the accounts of customers not updated. The United States continued its legacy for hosting phishing websites and became the large generator of the world's websites and domain names are hosted across the globe. Financial services becoming always targeted sector by the phishers[1].

RELATED WORK

Many of researchers have been analysed the statistics of suspicious URL's in some way. My approach bestows paramount ideas from earlier studies. Actually, I have reviewed an earlier work in the phishing site using URL salient features that pushed me to do this task.

Ma et al.[3,4] compared and analyzed several batch-based algorithms for classifying phishing URLs and showed that the combination of host-based and lexical features results in the highest classification accuracy. At the same time, they compared and analyzed the performance of batch-based algorithms to online algorithms, while using full features and get found that online algorithms especially Confidence-Weighted (CW), outperform batch-based algorithms.

Garera et al.[5] work said that uses logistic regression over hand-selected features to classify phishing URLs. The features which have the presence of red flag keywords in the URL, features based on Google's Page Rank, Google's Web page quality guidelines. It is very difficult to make a direct comparison with my approach without access to same URLs and features.

McGrath and Gupta[6] did not construct a classifier, but they performed a comparative scrutiny of phishing and non phishing URLs with respect to datasets. And they compared non-phishing URLs got from DMOZ Open Directory Project[7] to phishing URLs from Phish-Tank[8]. The features what they analysed including IP addresses, WHOIS thin records contain date and registrar-providing information, geographic information, lexical features of the URL like length, character distribution, and presence of predefined brand names[6].

PROBLEM OVERVIEW

Actually, URL's known as "Web links" are the primary factors which helps the user to find some information over the Internet. Basically, my target is to define the classification model that finds the phishing websites by assessing the lexical and host-based features of URL's. I analysed different types of classification algorithms on Waikato Environment for Knowledge Analysis (WEKA) workbench and JAVA.

DESIGN FLOW

The work what we took over concisely Host based, page based, the collection of URLs' with respect to lexical feature Extractions and analysis/assessment. With above said elements we can create a database, which is knowledge, mined employing using various types of machine learning methods. A particular classifier can be selected after analyzing the classifiers; and then it will implement on Java. We can see the design flow below Figure 3.

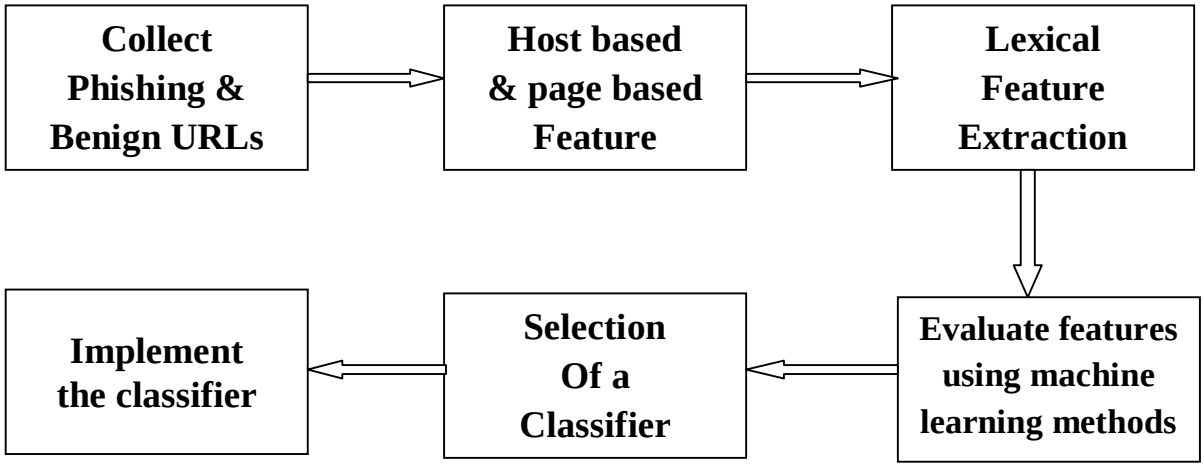

Figure 3 Design flow graph

The Collection of URL's

I collect the Data of URL's of benign websites from www.dmoz.org[7], www.alexa.com[9] and personal web browser history. Fairly the collected data sets from www.phishtak.com[8]. The data set consists of 20000 phishing URLs and 25000 benign URLs. With the help of Page Rank Checker[11]. I checked datasets that I collected, I got Page Rank of 240 benign websites and phishing websites each at the same

time. I collect WHOIS[12] information of 240 benign websites and phishing websites each. (Each website will be checked out individually).

The Host based Analysis

The Host-based features explain "where" the phishing websites can be hosted, "who" can manage, "how" can administer. According to above features eventually we get knew that phishing Web sites can be hosted unfamiliar web hosting centers, on machines that are unusual hosts and even disreputable registrars. The casual block diagram for the Host based Analysis is shown in Figure 4

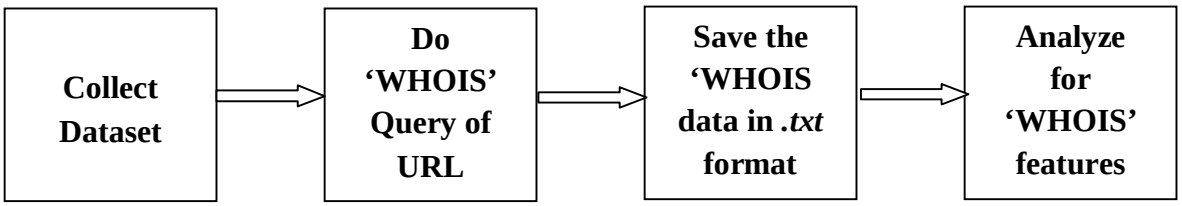

Figure4. Block diagram for host-based analysis

The following are the properties of the hosts which are identifier.

1. **WHOIS properties:** It gives the details of Date of registration, update and expiry, who will be the registrar and the registrant. If the phishing sites are being taken frequently, the registration Dates would be newer than for legitimate sites. A large number of phishing websites will contain IP address on their hostname[5] therefore the information of such hostnames would be helpful in efforts to find the phishing sites.

2. **Geographic properties:** It gives the details of city about the continent or country or city to which the IP address belongs to.

3. **Blacklist membership:** The huge percentage of URL's has been present in blacklists. In the context of Web browsing blacklists are databases or precompiled lists which contain or hold IP addresses, domain names or malfunctioning site URLs can be avoided. On the other hand, white lists hold websites that are known to be safe.

    a. ***DNS-Based Blacklists*:** Suppose a Web browser put a query that represents IP address or the domain name, to a sophisticate 'DNS server' it represents the response, as whether the query

would be present in the blacklist. SORBS[13] URIBL[14], SURBL[15] and Spamhaus[16] are examples of major DNS blacklist providers.

b. ***Browser Toolbars*:** It provides a client side based defending system for the users. Whenever a user visits a site (suppose it is a phishing website), the toolbar prevents the URL from the address bar and it shows cross references a URL blacklist that is often stored locally on the user's machine or on a server that the browser poses a query. Suppose the URL is present on the blacklist, then automatically the browser signs the user to a specified warning screen which provides information about the threat. McAfee SiteAdvisor[17], Google Toolbar[18] and WOT Web of Trust[19] are best examples of blacklist backed browser toolbars.

c. ***Network Appliances*:** Dedicated network hardware is one of predominant and familiar option for deploying blacklists. These appliances will serve as proxies between user machines/nodes within an enterprising network and the rest of the Internet. The end users within an organization visit websites, while the network appliances prevent outgoing connections and IP addresses against the Database of the black listed. IronPort attained/held by Cisco in 2007 and WebSense are the best instances of companies which produced blacklist backed network appliances.

Limitations of Blacklists:

The main benefit of blacklists is that querying is a low overhead operation: the lists of malicious/harmful websites are precompiled or database, so the only computational cost of deployed blacklists is the lookup overhead. However, need to construct/configure/structure these lists in advance give arise to their disadvantage that blacklists become stale/musty. Network administrators will block existed harmful sites or malicious sites, and apply the efforts to remove criminal/culpable enterprises behind those websites. Actually, they will have a consistent persistency to keep building a malicious website and to get found a new hosting infrastructure. Consequently, still new malicious/harmful URLs are introduced and blacklist providers should update their information lists again. By this reason in this process the culprits are being always ahead since building a website is very cheap. Further, free services for blogs example Blogger[20] and personal hosting example Google Sites[21], Microsoft Live Spaces[22] those provide another cheap/trivial/trifle source of disposable websites.

4. **Page/Popularity Based Property**: The Popularity features that define how familiar a web page is among the Internet users. Different kinds of popularity features represented in below:

    a. ***Page Rank:*** It is one kind of methods that Google employs/uses to describe a webpage's importance. Every month the PR of all WebPages on the web keep on changing at max when Google does its re-indexing. The Page Ranks[10] form a probability distribution over web pages; the sum of all web pages' Page Ranks will be equal to unity thereby.

    b. ***Traffic Rank details:*** Traffic Ranks of websites shows a site's familiarity. Alexa.com that ranks different websites as per the Internet traffic based on the basis of past 3 months. The Traffic which close to 1 it shows it can be accurate. Ranks more than 100000 are not as accurate as above said that reflect a severity error.

5. **Lexical feature analysis:** Lexical features are nothing but textual properties/features of the URL itself and its not point to the content of the webpage. URLs can be read by humans because the text strings that are parsed in a standard way by client programs. By the multistep resolution process the web browsers convert each URL into some of instructions that locate the address of the site to which belongs to a server and revile where the resource take place on that host. To customize this machine translation/conversion process, URLs have the following standard syntax.

<protocol>://<hostname><path>

An example of URL resolution is shown below:

The <protocol> version of the URL represents which of the network protocol can be used to fetch requested resource. The most prominent protocols are being utilized such as Hypertext Transport Protocol or HTTP (http), HTTP with Transport Layer Security (https), and File Transfer Protocol (ftp).

The <hostname> is represented entity for the Web server on the Internet. Sometimes it should act as a machine-readable Internet Protocol (IP) address, but in the prospect of human-readable domain name.

The <path> of a URL is the similar to the path name of a file on a local computer. Various punctuation marks (example: slashes, dots, and dashes) delimited/bound path tokens and show how site will be organized. Culprits/Criminals observe/cover/belie path tokens to avoid scrutiny/check, or they may create/build the tokens which can be guise of lawful/ legitimate web site.

The pattern/methodology what used in our work to get extracts the lexical features from the list of URL list is followed as follows:

i. Some of lawful/legitimate URLs gathered/accumulated/collected from the alexa.com and dmoz.org, first written on the notepad/word and that file was saved in the computer; afterwards Java Program (tool) is get executed. It will ask you a file as input. At first upload/provide/feed the benign URL list to the JAVA program. The program processed the list and the feature list which are attained/acquired/obtained.

ii. The Decision Vector ‘0’ can be added. The list will be saved in excel and csv format in the location on the computer as if specified in the program. The same procedure is done for phishing URL list. The same kind of process/method/procedure can be done for phishing URL-list as well.

iii. The Decision vector ‘1’ is added. The feature which consists of host length, path length, number of slashes, and number of path tokens etc. The below Figure shows the flowchart of feature extraction.

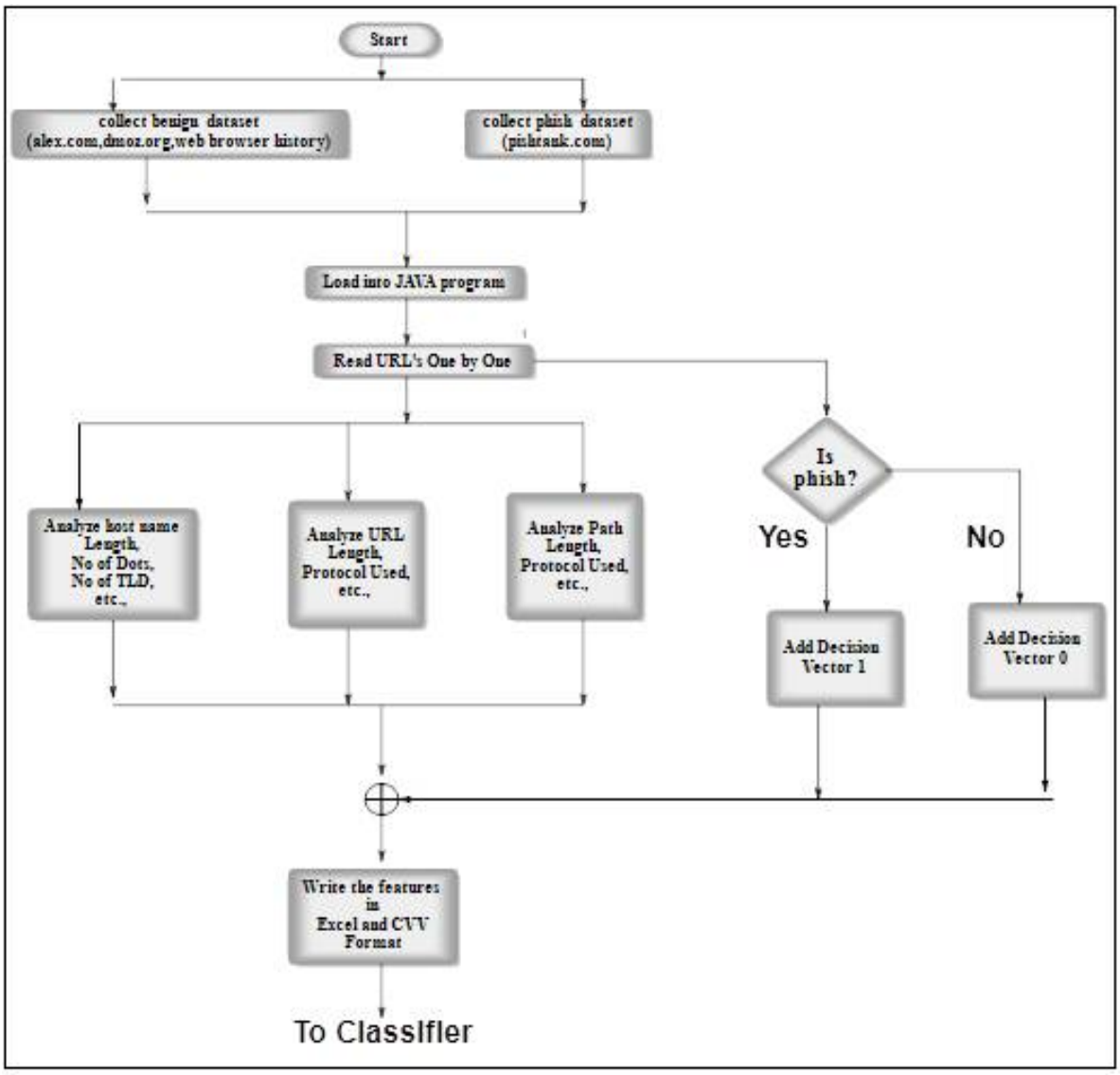

Figure 5 Flow chart for feature extraction

Machine Learning Algorithms

The assessment of different classifying algorithms is done by using the workbench the for-data mining, Waikato Environment for Knowledge Analysis (WEKA) [22] and using JAVA. Four types of input data files are there Attribute Relation File Format (.arff), Comma Separated Values (.csv), C4.5, binary is allowed in WEKA. In our work process (methodology) we use .csv file format. The input file to the WEKA was gotten by the JAVA program on appending 'YES' in the place of decision vector '1' (phish) and 'NO' in the place of decision vector '0' (benign) of the dataset establishment/generated by JAVA program from the input URL list. The evaluation/assessment was done using percentage split 60%.

The input to the classifiers in JAVA will be classified into four types' as .txt files

i. test.xls
ii. testresult.xls
iii. train.xls
iv. trainresult.xls

There are four machine learning algorithms for processing the feature set.

1. **Naive Bayes:** It is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes's rule) with strong independence (naive) presumption/assumptions. The parameter estimation for Naïve Bayes models will be used the max-likelihood estimation. It would only one pass over the training set and it is computationally very fast. Naive Bayes implementations on Multi class Prediction, prediction on Real time, Spam Filtering/ Text classification/ Sentiment Analysis and Recommendation System.

2. **J48 decision tree:** It is a Predictive- machine-learning model which decides/prescribes the target value (dependent variable) of a new sample based on various/different attribute values of the available data.

3. **K-NN:** It depends on the closest training examples in the feature space. An object can be classified by its neighbors on account of Max votes. K-NN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique, we generally look at 3 important aspects:

   i. Ease to interpret output
   ii. Calculation time
   iii. Predictive Power

4. **SVM:** The performs classification on finding/getting the hyper plane which maximizes the margin between two classes. The vectors described the hyper plane are the support vectors.

The program flow for the classifier performance is shown in Figure 6. at maximum.

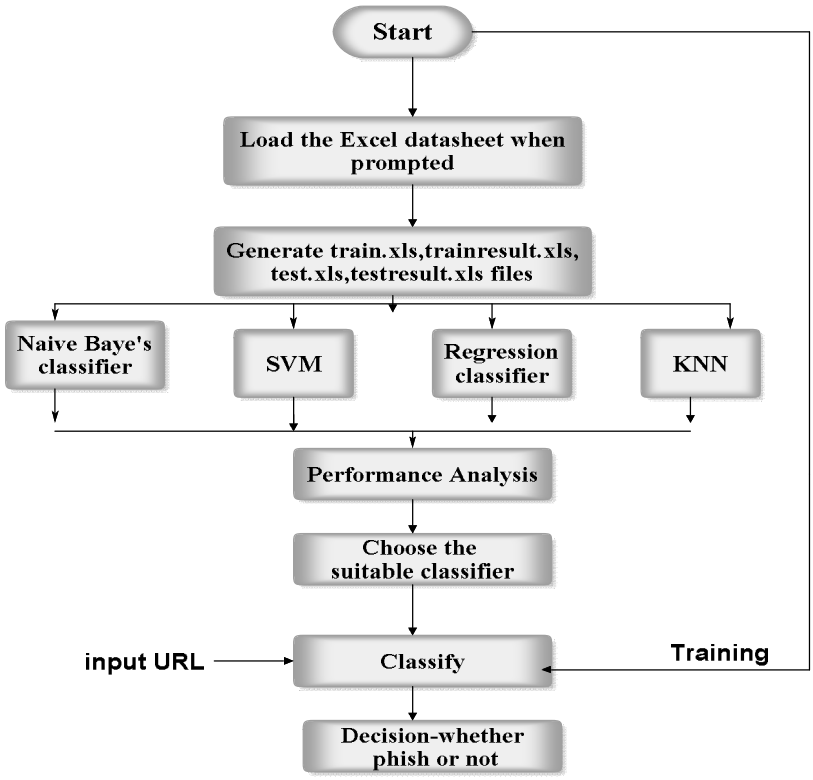

Figure 6. Program flow

RESULTS

The main findings of our preliminary work include:

- Phishing URLs and domains exhibit characteristics that are different from other URLs and domains.
- Phishing URLs and domain names have very different lengths compared to other URLs and domain names in the Internet.
- Many of the phishing URLs contained the name of the brand they targeted.

Page Rank of benign and phishing websites were collected using Google Page Rank Checker[11] and are presented in Page Rank will be employed / used for phishing websites are (i) Not Available (ii) Non-Existing and 0.

The N/A page rank (grey page rank bar) may be due to one of the following causes/reasons:

1. If the Web page is new, Google yet to be indexed.
2. If the Web page is indexed by Google, but it has to be ranked.
3. Web page was indexed by Google long back, but it is recognized as a supplemental (added to something else to make it complete) page.
4. If Web page whole website expelled by Google.

Supplemental Result is a URL retained/resided on Google's secondary database which contains pages of petty importance, as it can be scaled /measured by Google's Page Rank algorithm. Google used to place as "Supplemental Result" label at the bottom of a search result to represent that it is on the supplemental index; But it is very difficult to tell/predict whether it is on the supplemental index or the main one[11].

Page Rank for benign web sites ranging from 0 to 9 on scaling. Basically, we employed around 240 benign URLs and 240 malicious/harmful URL sites for the scheme/plot. It concluded from the graph represents the Page Rank is pretty good enough/So high for benign URLs as compared to phishing websites. One important thing newly registered website. Suppose we do the Page Rank check /analysis we would get 'N/A' (Not Available) message from the Page Rank Checker.

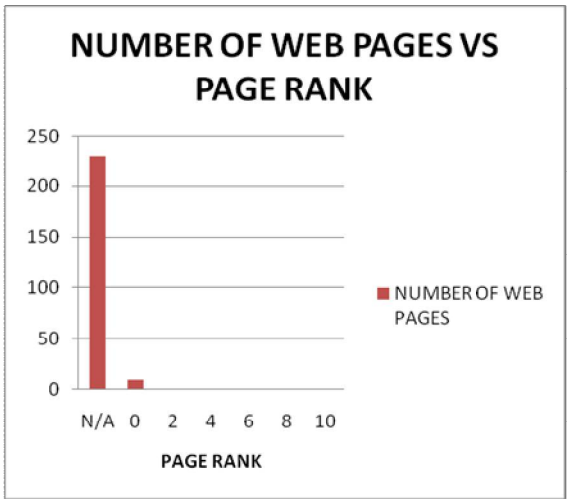

Figure 7 Number of phishing sites Vs PageRank

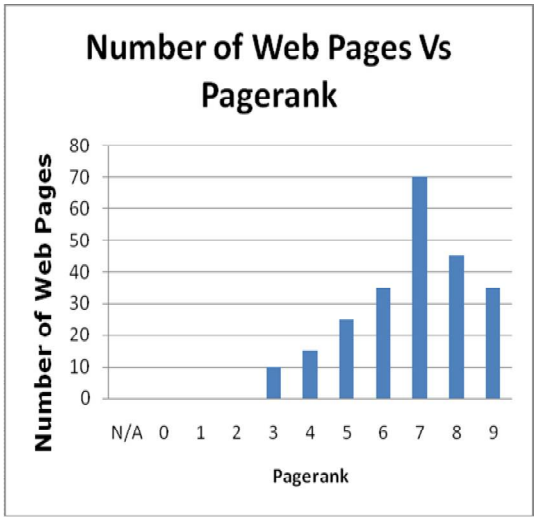

Figure8 Number of benign sites Vs Page Rank

We assessed the prepared URL feature dataset on using Naïve Bayes, J48 Decision Tree, k-NN, and SVM classified algorithms on WEKA. The percentage split is set to 60% i.e., 40% percentage of dataset was taken as training data and 60% a percentage as test data. The performance/accomplishment/process of the data is then assessed on the basis of Confusion matrix, Detection Accuracy, True Positive Rate and False Positive Rate. The efforts/result is tabulated on TABLE 1.

Test Options	Classifier	Confusion Matrix		Success Rate (%)	Error Rate(%)
Percentage split-60	Naïve Bayes	4438	3578	68.6	31.4
		260	3945		
	J48	7612	404	93.2	6.8
		428	3777		
	IBK	7042	974	88.3	11.7
		455	3750		
	SVM	7511	505	83.93	16.07
		1459	2746		
Percentage split-90	Naïve Bayes	1180	792	72.08	27.92
		61	1022		
	J48	1883	89	93.78	6.22
		101	982		
	IBK	1756	216	89.75	10.25
		97	986		
	SVM	1846	126	84.16	15.74
		355	728		

Table 1 Classifier performance WEKA

The analysis/evaluation of the dataset is done by using Java program also by setting the above said testing conditions and was tabulated in TABLE 2.

When we check/asses analyze the Success Rate in the analysis by WEKA and JAVA, it represented that there will be light changes/differences/variances in the values. The J48 Decision Tree that has the maximum Success Rate as compared to other selected classifying algorithms in WEKA. On using of only the lexical features, we have been able to attain a Detection Accuracy/Success rate of 93.2% for test split of 60%. When 90% of dataset is employed, we have gotten 93.78% Detection Accuracy. In

JAVA, by using Regression Tree, we have gotten 91.08% detection accuracy while using 60% of dataset for testing and 85.63% detection accuracy when employing/using 90% of data for testing.

Figure 9 shows a comparison of TP Rate, FP Rate and Detection Accuracy of SVM, Naïve Bayes, Regression Tree and k-NN classifiers. Figure 10 shows detection accuracy parameters of the classifiers with 60% and 90% test split.

TABLE2. Classifier performance JAVA

Test Options	**Classifier**	**Confusion Matrix**		**Success Rate (%)**	**Error Rate (%)**
Percentage split-60	Naïve Bayes	7281	303	74.2	25.8
		3633	4042		
	Regression Tree	1085	470	91.08	8.92
		1166	5839		
	KNN	1129	3025	79.55	20.45
		723	3284		
	SVM	9871	806	87.65	12.35
		1082	3531		
Percentage split-90	Naïve Bayes	13648	1018	83.5	16.5
		2764	5500		
	Regression Tree	15082	2999	85.63	14.37
		2951	8465		
	KNN	16451	5080	75.77	24.23
		1582	4384		
	SVM	16414	5848	74.48	25.52
		5	661		

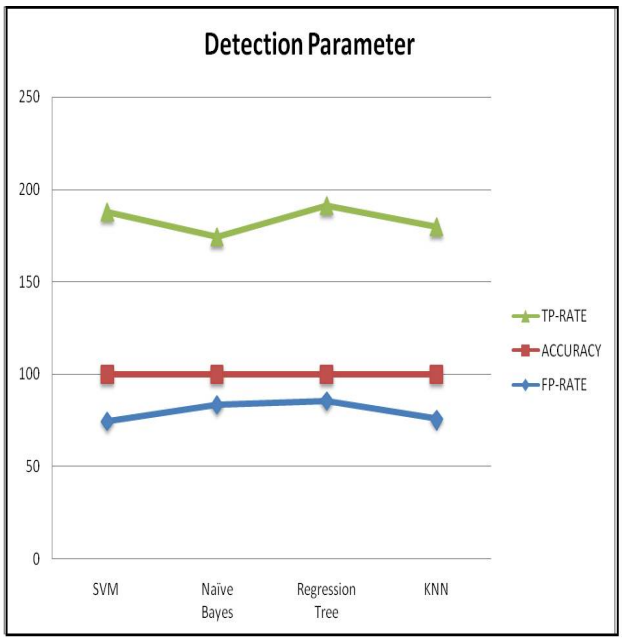

Figure9 Detection parameters

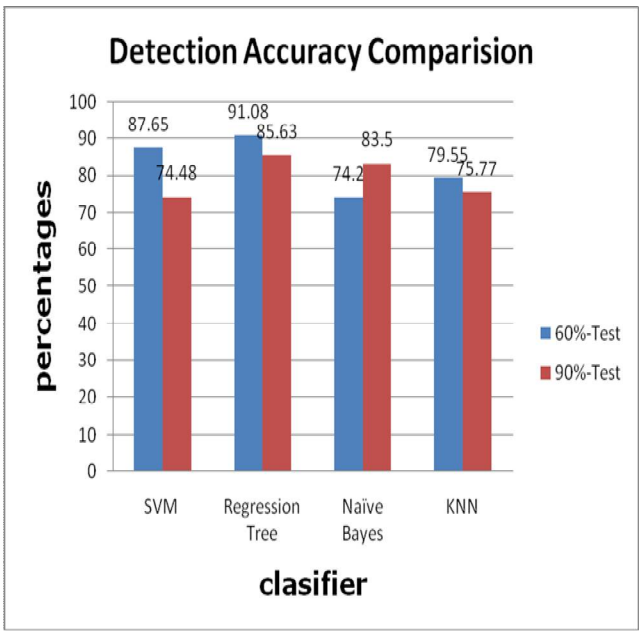

Figure10. Detection accuracy comparison

Apart from that another experiment was conducted to test whether an input URL is phishing or not. The URL is loaded into the Java program and retrieved /extracted URL features. A feature get will be created/generated .xls format. It can be used as test data and the classifier will make the decision whether 'Benign' or 'Phish' with its prescribed /exacted accuracy.

CONCLUSION

With the aid of page ranker under the page-based property will give an importance for the website to easily find out whether URL is phishing or benign and quantified the success rate and error rate by the different classifiers towards accurate performance on WEKA and JAVA. At eventually I have tested my various classifiers performance under the WEKA environment given different results and compared the results of classifiers among them. The J-48given best accurate success rate result is 93.2% under the percentage split-60. At the same time under the percentage split-90 the J-48 classifier has also been given best accurate result is 93.78% as compared with the other resultants which delivered by rest of classifiers.

I have also tested my various classifiers performance under the JAVA environment given different results and compared the results of classifiers among them. The Regression Tree given best accurate success rate result is 91.08% under the percentage split-60. At the same time under the percentage split-90 the Regression Tree classifier has also been given best accurate result is 85.63% as compared with the other resultants which delivered by rest of classifiers. Online machine learning algorithms will provide better learning methods as compared to batch-based learning mechanisms. As moving further, we will see in various aspects of online machine learning and collecting data to analyze the new strategies in phishing activities like fast changing DNS servers.

REFE RENCES

1. Anti-Phishing Working Group. Phishing Activity Trends Report. June, 2006. Available From URL: http://www.antiphishing.org/reports/apwg_report_june_06.pdf.
2. Binational Working Group , “Report on Phishing” , October 2006, Available From URL: https://www.justice.gov/archive/opa/docs/report_on_phishing.pdf.
3. Ma. J, Saul .L. K, Savage.S, and Voelker. G. M. “Beyond Blacklists: Learning to Detect Phishing Web Sites from Suspicious URLs”, Proc.of SIGKDD ’09.
4. Ma. J, Saul .L. K, Savage.S, and Voelker. G. M. “Learning to Detect Phishing URLs”, ACCM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, Article 30, Publication date: April 2011.
5. 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 Malcode (WORM), Alexandria, VA. WORM’ 07, November 2, 2007, Alexandria, Virginia, USA. Copyright 2007 ACM 978-1-59593-886-2/07/0011.
6. McGrath D. K., Gupta. M, “Behind Phishing: An Examination of Phisher Modi Operandi”, In Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET).
7. Sullivan Danny,”The Open Directory Project is closing” [2017 February 28]. Available From URL: https://searchengineland.com/rip-dmoz-open-directory-project-closing-270291.
8. J.-H. Li and S.-D. Wang, “PhishBox: An Approach for Phishing Validation and Detection,” Proc. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and

Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), The Orlando, FL, pp.557–564, Nov. 2017.

9. Fulham, Liz (May 10, 2018). "How & Why to Improve Your Alexa Ranking". Sales@Optimize. Archived from the original on November 16, 2017.
10. Southern, Matt (2016-04-19). "Google PageRank Officially Shuts its Doors to the Public". Search Engine Journal. Archived from the original on 2017-04-13.
11. Altman, Alon; Moshe Tennenholtz (2005). "Ranking Systems: The PageRank Axioms" (http://stanford.edu/ ~epsalon/pagerank.pdf) (PDF). Proceedings of the 6th ACM conference on Electronic commerce (EC-05). Vancouver, BC. Retrieved 2008-02-05.
12. Zhang W., Wang W., Zhang X., Shi H. (2015) Research on Privacy Protection of WHOIS Information in DNS. In: Park J., Stojmenovic I., Jeong H., Yi G. (eds) Computer Science and its Applications. Lecture Notes in Electrical Engineering, vol 330. Springer, Berlin, Heidelberg.
13. John Leyden (6 November 2009). "Controversial email blocklist SORBS sold". Retrieved 5 December 2009
14. Ward van Wanrooij1 , Aiko Pras., Filtering spam from bad neighborhoods, INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT Int. J. Network Mgmt 2010; 20: 433–444 Published online 15 October 2010 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/nem.753.
15. Gupta.Srishti, Ponnurangam Kumaraguru, “Emerging Phishing Trends and Effectiveness of the Anti-Phishing Landing Page” [cited 2014 June 14] Available from URL: http://precog.iiitd.edu.in/Publications_files/eCRS_Emerging_Gupta.pdf
16. SpamHaus (2014) Domain block list. Available From URL: http://www.spamhaus.org/dbl/
17. Maher Aburrous, Adel Khelifi “Phishing Detection Plug-In Toolbar, Using Intelligent Fuzzy-Classification Mining Techniques”, [JSCSE],2013, Vol. 3, No. 3, page No:54-61.
18. Google, Inc. Google Safe Browsing for Firefox. Accessed: June 13, 2006. Available From URL: http://www.google.com/tools/firefox/safebrowsing/.
19. Mark A. Hall, “Correlation-based Feature Selection forMachine Learning”, the university of waikato, Hamilton, NewZealand, 1999
20. Ricky Publico “How to Spot a Phishing Website“ 19 Apr 2017 Available from: https://www.globalsign.com/en-in/blog/how-to-spot-a-fake-website/

21. Schwartz, Barry. "Google Toolbar PageRank officially goes dark". Search Engine Land. Archived from the original on 2016-04-21
22. Thornton C, Hutter F, Hoos HH, Leyton-Brown K (2013). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 847–855.
23. Ian H. Witten; Eibe Frank; Len Trigg; Mark Hall; Geoffrey Holmes; Sally Jo Cunningham (1999). "Weka: Practical Machine LearninLearning Tools and Techniques with Java Implementations" .Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems. pp. 192–196. Retrieved 2007-06-26.
24. *Ian H. Witten; Eibe Frank; Mark A. Hall (2011).* "Data Mining: Practical machine learning tools and techniques, 3rd Edition". *Morgan Kaufmann, San Francisco*. Retrieved 2011-01-19.