International Journal of Electronics and Applied Research (IJEAR) **vol. 5, issue 2, Dec 2018**
Online (*http://eses.net.in/online_journal.html*) ISSN 2395-0064

Phishing URL detection system based on URL features using SVM

Bireswar Banik and Abhijit Sarma
Department of Computer Science,Gauhati University, Jalukbari, Assam India
** bireswarbanik02@gmail.com*
abhijit_gu@yahoo.com

Abstract

Phishing activities on the Internet are increasing day by day. It is an illicit attempt made by the attackers to steal personal information such as bank account details, login id, passwords etc. Many of the researchers proposed to detect phishing URLs by extracting features from the content of the web pages. But lots of time and space is required for this. This paper presents an approach to detect phishing URLs in an efficient way based on URL features only. For detecting the phishing URLs SVM classifier is used. The performances are evaluated for different size of datasets using different number of features. The results are compared with other machine learning classification techniques. The proposed system is able to detect phishing websites using URL features only with accuracy of 96.35%.

Key words*: Phishing URL; Lexical features; SVM; Intrusion Detection; Network Security; Machine Learning Techniques*

1. Introduction

Phishing is a powerful technique to mislead people either by giving a feeling that the site is legitimate or by showing some greedy approaches. The main strategy of phishing sites is to collect your personal information illegally like user ID, passwords, detail of your credit card, debit card or bank accounts [Sananse et al., 2015]. Now a day, it is affecting both financial and individual organizations a lot. Different policies are used by attackers to steal the information such as via email, advertisements, fake websites etc. [Gupta and Singhal, 2017].

As per as the APWG phishing trends report 2018 [https://www.antiphishing.org/resources/apwg-reports/], The total number of phish detected in first half of 2018 is 496,578 whereas the total number of phish detected in the second half of 2017 was 371,519. So, the number of phishing attacks is increasing

https://doi.org/10.33665/IJEAR.2018.v05i02.003

rapidly. The most targeted sectors of 2018 in phishing are payment sectors and SAAS/ webmail providers. As per report, a large number of websites having SSL certificate i.e. secure with HTTPS protocols are hosting phishing. The attackers tried to create HTTPS protocol for the phishing sites to make people believe that the site is legitimate. Cybercriminals of Brazil have taken the advantage of FIFA World Cup in 2018 for stealing and selling televisions. For these phishing attacks, Brazilian-commerce sites face a great loss from April to June [https://www.antiphishing.org/resources/apwg-reports/].

Many anti-phishing solutions are implemented to stop phishing activities, but still people are becoming victim of these attacks in their Daily life. It is very difficult to differentiate between the phishing and non-phishing/legitimate sites easily. In this work, we try to build a fast and efficient phishing URL detection system based on URLs only.

The rest of the paper is organized as follows. In section 2, we discuss some of the related works to detect phishing URLs. Section 3 presents the architecture of our proposed system, the detail of URL based feature selection and the performance metrics used in evaluation. The results of our experiments are analyzed in section 4.Conclusions are presented in section 5.

2. Related Works

Many of the researchers proposed different techniques for detecting phishing URLs. Some of them have maintained a list of domain name or IP addresses of previously detected phishing websites. A system named Phishnet is proposed [Prakash et al., 2010] where a blacklists of phishing URL was maintained. It will check whether IP address, hostname or the URL itself belong to that blacklist or not. They have also proposed five heuristics to detect phishing URLs. An approach of maintaining whitelist method is proposed in [Jain and Gupta, 2016] containing the domain name and corresponding IP address of legitimate sites instead of blacklist techniques. The system first checks whether a particular site is present in the list or not. If it is not, the system checks by extracting number of hyperlink contained in the site. If the number of hyperlink is NULL or zero or greater than certain threshold value, it is declared as phishing. Otherwise, it is declared as legitimate and added in whitelist.

The author of [Sonowal and Kuppusamy, 2017] provided a model known as PhiDMA containing 5 different layers. First layer is based on whitelist filter to check a particular site is legitimate or not. Second layer evaluates based on threshold value for five lexical features extracted from URL. If the site is greater

than that threshold, it is declared as phishing. Third and fourth layer is based on search engines results. The former one checks whether the given site's presence in the search engine list. The later one evaluates the percentage of similarity of the input URL with the search engines result using Longest Time Sequence and Edit list methods. If the value is greater than the threshold value, it has to go through the fifth level. The final layer evaluates the core based on a standard tool and compares it with the threshold.

Self Structured Neural Network for predict the phishing websites is proposed in [Mohammad et al., 2013]. They focus on improve the structure by using various number of neurons, different epochs and different number of hidden layers. 17 different features are used, along with the URL based feature, they also extract the features like whether the URL is present in WHOIS database or not, availability of DNS record, age of the URL, use of pop-up window, right click is disabled or not etc. For this, the content of the webpage is accessed. They classified the websites in to three categories viz. Legitimate, Suspicious and Phishy; Performance is evaluated for 1400 URLs only and the final structure is constructed using 500 epochs.

An association rule mining approach is proposed in [Jeeva and Rajsingh, 2016] to detect phishing and legitimate URLs. For this, 14 different features are extracted from URL. TF-IDF algorithm is used to find the words with high frequency in phishing URLs. Apriori and predictive apriori algorithm is used to generate the rules for detection. Rules generated by two algorithms are found different. The algorithms are compared using different number of inputs and apriori algorithm performed faster than predictive apriori algorithm. 93% of phishing URLs are determined correctly by apriori algorithm on a dataset of 1400 URLs.

A system to detect malicious URLs using Convolutional Neural Network (CNN) is proposed in [Abdi and Wenjuan, 2017]. First, they have checked whether the URL is in blacklist or not. If it is found, it is declared as malicious URL. Otherwise, goes for further evaluation. For this, two feature categories namely word2vec and TF-IDF are considered. Word2vec is used to convert each characters of the URL name into numbers numeric form. TF-IDF (Term frequency – inverse document frequency) algorithm is used to find the keywords whose frequency is more. The performance of SVM based TF-IDF is compared with LR and CNN based on word2vec feature.

Different works based on domain list, URL feature, features extracted from others websites like WHOIS, search engine etc., features from accessing the web content for detecting phishing URLs are discussed. The technique of maintaining a list of phishing or legitimate URL is not reliable as the

attackers may try different websites each time. Extracting features with the help of WHOIS database or different search engine is time consuming. Accessing the webpage content for large dataset of URLs requires lots of time and space. We have considered on the features extracted from URL only for developing our system.

3. Proposed Methodology

The design of our system is build up as shown in Figure 1. First, dataset of phishing and legitimate URLs are collected. Lexical features of these URLs are extracted. Feature selection method is used to find the important features only. This method provides ranking to each feature based on their contribution to detect phishing and non-phishing classes. The performance by taking different number of features is compared using different algorithms. The features of lower ranks are removed which are found to have low contribution to detect the classes. Then, the performances of various classification methods are analyzed for different numbers of URLs.

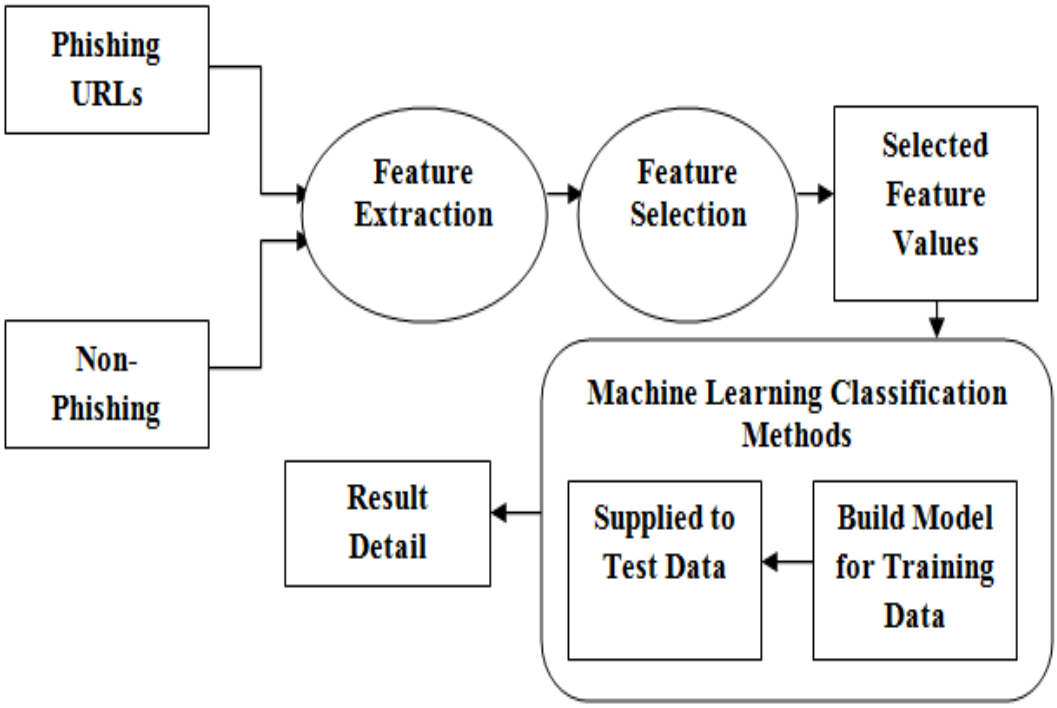

Figure 1: Block diagram of our proposed system

3.1 Features Selection

This paper focuses on only the lexical features extracted from URL. For this, we need to access only application header data from internet. Features extracted from content of the web pages are not considered because for retrieving the web content, we need to access the payload of the packets. But the packet payload in network is large in volume and is impracticable to process search large volume of traffic either real time or offline. That is why we concentrate on only the URL whose size is considerably small compare to the web pages. Different features used by other researchers are extracted from URLs. A feature selection technique is performed to select only the important features among them. In this technique, SVM is used to analyze the performances by taking all combinations of features. Rank of each feature is evaluated based on their ability to classify URLs correctly. The features along with their rank are shown in Table I.

Among all the features, length of the URLs [Jain and Gupta, 2017] shows highest rank as shown in Table I. Attackers use URL of larger length for hiding the doubtful part [Jeeva and Rajsingh, 2016] in URLs.In our dataset that the average length of a phishing URL is found as 101.3 whereas average length of a non-phishing URL is 34.5 only. The feature having second rank finds the ratio of total number of symbols (special characters) present in the URL to total number of alphanumeric characters present in the URL. The phishing URL usually contains more number of symbols. Phishing URLs usually contain specific characters which are not being commonly observed in case of non-phishing URLs. The next feature contains the number of suspicious characters present in the URL. Special characters like ‘`’,’%’,’#’,’^’,’$’,’&’,’-’,’*’,’:’ are considered as suspicious characters and their presence is more in phishing URLs. 4th ranked feature finds the ratio of length of the path of an URL to the length of the URL. If no path has been attached with the URL, it takes the value as zero. This ratio is comparatively greater than the other URLs [Bahnsen et al., 2017]. Rank 5 feature counts the number of suspicious words contained in URL. A list of suspicious words have been prepared from [Bahnsen et al., 2017; Astornio et al., 2016; Jeeva and Rajsingh, 2016; Jain and Gupta, 2017] and our observation. The tendency to contain these words in phishing sites is more. The list contain the words like ‘submit’, ‘secure’, ‘suspend’, ’confirm’, ’webscr’, ’account’, ‘login’, ‘signin’, ‘logon’, ‘cmd’, ‘update’, ‘wp’, ‘index’, ‘payment’, ‘home’, ‘paypal’, ‘webhostapp’, ‘dropbox’ etc. Our next feature finds the protocol used by the URL i.e. ‘HTTP’, ‘HTTPS’ and ‘FTP’ or some other protocols. As per APWG report

[https://www.antiphishing.org/resources/apwg-reports/], use of 'HTTPS' protocols in phishing URLs is increasing. In our dataset also, 2334 phishing URLs out of 12,799 URLs are phishing i.e. almost one sixth of phishing URLs have 'HTTPS' protocols. It is examined that the most of the URLs containing more than one or two hyphen is a phishing URL. It is used to make user fool that they are accessing legitimate URLs [Sonowal and Kuppusamy, 2017]. So, number of dash/ hyphen (-) is considered as one feature.8th ranked feature checks the last character of the URL is a symbol or not. In our survey, it is found that 649 URLs have special symbol other than slash as their last character and all of them are phishing URLs.

Table I: Features along with their Rank

Rank	Feature Name
1	URL length
2	Symbol to total character ratio
3	Number of suspicious symbols
4	Path length to URL ratio
5	Number of suspicious keywords
6	Protocols used
7	Number of dash(-)
8	Presence of symbol at last character
9	Redirection occurs
10	Presence of '@'
11	Number of slash (/)
12	Presence of IP address
13	Number of question mark
14	Number of subdomains
15	Presence of 'www'
16	Presence of 'http' word in URL
17	Presence of port number
18	Presence of Unicode characters

Feature having rank 9 checks whether redirection occurs or not in URL. The URL is redirected by double slashes ('//') in the URL. This is a binary feature to check the presence of '//' in URL where it ignores the cases like 'http://','https://'etc. protocols present in the starting of the URL. 10th rank feature is a binary feature which checks whether the symbol '@' present or not in URL. This feature is important because only the portion in the right side of '@' symbol in the URL is considered whereas the left side is ignored [Sonowal and Kuppusamy,2017; Mohammad et al., 2013]. Attackers use more number of slashes

to make user foolish so that suspicious URL of larger length look legitimate [Jeeva and Rajsingh, 2016]. 11th Rank feature counts the number of slash present in the URL where double slash (//) present together is not considered here. Our next feature checks whether any IP address is present in the URL or not. Presence of IP address instead of domain name is more in case of phishing URLs. Attackers try to steal confidential information through this type of URLs [Mohammad et al., 2013]. 12th and 13th number feature counts number of question mark (?) and number of subdomain present in the URL. 40% of the phishing URLs in our dataset have '?' symbol whereas only 2% of the non-phishing URLs have this symbol. A subdomain is the part of the domain name in the URL. Attackers add number of subdomain in URL to make user believe that the URL is legitimate [Jeeva and Rajsingh, 2016]. This feature counts the number of subdomain present in the URL. One of the most important features for detecting phishing URL is presence of 'www'. Rank 15 feature checks whether 'www' is present or not in the starting of the URL. It is found that the URL does not start with "www" has a high probability for being a phishing URL. Features having last three ranks are presence of 'http' keyword, presence of port number in URL and presence of Unicode character. Attackers may use keywords like 'http' or 'https' in the middle of the URLs to confuse the user and make them believe it as legitimate for the presence of these words. But in our dataset, we have found occurrence of 'http' in middle part of URL is once only. Some of the URLs may have port number using ':' symbol. So, we have checked whether any port number is present or not in the URL. As per [Jeeva and Rajsingh, 2016], phishing URLs may use Unicode symbol in the internet. So, we have checked whether any Unicode symbol is present or not. But we have not found presence of Unicode in URLs. Numeric value is stored in all features. Based on the extracted feature value, different classifiers are used to implement phishing URL detection system.

3.2 Performance Metrics

Parameters used for evaluating the performance of the stated models:

True Positive (TP): The number of phishing URLs that are classified as phishing URLs correctly by classifier.

True Negative (TN): The number of non-phishing URLs that are classified as non-phishing URLs correctly by the classifier.

False Positive (FP): The number of non-phishing URLs that are classified as phishing URLs incorrectly by the classifier.

False Negative (FN): The number of phishing URLs that are classified as non-phishing URLs correctly by the classifier.

TPR, TNR, FPR and FNR are evaluated in percentage (%) using above parameters. Accuracy, Recall (it is equivalent to TPR), Precision and F-Score are evaluated using following equations [Abutair and Belghith, 2017]:

$$\text{Accuracy, ACC} = \frac{(TP+TN)}{(TP+TN+FP+FN)} * 100\% \quad \text{---------- (1)}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} * 100\% \quad \text{---------- (2)}$$

$$\text{Precision} = \frac{TP}{(TP+FP)} * 100\% \quad \text{---------- (3)}$$

$$\text{F-Score} = 2*\left(\frac{Precision*Recall}{Precision+Recall}\right) \quad \text{---------- (4)}$$

4. Analysis & Results

For performance evaluation, a total of 32,652 URLs are collected. Out of this, 19,853 URLs are non-phishing URLs whereas 12,799 are phishing URLs. Phishing URLs are collected from phishtank [http://www.phishtank.com]. The dataset contains phish-id, phishing URL, submitted URL, submission time, verification time and other attributes. We have only collected URL from it. Non-phishing URLs are collected from DMOZ directory [https://github.com/gr33ndata/dmoz-urlclassifier/] containing serial no., URL and category.

Performances are evaluated using SVM (Support Vector Machine) classifier. The basic idea of SVM classifier [Astornio et al., 2016; Vanhoenshoven et al., 2016; Sahoo et al., 2017] is to draw hyperplane(s) to separate the classes. Suppose, we have 'n' number of features in the dataset. Then, each data has to plot a point in n-dimensional space where each feature value is the particular coordinate. The algorithm starts by identifying a subset of training dataset known as support vectors. The main aim is to separate the support vectors of two different classes in an efficient way. In two dimensional spaces, SVM focuses on to draw line in order to achieve the maximum distance from the support vectors of each class and minimize the wrong occurrences in each side. But drawing linear hyperplane is not suitable in case of n-dimensional space where n is relatively higher. SVM functions are taken from kernel function f(x, y)

where its value is chosen based on the situation. In our case, SVM is used as a binary classifier to separate the data points of phishing and non-phishing classes. Performance of Support Vector Machine (SVM) is compared with Naïve Bayes classification (NB), Multi-layer Perception (MLP), Logistics Regression (LR) and Decision Tables (DT) machine learning classifiers. A total of 18 features are extracted from URL as described in section 3.1. Feature selection method is used to find the rank of the features based to their contribution to detect phishing and non-phishing classes. Features with their corresponding rank are shown in Table I.

The accuracy of each feature set F_i is evaluated where i states the number of features in the set and the range of i is 1 to 18 as shown in Table 1 i.e. F_1 contains only the highest rank feature, F_2 contains the top 2 features and so on. Accuracy using different number of features (feature set F_i) is evaluated as shown in Figure 2 to Figure 5 using SVM, Naïve Bayes, Logistic Regression and Decision Table classifiers respectively. This experiment is performed for records of total 32,652 URLs using 3-fold Cross-Validation technique. It has been observed that the accuracy of the feature sets is changing from F_1 to F_{15} and it does not change more from F_{16} onwards. SVM achieves the highest accuracy among all the classifiers by taking 15 number of features So, for further evaluation, only the top 15 features of Table I are considered.

Table III depicts the result for k Cross Validation folds (CV) where k is 3,5,10 for SVM classifier. It shows the TP, TN, FN, FP and accuracy for 3 fold cross validation. In this technique, the dataset is split in to k parts. At the time of splitting dataset, it should take care that the proportion of number of records of each class is approximately equal. Here, a record contains the feature values of each URL. Each classification model is run k times considering one part as testing dataset and remaining parts as training dataset each time. So, k different evaluation results are found and the final result is shown aggregating the result of all folds. Detail of fold wise evaluation of SVM for F_{15} feature set using 3 fold cross validation is shown in Table II. Our further evaluation is done using 10 fold cross validation technique.

Accuracy of different classifiers is evaluated for different size of dataset of 10000, 15000, 20000, 25000 and 30000 URLs for F_{15} set as in Table IV. It is found that the accuracy of almost all classifiers increased as the number of records increased. SVM is showing highest accuracy for all sizes of dataset. Performance of SVM classifier model is compared with four different machine learning supervised classification techniques viz. Naïve Bayes, Logistic Regression, Decision Tables and Multi-layer Perception using 10 cross validation folds for dataset of 32,652 URLs as in Table V. All of the

classifiers have very low FP value which shows that our selected features have high significance for detection of phishing and non-phishing URLs correctly. Comparison of TPR, Accuracy and F-score value of different classifiers is shown graphically in Figure 6. Among all of them, SVM has significantly higher accuracy and F-score value. The performance of our result is compared with some previous works based on URL features to detect phishing URLs. Table VI shows the comparison of accuracy / recognition rate of our method using SVM with [Agarwal and Mangal, 2016], [Jeeva and Rajsingh, 2016], [Sonowal and Kuppusamy,2017] and [Zouina and Outtaj, 2017]. Our method has taken comparatively a large dataset and shows highest accuracy and low FPR value. So, it can be considered as an efficient approach for detecting phishing URLs.

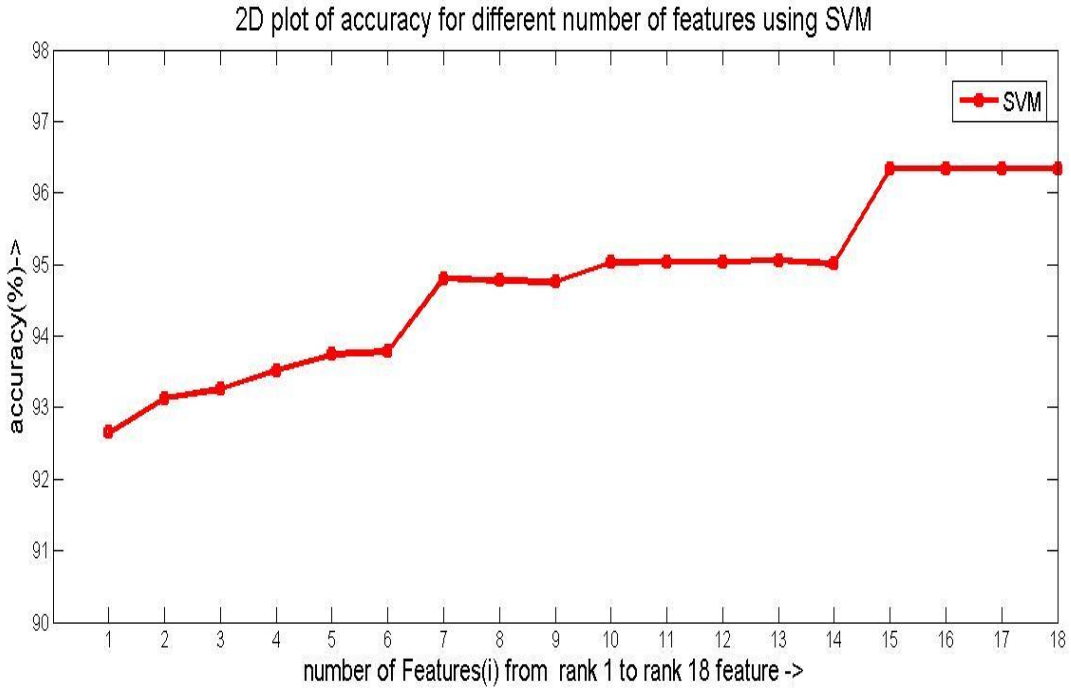

Figure 2: Accuracy of feature sets F1,F2,…,F18 for 3-fold cross validation using SVM

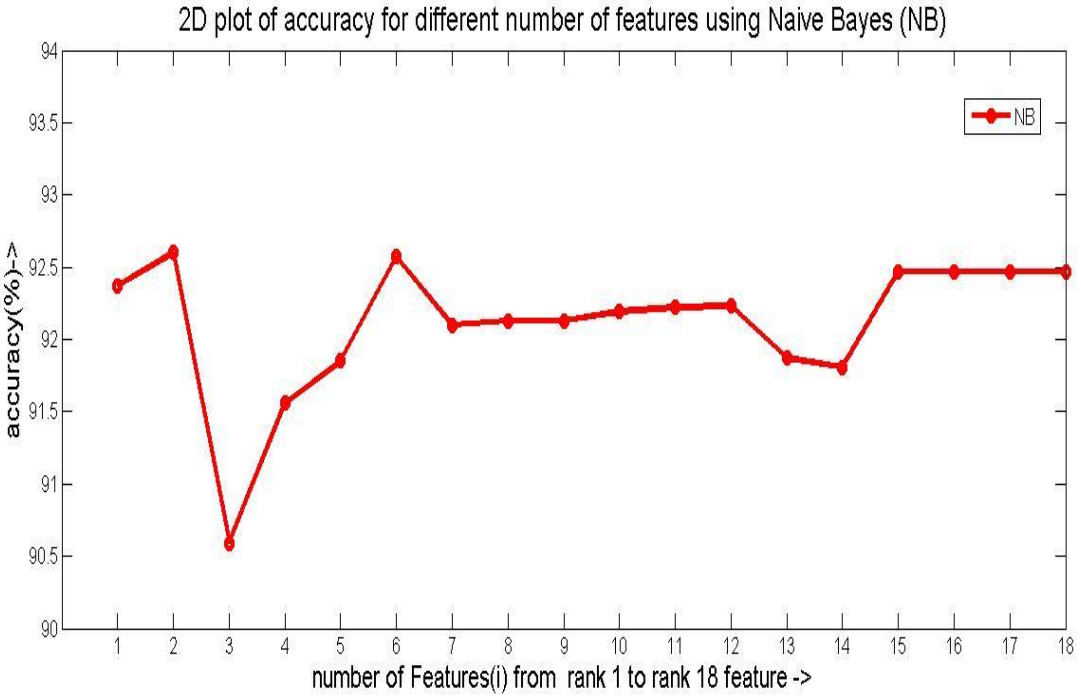

Figure 3: Accuracy of feature sets F1,F2,…,F18 for 3-fold cross validation using Naïve Bayes (NB)

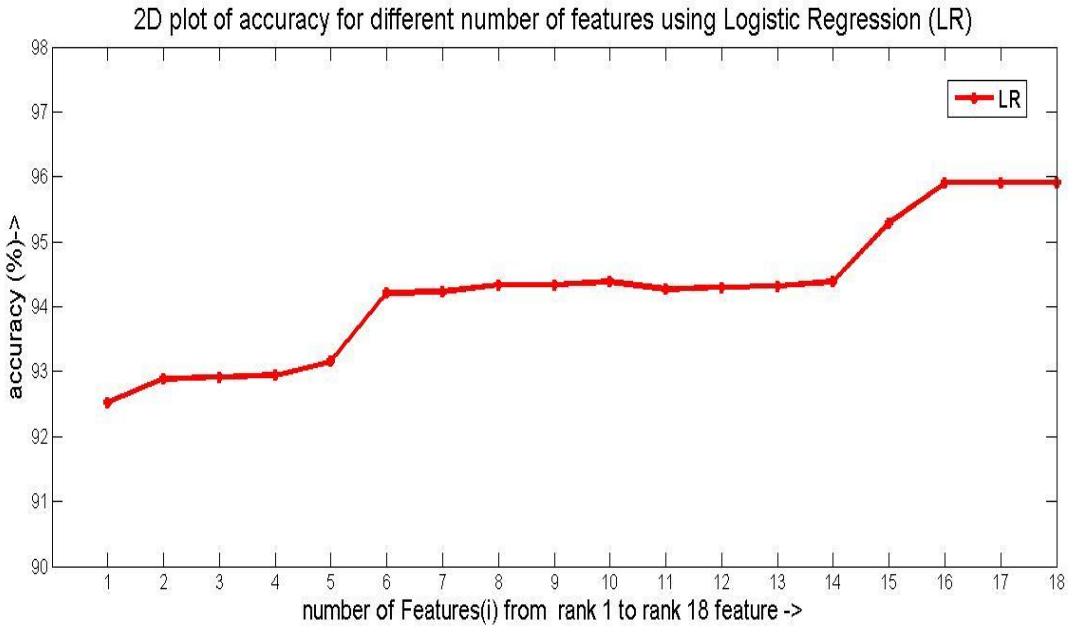

Figure 4: Accuracy of feature sets F1,F2,…,F18 for 3-fold cross validation using Logistic Regression (LR)

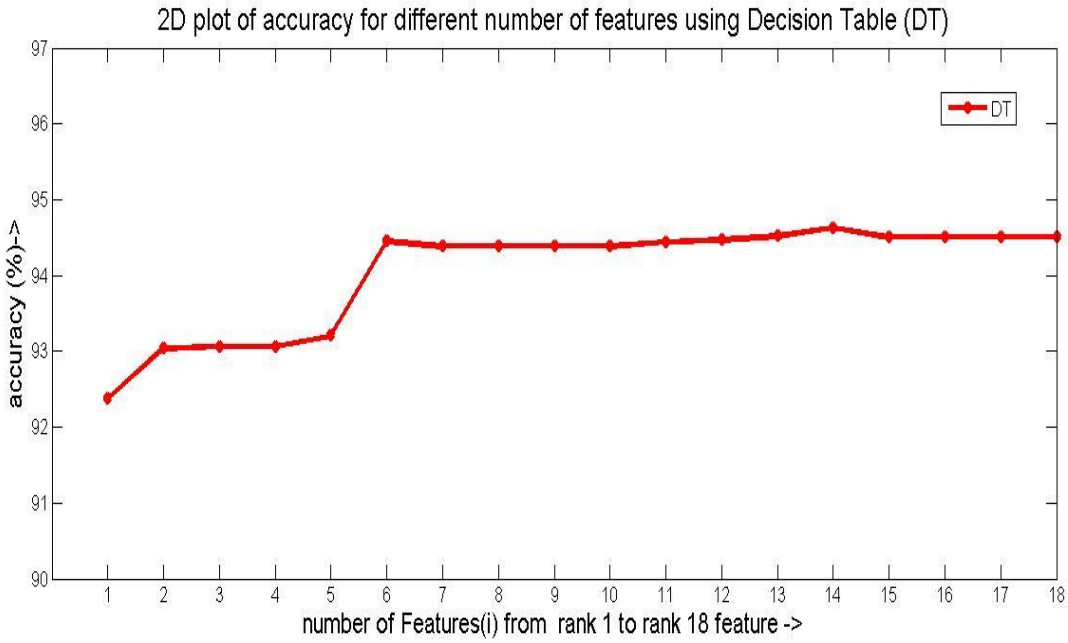

Figure 5: Accuracy of feature sets F1,F2,…,F18 for 3-fold cross validation using Decision Table (DT)

Table II: Fold wise performance using SVM classifier for 3-fold CV

Fold no	Total URLs	TP	TN	FP	FN		TPR (%)	TNR (%)	FPR (%)	FNR (%)	ACC (%)
1	10,884	4037	6445	173	229		94.6	97.4	2.6	5.4	96.30
2	10,884	4029	6443	175	237		94.4	97.4	2.6	5.6	96.21
3	10,884	4044	6457	160	223		94.7	97.5	2.5	5.3	96.48
Total	32,652	12110	19345	508	689	Average	94.6	97.4	2.6	5.4	96.33

Table III: Result using SVM feature on 15 features for CV=3, 5, 10

No of Cross Validation folds (CV)	TP	TN	FP	FN	ACC (%)
3	12110	19345	508	689	96.33
5	12108	19337	516	691	96.30
10	12120	19339	514	679	96.35

Table IV: Comparison of accuracy (%) of different classifiers for different size of dataset for top 15 features using 10 CV

DATASET SIZE (Number of URLs)	SVM	NB	LR	DT	MLP
10000	94.06	85.96	93.04	92.76	93.2
15000	95.65	89.25	94.79	94.82	95.09
20000	96.19	91.12	94.19	94.44	95.28
25000	96.26	92.02	95.92	95.75	96.08
30000	96.31	92.67	94.99	95.88	96.13

Table V: Comparison of performances of different classifiers using 10-Fold cross validation (CV)

Methods	TP	TN	FP	FN	ACC (%)	Precision (%)	Recall (%)
Support Vector Machine (SVM)	12120	19339	514	679	96.35	95.9	94.6
Naïve Bayes (NB)	10788	19404	449	2011	92.47	96.0	84.3
Logistic Regression (LR)	11644	19473	380	1155	95.29	96.8	91.0
Decision Tables (DT)	11996	19257	596	803	95.71	95.3	93.7
Multilayer Perception (MLP)	12089	19232	621	710	95.92	95.1	94.5

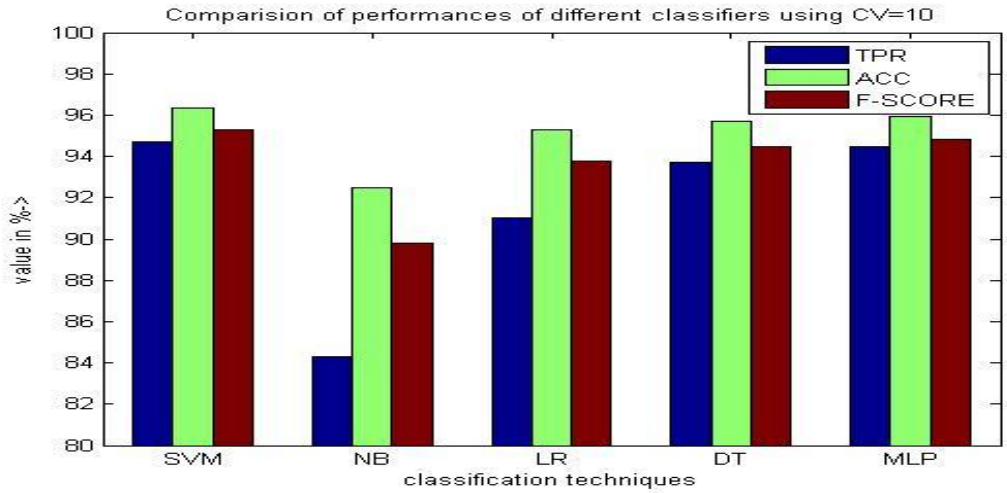

Figure 6: Comparison of TPR, ACC, F-score of different classifiers

Table VI: Comparison of our approach with other research works using URL feature

Approaches	Methods used	Recognition Rate (Accuracy)
Sonowal and Kuppusamy,2017	URL feature approach for 1662 URLs only	87.18%
	PhiDMA model	92.72%
Zouina and Outtaj, 2017	Using SVM for 2000 URLs only	95.80%
Agarwal and Mangal, 2016	Lexical features using Regression	91.5%
Jeeva and Rajsingh, 2016	Using association rule mining for 1400 URLs only	93%
Our approach	Using SVM for 32,652 URLs	96.35%

5. Conclusion

We explored the performance of different standard classification techniques to detect phishing and non-phishing URLs for different size of datasets. Our focus is to build a fast and efficient phishing URL detection approach. For this, lexical features extracted from the URLs are only considered. Accuracy for different size of datasets using different classifiers is evaluated. SVM has achieved the highest accuracy of 96.35% for 32,652 URLs using 10 cross validation fold. So, it gives an efficient result compare to other algorithms.

In future, we will try to implement other machine learning techniques using different feature sets and taking large datasets which can help to detect URLs more accurately by understanding the behavior of phishing URLs.

Reference

Abdi, F. D., & Wenjuan, L. Malicious URL Detection Using Convolutional Neural Network. International Journal of Computer Science, Engineering and Information Technology,7(6), 01-08. doi:10.5121/ijcseit.2017.7601,2017.

Abutair, H. Y., & Belghith, A. Using Case-Based Reasoning for Phishing Detection. Procedia Computer Science,109, 281-288. doi:10.1016/j.procs.2017.05.352, 2017.

Agarwal, P., & Mangal, D. A Novel Approach for Phishing URLs Detection. International Journal of Science and Research (IJSR), https://www.ijsr.net/ archive/v5i5/NOV163523.pdf, Volume 5, Issue 5, May 2016,1117 -1122, 2016.

Astorino, A., Chiarello, A., Gaudioso, M., & Piccolo, A. Malicious URL detection via spherical classification. Neural Computing and Applications,28(S1), 699-705. doi:10.1007/s00521-016-2374-9,2016.

Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & Gonzalez, F. A. Classifying phishing URLs using recurrent neural networks. 2017 APWG Symposium on Electronic Crime Research (eCrime). doi:10.1109/ecrime.2017.7945048, 2017.

Baunfire.com, S. B. (n.d.). APWG Phishing Attack Trends Reports. Retrieved from https://www.antiphishing.org/resources/apwg-reports/.

DMOZ open directory project, retrieved from https://github.com/gr33ndata/dmoz-urlclassifier/.
Gupta, S., & Singhal, A. Phishing URL detection by using artificial neural network with PSO. 2017 2nd International Conference on Telecommunication and Networks (TEL-NET). doi:10.1109/tel-net.2017.8343553, 2017.

Jain, A. K., & Gupta, B. B. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security,2016(1). doi:10.1186/s13635-016-0034-3, 2016.

Jain, A. K., & Gupta, B. B. Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems, 68(4), 687-700. doi:10.1007/s11235-017-0414-0, 2017.

Jeeva, S. C., & Rajsingh, E. B. Intelligent phishing url detection using association rule mining. Human-centric Computing and Information Sciences,6(1). doi:10.1186/s13673-016-0064-3, 2016.
Mohammad, R. M., Thabtah, F., & Mccluskey, L. Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications,25(2), 443-458. doi:10.1007/s00521-013-1490-z, 2013.

Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. PhishNet: Predictive Blacklisting to Detect Phishing Attacks. 2010 Proceedings IEEE INFOCOM. doi:10.1109/infcom.2010.5462216, 2010.
PhishTank, http://www.phishtank.com.

Ramesh, G., Gupta, J., & Gamya, P. Identification of phishing webpages and its target domains by analyzing the feign relationship. Journal of Information Security and Applications,35, 75-84. doi:10.1016/j.jisa.2017.06.001, 2017.

Sananse, B.E., E.R., M., Shahani, T., & Sarode, T.K. Phishing URL Detection : A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications, 123(13), 46-50. doi:10.5120/ijca2015905665, 2015.

Sahoo, D., Liu, C., Hoi, S.C.H. (2017). Malicious URL detection using machine learning: A survey. arXiv:1701.07179v2 [cs.LG], 2017.

Sonowal, G., & Kuppusamy, K.PhiDMA – A phishing detection model with multi-filter approach. Journal of King Saud University - Computer and Information Sciences. doi:10.1016/j.jksuci.2017.07.005, 2017.

Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., & Koppen, M. Detecting malicious URLs using machine learning techniques. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). doi:10.1109/ssci.2016.7850079, 2016.

Zouina, M., & Outtaj, B. A novel lightweight URL phishing detection system using SVM and similarity index. Human-centric Computing and Information Sciences,7(1). doi:10.1186/s13673-017-0098-1, 2017.