ISBN: 978-1-948012-17-1

Ethics and Information Technology (ETIT)

DOI: <http://doi.org/10.26480/etit.02.2020.158.163>

CrossMark

ENHANCED MODEL FOR DETECTION OF PHISHING URL USING MACHINE LEARNING

Siddhesh Masurkar^{*}, Vipul Dalal

Vidyalankar Institute of Technology, Mumbai, India.

^{*}Corresponding Author Email: siddhesh1199@gmail.com; vipul.dalal@vit.edu.in

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 25 October 2020
Accepted 26 November 2020
Available online 03 December 2020

ABSTRACT

Due to the increase in the use of the internet and the advancement of technology, more people are browsing on the internet. No need to mention that most people are not well aware of basic precautions, preventive measures that have to be taken while performing transactions over the net. This lack of knowledge among most people is well exploited by the phishers or attackers to acquire valuable private information from the user. Those security threats can cause user's mobile or computer to be hacked by the remote person usually called a hacker by executing some malware in the computer or mobile when the user clicks on the links purposely sent to the user. The phisher or Attacker or Hackers use sophisticated technics to cheat the user or to make scam in the virtual world by creating a fake site which looks almost similar to the original one. Therefore, if proper attention is not provided while filing user credentials or clicking on unknown links, those the security of the user can be compromised and its private information can be sent to the person who can take advantage of it. Phishing is new kind of an attack which is prevalent nowadays on the internet to jeopardize privacy of the user and subsequently of an organization. Therefore, many works are going on to detect and identify the phishing sites so user can be saved from entering his personal information and to prevent security attacks. But at the same time the way of designing phishing links are also evolving by improving content of the link in such way that the detection rate would be low. This paper is presenting a detection system which give emphasis on boosting algorithm. At the same time various other classifiers are used on same set of datasets to evaluate the performance of the Boosting algorithm. We implemented five classification algorithms and some unusual features from the URLs. The experimental results exhibit better accuracy and other parameters of the Boosting algorithm as compared to other traditional classifiers.

KEYWORDS

Phishing Websites, URL, features, Boosting, Machine learning algorithm, classification, Feature Selection.

1. INTRODUCTION

In today's world, Internet plays significant role in everyone's life. It has its convenience in respect of all daily routines like entertainment, communication, bank transaction, online shopping, etc. It also has its advantages and disadvantages. Every device is connected to internet so as to control it from remote place. Therefore in contemporary society, keeping personal information safe on such device or computer is more difficult for every person. Any person with bad intent and sufficient knowledge can gain the access to such device or computer or to spread virus or malware in any system causing.

There are many crimes happens on this virtual world. The main reason is that the attacker can be anonymous behind the curtain of this virtual world while carrying such bad activities. Phishing is a kind of attack wherein focused is on sensitive information and data entered by the user on a site which is designed in such a way that it disguises user as it is a trusted and real site which he used to surf. Once such information is obtained, it can be misused to make happen scams, make transactions. The APWG phishing Activities Trends Report say that in last semester approx. 4 million phishing sites are reported in year 2019. The most of the attacks were on SaaS and webmail which is 30.8%, payment site (19.8%) and

Financial Institution (19.04%). Fig 1 shows the steady increment in numbers of phishing sites detected in period of 2017 to 2019.

In recent years, many Machine Learning algorithm has made very good progress in identifying phishing links. In these method various features are extracted from the URL and one or more classification algorithms are used. Accordingly phishing sites are detected by training those extracted features with classification models. Earlier scenario was different wherein black list of phishing and whitelist were used. But since phishers are incorporating new technics for making phishing sites, those method were not sufficient and therefore employment of machine learning technique is provide better results than black list and white list of URLs.

The proposed approach extracts parameters of the URL given by the user. The extracted features are given to the Boosting algorithm. Those features are used by Boosting algorithm to find out a site requested by user whether input site is genuine or not. The main contribution of paper is summarized as below:

1) We have proposed the extraction of URL and selecting a important features from the URL data set to detect phishing, which is good phishing detection.

Quick Response Code	This paper was presented at	Access this article online
International Conference on Contemporary Issues in Computing (ICIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/etit.02.2020.158.163

2) We have proposed to use Boosting algorithm Gradient Boosting algorithm and Extreme Gradient Boosting algorithm (XGBoost) to test the data set.

3) We have examined and evaluated the performance of above said classifiers against some traditional classifiers namely naïve Bayes, Random forest and Support Vector Machine (SVM) with feature selection and with all features.

Figure 1: Numbers of Phishing sites detected

This paper is divided into sections as further: Section 2 shows the literature review, Section 3 shows the problem statement and Objective of the system. Section 4 discusses the proposed system which includes the details of working of the system and URL features and their conditions to create the dataset. Section 5 describes the methodology used in the system and Experimental result is shown in the Section 6. Section 7 gives conclusion and future work.

2. LITERATURE REVIEW

A study explains the system to detect the phishing URL using C4.5 decision tree algorithm. The built system is robust and precise and URL features are making it more accurate. They extracted 8 features and built a model which applied on 3 datasets. They attained average accuracy of 89.4% (MacHado et al., 2018).

Others show that the author has extracted 11 features from URL and web and used ELM technique. It has 2 stages. They extracted 11 features from 5 classes and 4 different classifiers namely adaboost, Random forest, bagging and SMO were used. First is building content classifier and second is combining text and other hybrid classifier. It has third party dependency to extract the features. The limitation associated with their approach is they lack real time demand of phishing detection (Zhang et al., 2017).

A projected the multi forward feed forward ANNs and their variants to detect the phishing sites. They have used 30 features which are extracted from the URLs. Those features are categorized into 4 classes and based on each class along with feature selection technique. It has been further observed that when all features are used for detection, it is providing best result among all which is 96.07%. But drawback of the system is it

consumes more time to find URL is phishing or not (El-Alfy, 2017).

Others projected URL identification strategy with the help of Rstudio tool to provide better analysis. They have obtained dataset from phishtank.com and are using only limited dataset. Out of 31 features they are using only 8 features. On these 8 features, information gain and ranker method is used to find important features. Dataset is divided on 70:30 ratio for training and testing respectively. The random forest classifier is used in the approach which is providing accuracy of 95% (Parekh et al., 2018).

A method of detection using heuristic features by authors used only ranking features which are obtained from the URL namely PageRank, AlexaRank, AlexReputation along with some basic features like Primary Domain, Sub Domain, Path Domain. They calculate 6 values of heuristic. They use dataset of 11660 phishing sites. Then they have calculated weights of these heuristics. The flow of the approach is divided into 6 phases which is implemented in PHP and MYSQL. 10 dataset is tested against the built model. According to result the threshold value of 0.4, 0.5 and 0.6 is having accuracy above 97% which is 97.16%, 97.10% and 97.09% respectively (Nguyen et al., 2014). In this paper they have used 27 features to build the classifier which is based on the fuzzy system. Their approaches mainly concentrate on the banking domain. They have used CBA and MCAR classifiers (Aburrous et al., 2010).

The authors in projected email classification model with mining concept which extract 23 keywords from the email. They have used 2500 emails out of which 1260 emails and others are valid emails. In text mining they have mined some keyword and using t-statistic methods the statistic values of those keyword is calculated. Using those features model is built using 7 classifiers and they all evaluated on accuracy, specificity, sensitivity with and without feature selection. It is observed that all model using all features is giving best accuracy of 98.12%. The Genetic Programming is yielding best result (Pandey et al., 2012).

The study presents the use of Bayesian classifier to detect phishing email. They are removing HTML tag and Stopword. After that they are tokenizing the URL and calculating the token frequency in URL. Accordingly Bayesian model is built and evaluated. That model is tested on the 3 datasets with highest accuracy is 96.46%. They used about 4700 emails as a dataset (Rathod et al., 2015).

A recent study used 9 features which consist of 5 email structure and 4 is based on sender behavior based. They have considered 500 phishing and 500 valid emails and SVM is used to build the model along with Support Vector Machine classifier to determine emails as genuine or not. The accomplished accuracy of model is 95.25% in results, but, its drawback is that comparatively very small training dataset (1000 emails with 50% spam and 50% ham) is used (Form et al., 2015).

Some study presents an approach to distinguish attacks of phishing email using NLP and classifiers. The text semantic analysis is used to find harmful content of the email. They have used 5009 phishing emails and 5000 valid emails (Peng et al., 2018).

Table 1: Comparison between Existing works

Existing work	Number of features	Approach used	Dataset used	Accuracy
MacHado et al. [1]	8	C4.5 Algorithm	3000 Phishing and Legitimate	89.4%
Zhang et al. [2]	11	AdaBoost, Bagging, Random Forest, and SMO	7788 URLs	98%
El-Alfy [3]	30	PNNs and K-medoids clustering	4898 Benign 6157 Phishing	96.07%
Parekh et al. [4]	31	Random Forest	-	95%
Aburrous et al. [6]	27	C4.5, JRip, PART, PRISM, CBA and MCAR	1006 Phishing	86.3%
Pandey et al. [7]	23	DT, LR, SVM, MLP, GMDH, PNN and GP	1260 phish 1240 valid	98.12%
Rathod et al. [8]	-	Bayesian Classifier	4700 data set	96.46%
Form et al. in [9]	9	C4.5 algo	500 spam 500 valid	95.25%
Peng et al. [10]	9	Seahoud and Netcraft	5009 phishing 5000 valid	95%

3. PROBLEM STATEMENT

The phishing attack is great concern to the online industries as per the report on Anti-Phishing Working. It has depicted the gradually increase in phishing attack in year of 2019. This number goes up to 1.2 million (APWG, 2020). These attacks are mostly targeting an industries which runs on

internet or which deals through internet compromising private information of the user. .com is the major domain which is used for phishing sites, but other for top level domains are .uk, .ru, .xyz, and .net. These five domains makes 75% of total phishing websites. The Business Email Compromise (BEC) scam is one of the major problems of 2019 (APWG, 2020). In this scam, industries are fooled by the attacker into

removing large amounts of money using a technique called as spear phishing techniques. According to RSAs Fraud Report, Phishing accounted for 37% of all fraud attacks in Quarter 4 of 2019. Phishing volume has increased 6% since last year. To tackle this threat, many algorithms have been developed which make use of Machine Learning. Generally in this technique, various features are extracted from the URL under various head which are used distinguish fake site from real one and accordingly model is trained. This model is used to find phishing sites.

In literatures, authors have used many methods and achieved the accuracy in the range of 86%-98% with varying sets of features ranging from 4 – 31 with feature selection or not as shown in table 1. Inclusion of large number of features may take large amount of time to determine which URL, could increase the complexity and shows the dependency on third party. Since life cycle of phishing site is small, there is a need of fast and intelligent detection mechanism.

Applying machine learning algorithm to detect Phishing Sites. The research questions we plan to answer in this work are:

1. How effective is the new approach compared to existing approaches?
2. What is the effect of using feature Selection with new approach?

3.1 Objectives

The fundamental objective of this system is to develop a Phishing detection mechanism. The system can work on limited features to provide with or without feature selection to provide better accuracy. Phishing detection mechanism intends to use information mining on URL dataset to aid the distinguish phishing URLs.

3.2 Specific Objectives

- Extract the URL features from the dataset and select best features.
- Build the model using Boosting algorithm and other widely used classifier.
- Evaluate the performance of Boosting classifiers on different parameters against traditional classifiers.

4. PROPOSED SYSTEM

Figure 2 show the data flow and process of the proposed system. The main focus in the proposed system in determining the spoofed web pages and any phishing attack. This is the hybrid approach in which Boosting algorithm is tested with all features and with limited sets of features selected by selection methods. The overview of the approach is shown in Figure 3.

4.1 Structure of The URL

URL is mainly known as site address using which any website can be visited. A URL is an address given to unique resources on the web. It is human-understandable text that was used to replace the IP addresses that computers understands and use to communicate with other systems in communication. A URL comprises of a protocol, domain name, and path (which contains the specific subfolder structure where a page is placed) and has the following basic format:

2								
http://www.mytimes.example.com:80/path/myfile.html?key1=value1#Somewhere								
1	3	4	5	6	7	8	9	
1 - Protocol				4 - Primary Domain		7 - Path		
2 - Domain				5 - TLD		8 - Parameters		
3 - Sub Domain				6 - Port		9 - Anchor		

Figure 2: Structure of the URL

4.2 Data Set

Figure 2 shows the general format of the URL which usually consist of *Protocol*, *Primary Domain*, *Domain*, *Subdomain*, *Top Level Domain*, *Port*, *path*, *Parameters* and *Anchor*. These parts of the URL gives valuable information using which spoofed URL can be created quickly. Therefore same information which is obtained from the URL can be used to identify whether URL is harmful or real. For this purpose 23 different features are extracted from the URL. We are incorporating those feature to build the model using tradition algorithm and Boosting algorithm. To create a dataset 10,000 phishing URLs are obtained from the phishtank.com from the year 2019 [20] and 3000 benign URLs are fetched from the Alexa and DOMZ (The Web Information Company). From those collected URLs 2 training dataset are created. The Description is shown in Table 2. These

training data set are given to the feature extractor to extract the 23 features form it.

Table 2: Description of the Training dataset

Data Set	Size of Data Set	No of Phish URL	No of Benign URL	Time to Train (in seconds)	Time to train (in minutes)
Data Set 1	10000	8000	2000	6657.12	110.952
Data Set 2	5000	4000	1000	3146.928	52.448

The 23 features which are extracted from the URL are majority lexical features except Rank country and Rank host. Those URL features are first separated by tokenization and parsing with the help of special character present in the URL. Further by processing it values corresponding to those features are obtained. Those features are as follow: Rank Host (**f1**), Rank Country (**f2**), Length of url, (**f3**), Length of host (**f4**), Numbers of Dots (**f5**), Executable file in url (**f6**), presence of '@' in url (**f7**), presence of '/' in url (**f8**), presence of '-' in url (**f9**), Phishing TLD I url (**f10**), presence of http in path (**f11**), shortening services used (**f12**), Average length of the token (**f13**), Numbers of token in url (**f14**), Largest token size (**f15**), Average length of domain token (**f16**), Domain token count (**f17**), Largest domain (**f18**), Average length of the path (**f19**), path token count (**f20**), Largest path token (**f21**), Presence of Security Sensitive words (**f22**), Presence of IP address (**f23**). Among these features, features f5, f6, f7, f8, f9, f10, f11, f12, f22 and f23 are having value either 0 or 1. Other features are having value greater than 0.

We have used 3 sets of testing data sets. Each data set comprises of 1000 url each. The features of those url are obtained by the feature extractor. Details of those testing data set is given in Table 3 herein below.

Table 3: Description of the Testing dataset

Data Set	Size of Data Set	No of Phish URL	No of Benign URL	Time to Train (in seconds)	Time to train (in minutes)
Data Set 1	1000	800	200	615.318	10.255
Data Set 2	1000	850	150	623.444	10.39
Data Set 3	1000	900	100	645.603	10.76

4.3 Approach of The System

Figure 3 show the data flow of our approach. It requires training set and testing set which is obtained as stated above.

Figure 3: Approach of the Proposed System

Both training data set and testing data set is tokenized first by string operation in feature extractor. Then all those tokens are processed and f3-f23 features are obtained. Features f1 and f2 are obtained from the alexa site.

The relation of features of phishing and benign sites are shown in Fig 4 (a) to (h) It shows that the differentiation between phishing and legit sites. Upon the values of these 2 types of url it is easy to train the model. Once features are extracted, those features are saved in their respective .CSV files.

Figure 4: Relation between phishing and benign urls

Based on all 23 features, RF, NB, SVM, GBM, XGBOOST models are trained. Then Feature selection initial 23 features are processed and value of importance of those features is calculated. We have considered top 10 features after feature selection and using those 10 features model is built. Fig 5 illustrate the some features with their relative value of importance. After models are built, test data set 1, 2 and 3 are tested with all the models and models predict whether URL is phishing or not. The result of the experiment is discussed in Section 6.

Features	Importance Value
rank_host	5.58E+08
Length_of_url	6.04E+04
largest_path	2.24E+04
avg_path_token	1.25E+04
largest_token	1.21E+04
path_token_count	8.02E+03
token_count	5.25E+03

Figure 5: Importance of top ten features

5. METHODOLOGY

5.1 Boosting

In boosting, initially the dataset is classified into several subsets. The base learner model is trained with one of the subset to get the outcome the classification. If the elements which were incorrectly classified by the previous base learner, then are passed to the next base learner model. This process continues depends on the base learners created. Then, the ensemble process enhances their exhibition by consolidating the feeble models together utilizing a cost work. Boosting Technique works on:

- Gradient Boosting
- XgBoost

5.2 Feature Selection

Filter methods select features based on a measure of the performance regardless of the used data modeling algorithm. Once best features are identified, the classification algorithms can use them. Filter methods put a rank for individual features or evaluate entire feature subsets. The developed measures for feature filtering can be classified into: information, distance, consistency, similarity, and statistical measures. Feature selection can be performed in a univariate or multivariate fashion (i.e. performing feature selection based on a single feature at a time, or considering subsets of features at a time).

The most common approach to univariate feature selection, which has been used extensively for years in the context of binary classification problems, is the simple t-test used to measure the degree of feature expression difference between two types of samples (groups). The basic idea with univariate feature selection is to create a ranked-list of feature based on their individual scores.

6. EXPERIMENTAL RESULT

Datasets of phishing sites and legitimate sites are collected from publically available sources from year of 2019. We have used Support Machine Vector (SVM), Random Forest, and Naive Bays along with Gradient Boosting and XG Boosting algorithm. Again the performance of classifiers is tested some selected features which obtained from features selection technique (The Web Information Company; DMOZ; Phish Tank).

To evaluate and compare the performance of model of Boosting algorithm with other model, the following evaluation matrix is used, Accuracy, Precision, recall and f - score. Accuracy measures the ratio of websites which are correctly predicted. Precision measures the fraction of websites correctly predicted as phishing. Recall metric measures the fraction of phishing websites identified by the model.

Table 3: Confusion Matrix		
	Predicted Positive Class	Predicted Negative Class
Actual positive class	TP (True Positive)	FN (False Negative)
Actual negative class	FP (False Positive)	TN (True Negative)

Table 3 shows the confusion matrix in which TP (True Positive) is a case in which a classification model correctly finds a website as phishing, TN (True Negative) is a case where a website is wrongly classified as Valid URL. FP (False Positive) is a case where a website is wrongly classified as phishing and lastly FN (False negative) is when the model wrongly classified a website as valid URL while it is actually phishing. The mathematical equations of the performance metrics are given below respectively.

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

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

We conducted 2 experiments in which 3 testing datasets of 1000 URLs each are tested with all the classifiers namely NB, SVM, RF, GBM and XGBOOST which are trained with 5000 and 10000 URLs.

Table 4 and 5 shows the result of the performance of those classifiers based on various parameters.

Table 4: Performance evaluation of Classifiers with all features on 5000 URLs

Classifier	Dataset	Accuracy	Precision	Error	TP	TN	FN	FP	FPR	Specificity	Time
NAIVE	1	0.363	0.97	0.63	168	195	632	5	0.025	0.975	0.0468
NAIVE	2	0.51	0.99	0.48	369	149	481	1	0.006	0.99	0.0311
NAIVE	3	0.43	1	0.56	332	100	568	0	0	1	0.015
SVM	1	0.8	0.8	0.2	800	0	0	200	1	0	0.561
SVM	2	0.85	0.85	0.15	850	0	0	150	1	0	0.577
SVM	3	0.9	0.9	0.1	900	0	0	100	1	0	0.639
RF	1	0.946	0.93	0.054	800	146	0	54	0.27	0.73	0.219
RF	2	0.968	0.964	0.032	849	119	1	31	0.206	0.793	0.217
RF	3	0.974	0.972	0.026	899	75	1	25	0.25	0.75	0.207
GBM	1	0.994	0.99	0.006	800	194	0	6	0.03	0.97	0.046
GBM	2	0.985	0.98	0.015	845	140	5	10	0.06	0.93	0.0311
GBM	3	0.993	0.99	0.007	894	99	6	1	0.01	0.99	0.0311
XGBOOST	1	0.998	0.99	0.002	800	198	0	2	0.01	0.99	0.068
XGBOOST	2	0.984	0.98	0.016	843	142	7	8	0.06	0.94	0.033
XGBOOST	3	0.997	0.996	0.003	898	99	2	1	0.01	0.99	0.028

Table 5: Performance evaluation of Classifiers with all features on 10000 URLs

Classifier	Dataset	Accuracy	Precision	Recall	Error	TP	TN	FN	FP	FPR	Specificity	Time
NAIVE	1	0.517	0.984	0.402	0.483	322	195	478	5	0.025	0.975	0.0155
NAIVE	2	0.61	0.99	0.542	0.39	461	149	389	1	0.006	0.993	0.0155
NAIVE	3	0.553	1	0.5	0.447	453	100	447	0	0	1	0.0311
SVM	1	0.801	0.8	1	0.199	800	1	0	199	0.995	0.005	1.076
SVM	2	0.852	0.851	1	0.148	850	2	0	148	0.986	0.013	1.0138
SVM	3	0.9	0.9	1	0.1	900	0	0	100	1	0	0.986
RF	1	0.992	0.99	1	0.008	800	192	0	8	0.04	0.96	0.207
RF	2	0.973	0.986	0.98	0.027	834	139	16	11	0.0733	0.926	0.208
RF	3	0.974	0.99	0.97	0.026	875	99	25	1	0.1	0.99	0.203
GBM	1	0.998	0.99	1	0.002	800	198	0	2	0.01	0.99	0.0311
GBM	2	0.985	0.99	0.99	0.015	842	143	8	7	0.04	0.95	0.0312
GBM	3	0.991	0.99	0.99	0.009	892	99	8	1	0.01	0.99	0.0156
XGBOOST	1	0.998	0.997	1	0.002	800	198	0	2	0.01	0.99	0.027
XGBOOST	2	0.989	0.995	0.991	0.011	843	146	7	4	0.026	0.9733	0.031
XGBOOST	3	0.997	0.998	0.997	0.003	898	99	2	1	0.01	0.99	0.081

It is shown from the Table 4 and 5 that with increase in URL instances the accuracy of all the model is increased. Among those XGBoost algorithm is giving best accuracy reaching upto 99.8%. The time to test the dataset is significantly decreased with XGBoost algorithm. Figure 6 (a) and 6(b) describe the accuracy for RF, NB, SVM, GBM and XGBOOST with all features tested against 5000 and 10000 URLs and Time required to identify phishing URL or not.

simultaneously offering good accuracy with slight increase in computation time.

Figure 6: Accuracy for all classifiers with all feature

The time required for all classifier is also shown in Figure 7(a) and 7(b) which describe, though Naive Bays taking less time but XGBoost is

Figure 7: Time required for each classifier to produce result

With feature selection method, XGBoost is providing very good accuracy upto 99.9% accuracy in lesser time.. It further shows that in both cases the XGBOOST classifier is providing best accuracy among all other classifiers. Along with accuracy the precision is also consistent with Boosting algorithm. Therefore results have shown that the classification model with XGBOOST whether large set of features or with minimum features providing overall better result. Table 6 here below show that the present work outperform other model shown in the Table 1.

Table 6: Performance of Present work

Existing work	Number of features	Approach used	Dataset used	Accuracy
MacHado et al. [1]	8	C4.5 Algorithm	3000 Phishing and Legitimate	89.4%
Zhang et al. [2]	11	AdaBoost, Bagging, Random Forest, and SMO	7788 URLs	98%
El-Alfy [3]	30	PNNs and K-medoids clustering	4898 Benign 6157 Phishing	96.07%
Parekh et al. [4]	31	Random Forest	-	95%
Aburrous et al. [6]	27	C4.5, JRip, PART, PRISM, CBA and MCAR	1006 Phishing	86.3%
Pandey et al. [7]	23	DT, LR, SVM, MLP, GMDH, PNN and GP	1260 phish 1240 valid	98.12%
Rathod et al. [8]	-	Bayesian Classifier	4700 data set	96.46%
Form et al. in [9]	9	C4.5 algo	500 spam 500 valid	95.25%
Peng et al. [10]	9	Seahoud and Netcraft	5009 phishing 5000 valid	95%
Present work	7	XGBOOST	8000 Phishing 2000 Valid	99.8%

7. CONCLUSION

In this approach, our aim to analyses the performance of Boosting algorithm for distinguishing the genuine and valid urls. From Table 1, 4 and 5 it is clearly highlighted that the performance of Boosting algorithm is better than other classifiers. We also found that with feature selection, the performance of almost all classifiers is bit increased. In both situations XGBOOST is giving best performance and Naïve Bayes has given worst performance.

As a future work, we shall incorporate spam email detection with same technique. Again the performance of boosting algorithm will be tested on different category of features to improve the accuracy. The present method are not using other features like html page features, event based features, domain features in phishing URL detection. Though these features are also important group to be considered, out future work is based upon incorporation of those remaining features.

REFERENCES

- Aburrous M., Hossain M. A., Dahal K., and Thabtah F., 2010. Predicting phishing websites using classification mining techniques with experimental case studies, in Proc. 7th Int. Conf. Inf. Technol. 176–181.
- APWG report available at: https://docs.apwg.org/reports/apwg_trends_report_q4_2019.pdf. Accessed 24 April 2020.
- APWG report available at: https://docs.apwg.org/reports/apwg_trends_report_q3_2019.pdf. Accessed 24 April 2020.
- DMOZ Open Directory Project. Online available at: <http://www.dmoz.org>
- El-Alfy E-SM. 2017 Detection of phishing websites based on probabilistic neural networks and K-Medoids clustering. Comput J.
- Form L.M, Chiew K.L, Szeand S.N, Tiong W.K, 2015, Phishing Email Detection Technique by using Hybrid Features, IT in Asia (CITA), 9th International Conference.
- Huang, H., Qian, L. and Wang, Y. 2012. 'A SVM based technique to detect phishing URLs', International Technology Journal, 11(7), 921–925.
- Li, Y., Xiao, R., Feng, J. and Zhao, L. 2013. 'A semi-supervised learning approach for detection of phishing webpages', Optic, 124(23), 6027–6033.
- MacHado L. and Gadge J., 2018, Phishing Sites Detection Based on C4.5 Decision Tree Algorithm, in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 1–5.
- Nguyen L. A. T., To B. L., Nguyen H. K., and Nguyen M. H. , 2014, A novel approach for phishing detection using URL-based heuristic, 2014 Int. Conf. Comput. Manag. Telecommun. ComManTel 2014, 298–303.
- Ozarkar P, & Patwardhan M, 2013. Efficient Spam Classification by Appropriate Feature Selection, International Journal of Computer Engineering and Technology (IJCTET), ISSN 0976 –6375(Online) 4(3).
- Pandey M. and Ravi V., 2012. Detecting phishing e-mails using Text and Data mining, IEEE International Conference on Computational Intelligence and Computing Research 2012
- Parekh S, Kotak S, Parekh D, Sankhe S. 2018. A new method for Detection of Phishing Websites: URL Detection Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2
- Parekh S., Parikh D., Kotak S., and Sankhe P. S., 2018, A New Method for Detection of Phishing Websites: URL Detection, in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 949–952
- Pattewar T.M, Rathod S.B, 2015, A Comparative Performance Evaluation of Content Based Spam and Malicious URL Detection in E-mail, IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS).
- Peng T., Harris I., and Sawa Y., 2018, Detecting Phishing Attacks Using Natural Language Processing and Machine Learning, Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, 300–301.
- Phish Tank. Online available at: <https://www.phishtank.com/>
- Rathod S.B., Pattewar T.M, 2015, Content Based Spam Detection in Email using Bayesian Classifier, IEEE ICCSP conference.
- Tak G.K and Ojha G., 2013. Multi-Level Parsing Based Approach against Phishing Attacks with the Help of Knowledge Bases, International Journal of Network Security & Its Applications (IJNSA), 5(6).
- The Web Information Company. Online available at: www.alexa.com
- Zhang X., Zeng Y., Jin X., Yan Z., and Geng G. , 2017, Boosting the Phishing Detection Performance by Semantic Analysis.
- Zhang Y., Hong J. I., and Cranor L. F., 2007, Cantina: a content-based approach to detecting phishing web sites, in the 16th international conference on World Wide Web, 639–648.