A Novel Machine Learning Approach to Detect Phishing Websites

IshantTyagi, Jatin Shad, Shubham Sharma
Deptt.of CSE&IT
Jaypee Institute of Information Technology
NOIDA, India
ishanttyagi11@gmail.com, jatinshad@gmail.com,
shubham15104026@gmail.com,

Siddharth Gaur, Gagandeep Kaur
Deptt.of CSE&IT
Jaypee Institute of Information Technology
NOIDA, India
siddharthgaur1996@gmail.com.
gagandeep.kaur@jiit.ac.in

***Abstract*—Phishing can be described as a way by which someone may try to steal some personal and important information like login id's, passwords, and details of credit/debit cards, for wrong reasons, by appearing as a trusted body. Many websites, which look perfectly legitimate to us, can be phishing and could well be the reason for various online frauds. These phishing websites may try to obtain our important information through many ways, for example: phone calls, messages, and pop up windows. So, the need of the hour is to secure information that is sent online and one concrete way of doing so is by countering these phishing attacks. This paper is focused on various Machine Learning algorithms aimed at predicting whether a website is phishing or legitimate. Machine learning solutions are able to detect zero hour phishing attacks and they are better at handling new types of phishing attacks, so they are preferred. In our implementation, we managed an accuracy of 98.4% in prediction a website to be phishing or legitimate.**

Keywords—phishing, R, machine learning algorithms

I. INTRODUCTION

Internet has tremendously changed the way we work and communicate with each other. There are applications like e-mail, file transfer, voice communication, You Tube etc. that are available for users to use. But with its humongous success has come its weaknesses and vulnerabilities. The protocols and applications responsible for its success are being exploited by malicious users and hackers for gaining limelight. Phishing websites is one such area where administrators need new techniques and algorithms to protect naïve users from getting exploited. Phishing is an attempt of fraud aimed at stealing our information, which is mostly done by emails. The ideal way to save ourselves from these phishing attacks is by observing such an attack. These phishing emails mostly come from trusted sources and try to retrieve our valuable information, for instance our passwords, bank details or even SSN. Many a times, these attacks come from sites where we have not even made any type of account. The procedure followed by phishers includes us reaching their website through the means of an email. In those emails, they make us click on a certain link that directs us to their websites. Asking for personal information is something that legitimate websites would hardly do. The looks of these phishing websites are quite similar to their respective legitimate ones and the only distinguishing factor is their URLs. Various initiations appearing from social websites, banks and online payment portals are used to deceive users. These phishing emails mostly contain links to websites that are affected with malware. Some of the ways to tackle these phishing attacks include generating awareness among people and training the users.

II. RELATED WORK

The authors in this paper [1] talk about two approaches of phishing detection in general: Blacklists and Machine Learning. In Blacklists, some blacklist providers were mentioned and in machine learning Google Safe Browsing API, DNS Based Blacklist, and Phisnet. Also mentioned were some of algorithms with a brief introduction of application of machine learning in phishing detection. A couple of techniques (Cybersquatting and typosquatting) were briefly discussed which are often used by phishers for URL manipulation. Then, some of the features of websites were considered which may classify the website as phishing or not. Although, there was no mention of any of results after application of machine learning algorithms and no clear methodology about feature extraction was mentioned.

The authors in this paper talk about neural network as the method for detecting phishing websites [2]. Supervised learning algorithms were used namely Adaline network, Backpropagation network, and Support Vector Machine. Those algorithms were applied independently, and also combinations of algorithms were applied. Data pre-processing, data cleaning, and feature extraction (of website) were the key points of the work done. About 15 features of these websites were extracted. Highest accuracy was observed by using Adaline network with Support Vector Machine. Although, we have considered as many as 30 attributes and their extraction was done through Python.

In this paper [3], the detection of phishing websites was done through some of the many features that one can extract about a URL [3]. By only extracting those features they are making a decision on a website to be phishing or legitimate. Then they have created a simple application where a user can enter a URL and find out whether that website is fake or real. The dataset used by them consists of only about 100 URL's taken from Phistank and Yahoo directory database. Out of those 100

978-1-5386-3045-7/18/$31.00 ©2018 IEEE

URL's, 59 were legitimate and 41 were phishing. They applied their methods on these 100 URL's only and obtained a result comprising of 68 legitimate URL's and 32 phishing ones. So, their accuracy rate was only 96%. Our accuracy was ahead of theirs at 98.4%.

In this research paper [4], the authors have used machine-learning algorithms to detect phishing websites using features from X.509 public key certificates. They have collected and tested certificates from all confirmed phishing websites associated with PhishTank entries, regardless of whether HTTPS was included in the listed URL. They have used features like *NotBefore*, *NotAfter*, *Date-Downloaded,Issuer, Subjectand Domain Name*, etc. They have extracted the features using Python. The machine learning algorithms, that they have used are – Decision Trees, Random Forest, Naïve Bayes, and logistic regression. Their best prediction came from random forest with accuracy of 95.5%.

In this paper [5], Phishtank and OpenPhish were the sources for phishing dataset. These datasets include columns such as phishing URL, target brand name, IP etc. In this paper, machine learning algorithms used are - J48, Support Vector Machine (SVM), and Logistic Regression (LR). J48 has the highest accuracy with accuracy of 96.96%. Our project is different from this research paper in ways like we have used different data set, different machine learning algorithms, different pre-processing of the data set. Our accuracy is better. Moreover we have tried to extract the features of any new URL by using Python for checking whether the new URL is Phishing or Legitimate.

The approach assesses the relatedness of words that compose a URL [6]. They leveraged search engine query data to establish relatedness between words and show that this is more suited to Internet vocabulary than existing methods. It first checks the obfuscation of URL which can be done with domain name, keywords, IP address and URL shortening. Secondly, they formed sets of words related & associated to registered domain & remaining part of URL. It then defines 12 features, derived from these sets and applied ML algorithms including SVM, Random tree & random forest. Worst accuracy is of SVM at 86.31% whereas best is of RF at 95.22%. Their shortcoming is that Wordnet only contains English dictionary words whereas internet has many languages and secondly average time per URL was 4.2 seconds which is quite large for real time usage.

[7] They proposed a method that makes prediction based on the features of URL and the ranking of site. Alexa Reputation was calculated using URL. Root mean squared error was also calculated to find accuracy of different values. The accuracy rate of this technique is 97.16% with a threshold of 0.4. Their shortcoming were the important features like status bar customization, submission of information, website forwarding that were neglected which may lead to wrong prediction.

[8] They extracted websites' URL features and analyzed subset based feature selection methods and classification algorithms for phishing websites detection. First they determined most targeted brand names and their legit URL via Google and their real phishing URLs from PhishTank website. The authors created matrix of 133 features and trimmed them using Feature Selection technique. Then two classification algorithms namely Naïve Bayes and Sequential Minimal Optimization (SMO) are used for classification to which the dataset was given as input. Best accuracy was observed using SMO algorithm was 95.39% whereas Naïve Bayes only gave accuracy of 88.17%. Their shortcoming was that the accuracy percentage is low.

III. Methodology

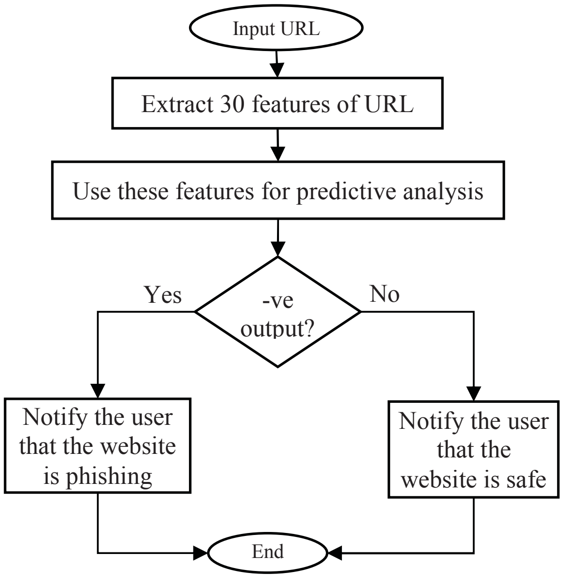
Input URL
Extract 30 features of URL
Use these features for predictive analysis
-ve output?
Yes
No
Notify the user that the website is phishing
Notify the user that the website is safe
End

Fig 1: Phishing detection methodology

A. Machine Learning Algorithms used

In this subsection, a brief overview of the machine-learning algorithms used for classification has been explained. In our experiments several well-known classification algorithms were tested. This was done using the open source programming language R. The algorithms have been ranked based on their overall performance. In this section, these algorithms are briefly described.

1. Decision Tree: Before adding a node (feature) to the decision tree, the algorithm calculates the information gain for all features according to its entropy. It then creates a single decision based on the feature with the highest value. Then the feature selection is iterated on the remaining features. It divides on the feature with the highest value. The value obtained by terminal is the mode of observations. Thus prediction for new data is made with that mode value. The depth of the tree is defined by the user.

2. Random Forest: Random Forest adds randomness to the generation of decision trees. Instead of relying on one single decision tree to cover the entire dataset and features, this approach selects features and training data randomly from the given sets and constructs a series of decision trees based on these randomly selected inputs. The output of Random Forest is then calculated by the outputs of the contained decision trees. For a new data, each tree gives a classification. The forest chooses the classification having the most votes.

3. Gradient Boosting (GBM): Boosting is applied along with other machine learning algorithms to improve the model fitting. It takes a lot of weak predictors;add weightage to them to makethem strong predictors. Weight is calculated on the basis of errors. It increases weight of missed classifications, so it converts weak learner to strong learner. Boosting with tress is known as Gradient Boosting.

4. Generalized Linear Model: In GLM, Logistic Regression is commonly used. Logistic Regression is a classification algorithm used to predict a binary outcome. In this a logit function is used. The fundamental equation for GLM is$g\big(E(y)\big) = a + bx1 + cx2 + \cdots$, where a, b, c are the coefficients or slope with respect to x1, x2,… This slope is computed by using the correlation formula. $a = cor(y, x) * \left(\frac{sd(y)}{sd(x)}\right)$, where cor is correlation, sd is standard deviation. Correlation between y and x was calculated using$cor(y, x) = \frac{cov(y,x)}{sd(x)*sd(y)}$, where cov is covariance. We also computed$cov(y, x) = \frac{\sum(x-mean(x))*(y-mean(y))}{n-1}$, where n is number of observations. Values of b & c are calculated similarly. In logistic regression, probability of outcome variable is determined. The logit function is established by using Probability of Success (P) and Probability of Failure (1-P).Since probability is always positive, the linear equation will be in exponential form i.e. $P = e^{a+bx1+cx2+\cdots}$. To make P less than 1, we do,P $=\frac{e^{a+bx1+cx2+\cdots}}{e^{a+bx1+cx2+\cdots}+1}$, where $P = \frac{e^{y}}{1+e^{y}}$ and y is computed as log $\left(\frac{p}{1-p}\right) = y$, wherelog $\left(\frac{p}{1-p}\right) = y$is link function.

5. Generalized Additive Model (GAM): This model works by combination of two different models and making it a single model. The best result was given by combination of GLM & Random Forest. At first, model fitting of both these algorithms is applied separately. Then GAM is applied on the combined result of these algorithms.

B. Pre-Processing and Cleaning of Dataset

1. *Variability Inflation Factor (VIF):* VIF is the increase in variance for the ithregressor compared to the ideal setting when it is orthogonal i.e. not related to others. This variability occurs due to high correlation of that variable along with other variables apart from result variable. The variables having high VIF value were removed from dataset which further enhanced the result prediction rate.

2. *Principal Component Analysis (PCA):* This technique combines highly correlated variables by using suitable combinations. Hence it creates new variables in dataset removing highly correlated variables. This newly formed dataset gave the best prediction for all algorithms.

3. *K-Nearest-Neighbors (KNN):* It is an instance based algorithm. For each unknown instance, its category is determined by a majority vote of the K training instances that are closest to that instance (based on the features). Our model calculates five neighboring instances to generate the predicted class (K=5). KNN algorithm was used on the dataset to extract the values of attributes which were difficult to find because of limited computing resources. Those attributes were: Number of links pointing to a page, Server Form Handler (SFH), andStatistical based features.

IV. Evaluation Methods

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

Accuracy is defined as the ratio of correct predictions to the total predictions (both correct and incorrect).

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

Precision can be defined as the ratio of correctly predicted positive observations to the total predicted positive observations.

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

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

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

Where TP is number of cases which were positive and were also predicted positive, TN is number of cases which were negative and were also predicted negative, FP is number of cases which were negative but predicted positive and FN is the number of cases which were positive but predicted negative.

V. Attributes Of Url

There were as many as 30 attributes of a website that were considered for detection purpose. Those extracted using python and used for prediction for a new URL.

A. Having IP address

By transforming a website's URL to an IP address, users can be fooled. An example is http://125.98.3.123/fake.html. As it is having IP address, it should make users suspicious provided they are aware of this fact. Sometimes, phishers change IP to hexadecimal code, for example: "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

B. Long URL length

Using long URL is a way of hiding doubtful part. For instance:
http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website. html
The URLs having length greater than 75 are phishing, while those having length between 54 and 75 are suspicious, and those having length less than 54 are legitimate.

C. Shortened URL

Phishers often shorten a URL on the "World Wide Web" i.e. making a website URL much shorter but it still leads to the required website. This can be done by using a "HTTP Redirect" on a domain name that is short, which directs to the webpage with long URL. An example is: "http://portal.hud.ac.uk/" which may be shortened to "bit.ly/19DXSk4". So, if a website shortens itself, then it can be considered as phishing.

D. Having @ symbol

If someone uses "@" symbol in the URL, then the browser tends to ignore everything before the "@" symbol and the actual address is the one which succeeds the "@" symbol.

E. Double slash redirecting

The presence of "//" implies that the user will be directed to another website. What phishers do is that the put the address of their malicious website beyond the original "//" and users they redirect the users to their required webpage. So, finding out the location of "//" can be useful to find out whether we are being redirected or not. If the position of "//" is at 6th or 7th position(HTTP of HTTPS), then we can be assured that we are not being redirected.

F. Prefix suffix

Phishers add prefix of suffix which are separated by a dash ("-") symbol in the domain name to mislead the users. Although, in reality, one can hardly find any "-" symbol in legitimate URLs.
For instance in the given weblinkhttp://www.Confirme-paypal.com/, phishers have added dash symbol between Confirme and paypal and its difficult for regular user to detect the trick.

G. Iframe tag

Another trick that phishers do is that they hide a webpage within a legitimate webpage by using "iframe" tag in the HTML script. So, phishers can make use of the "frameBorder" attribute which causes the browser to render a visual delineation.

H. Anchor tag

An anchor is a HTML element denoted by the <a> tag. If the percentage of anchor tag in URL is less than 31, then we call it phishing. If it is between 31 and 67, then it is categorize it as suspicious, and phishing otherwise.

I. Disabling right click

By disabling the right click feature on a website, phishers defy users of viewing the source code of various websites. In order to check this, one can search for event "event.button==2" in the source code of the website and check if the right click is disabled.

J. Links in <Meta>,<Script>, and <Link> tags

Legitimate websites use <Meta> tags to show metadata about the HTML document, <Script> tags are used to create a client side script, and <Link> tags to obtain other web resources. One can assume that these tags are linked to the same domain of the webpage.
So, if the percentage of "<Meta>", "<Script>", and "<Link>" tags is less than 17%, then we call our website legitimate. For it to be in category of suspicious, it must have its percentage less than 81 but at the same time greater than 17. If its percentage exceeds 81, then it falls in the category of phishing.

K. Age of domain

By taking advantage of the fact that a phishing website live for shorter durations of time, one can classify a website as phishing or not. The cutoff time for age of domain of a website is 6 months.

L. record

If the DNS record for the website is not found, the identity that is claimed is not recognized by the WHOIS database, then the website is categorized as phishing.

M. HTTPS with SSL

Even though it is important to have HTTPS certification for any website, but only having it does not confirm its legitimacy. Some other factors are also considered. One of them is: certificate issuer should be a trusted one, for example:"GeoTrust, Network Solutions, Thawte, Comodo, Doster and VeriSign". Another factor is that the age of certificate should be a minimum of two years as phishing websites are short lived.

N. Domain registration length

Domain registration length should be at least one year as we are aware of the fact that trustworthy domains are regularly paid for several years in advance.

O. Website traffic

This feature generally decides a website to be phishing or not on the basis of its popularity. We find the popularity of the website by finding its rank through the Alexa database. Legitimate websites ranked among the top 100,000. So, any website with rank greater than 100,000 will come under the banner of phishing.

P. Statistical based reports feature

This feature basically searches the domain name and IP address of the website and searches it among the "Top 10 Domains" as well as "Top 10 IPs". If any matches occur, the website is considered as phishing. The statistics were provided by PhishTank's statistical reports published in the years 2010 to 2012.

Q. Using non standard port

If a certain service (like HTTP) is up or down, then this feature comes into the picture. It is advised to merely open those ports only which are required. Some firewalls, Proxy and Network Address Translation (NAT) servers will, by default, bar almost every port and open the selected ones only. Opening all the ports puts user's information in danger as the attacker may run service of its choice.

R. Abnormal URL

WHOIS database is used for pulling this feature. The identity is part of URL, for an innocent website. So, in a website, if URL doesn't contain host name, then it is considered phishing, and else it is acceptable.

S. Sub Domain and Multi Sub Domains

This feature is based on the number of dots in the URL. For example: http://www.iitd.ac.in. Here the "in" is country-code Top Level Domain (CCTLD) .The "ac" part is an abbreviation for "academics", the joined "ac.in" is called a Second-Level Domain (SLD) and "iitd" is the real name of domain. At first we neglect (www.) part and then (CCTLD) (if present) from URL. At last, we count the remaining dots. If the count is greater than one, then the URL is marked as "suspicious" since it has one sub domain. Else if the count is greater than two, it is marked as "Phishing" as it will have multiple sub domains. Else it is legitimate if it has no sub domains.

T. Favicon

Many phishers use fake favicon but it can potentially expose them. A favicon is an icon affiliated with a particular webpage. Favicon is displayed as a pictorial reminder of website's identity in address bar by current users as graphical browsers and newsreaders. If there is a mismatch in the favicon of the domain and the URL in the address bar, then one can be assured that it is a phishing attempt.

U. Request URL

This feature basically checks whether the content requested on the website is from another website or the same website. SO, if the percentage of request URL is greater than 61%, then the website is phishing, if it is between 22 and 61, then suspicious, and legitimate otherwise.

V. Server Form Handler (SFH)

If the Sever Form Handlers are containing any empty string or "about: blank", they should be considered doubtful. Also, if the domain names of webpage and SFH are not matching, then the website might well be considered as phishing as it rarely happens that external domains handle submitted information.

W. Submitting Information to Email

We often submit our personal information to various web forms. User's private information can be averted to attacker's email. This may be done in two ways: first by using mail () in PHP and second being "mailto". So, if either of these twofunctions is used, then the website can be phishing.

X. Website Forwarding

The more a website is redirected, the more chances are of it being phishing. It has been observed that a legitimate website has been redirected only once, while phishing websites have been redirected at least 4 times.

Y. Status Bar Customization

Phishers often deceive us by showing a deceptive URL in the status bar by the help of JavaScript. For checking it, one may extract the source code of the webpage, and then check the status bar for any changes by the "onMouseOver" event.

Z. Using Pop-up Window

If a pop up window in website asks for private information, then this website might as well be a phishing one as usually, legitimate does not ask their users to submit any important piece of information through a pop up window.

AA. PageRank

PageRank is an algorithm to rank websites in search engine results by Google search. It ranges from 0 to 1. The better the PageRank value better is the webpage. It has been found that 95% phishing websites do not have PageRank, and the remaining 5% phishing websites have a PageRank upto 0.2. So, if the PageRank is less than 0.2 or does not exist for a website, then that website will come under the category of "phishing".

BB. Google Index

If a website is indexed by Google, then it is considered legitimate. Being indexed by Google here means the website is displayed on search results.

CC. Number of Links Pointing to Page

More the number of links referring to a webpage, more is the website secure. According to evidence, 98% of illegitimate dataset have no links referring them. So, if the number of inks pointing to a webpage is 0, then the website is phishing, if they are between 0 and 2, then the website is suspicious, and legitimate otherwise.

DD. The Existence of "HTTPS" Token in the Domain Part of the URL

The phishers may attach the "HTTPS" portion to the domain part of their website's URL cheat on users. One example is: a phishing version of paypal, http://https-www-paypal-it-webapps-mpp-home.softhair.com/. As we can see, this URL has HTTPS in it, although it is not secured.

VI. RESULTS

We applied five machine learning algorithms on our dataset. The results obtained for the top three algorithms are mentioned in Table I, Table II and Table III.

The accuracy of Generalized Linear Model (GLM) was 93.33% and that of Generalized Additive Model (GAM) was 96.74%. Then, some methods were used for increasing our accuracy. After removing the attributes with high VIF values, we calculated the accuracy for three top algorithms.

TABLE I: Accuracy chart of different algorithms

Algorithm used	Accuracy after implementing independently (%)	Accuracy after implementing after removing variables with high VIF (%)	Accuracy after implementing after applying PCA (%)
Decision Tree	87.82	87.82	89.53
Random Forest	96.71	96.71	**98.40**
GBM	94.18	94.48	95.32

The increase in accuracy was minute. Then, after applying Principal Component Analysis (PCA), we managed to better our accuracy by a decent margin. The accuracy for random forest algorithm after applying PCA went as high as 98.4 %.

TABLE II: Recall chart of different algorithms

Algorithm used	Recall after implementing independently (%)	Recall after implementing after removing variables with high VIF (%)	Recall after implementing after applying PCA (%)
Decision Tree	89.95	89.65	90.01
Random Forest	96.89	97.28	98.59
GBM	94.43	95.22	95.20

The algorithms were implemented in same fashion as before and recall rates were calculated. As it is evident, the rates went down slightly after removing variables with high VIF values, but they increased considerably after applying PCA.

TABLE III: Precision chart of different algorithms

Algorithm used	Precision after implementing independently (%)	Precision after implementing after removing variables with high VIF (%)	Precision after implementing after applying PCA (%)
Decision Tree	81.96	81.96	85.32
Random Forest	95.71	95.03	97.70
GBM	92.24	92.17	93.95

Like accuracy, the precision rates too were highest for random forest after applying PCA. The algorithms were first implemented independently, once again after removing variables with high VIF values, and then after applying PCA.

VII. CONCLUSION

This paper mainly comprises of machine learning techniques to detect the phishing websites. Phishing websites mostly retrieve user's information through login pages. They are mainly interested in the bank details of the users. Out of the many features considered, the most important one was HTTPS with SSL i.e. whether a website uses HTTPS, issuer of certificate is trusted or not, and the age of certificate should be at least one year. In the future, we would like to extend our project by creating an extension to block the detected phishing website whenever the user clicks on their link.

REFERENCES

[1] Ebubekir Buber , ÖnderDemir , OzgurKoraySahingoz "Feature Selections for the Machine Learning based Detection of Phishing Websites", in 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017

[2] P. Singh, Y.P.S Maravi, S. Sharma, "Phishing websites detection through supervised learning networks", in IEEE International Conference on Computing and Communications Technologies (ICCCT)., pp. 61–65, 2015

[3] A. A. Ahmed, N. A. Abdullah, "Real time detection of phishing websites", In 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016

[4] Z. Dan Dong, A. Kapadia, J. Blythe and L. J. Camp, "Beyond the Lock Icon: Real-Time Detection of Phishing Websites Using Public Key Certificates" In IEEE APWG Symposium on Electronic Crime Research, pp. 1-12, May 2015

[5] S. Marchal, J. Francois, R. State, and T. Engel, "PhishScore: hacking phishers' minds," in proceedings of the 10th International Conference on Network and Service Management 2014 (CNSM 2014), vol. 11, no. 4, pp. 458-471, 2014.

[6] Luong Anh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen, Minh Hoang Nguyen, "A novel approach for phishing detection using URL-based heuristic." In IEEE International Conference on Computing, Management and Telecommunications (ComManTel), pp. 298-303, 2014

[7] Mustafa Aydin, Nazife Baykal, "Feature Extraction and Classification Phishing Websites Based on URL", in IEEE International Conference on Communications and Network Security (CNS), pp.769 – 770, 2015.

[8] Rami M. Mohammad, FadiTabah, Lee McCluskey, "Phishing Website Features" Unpublished. Available via: http://eprints.hud.ac.uk/24330/6/RamiPhishing_Websites_Features.pdf.