International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 3, Issue 10, ISSN No. 2455-2143, Pages 26-31
Published Online February 2019 in IJEAST (http://www.ijeast.com)

URL PHISHING DATA ANALYSIS AND DETECTING PHISHING ATTACKS USING MACHINE LEARNING IN NLP

Dr.G.Ravi Kumar
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

Dr.S.Gunasekaran
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

Nivetha.R
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

Sangeetha Prabha.K
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

Shanthini.G
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

Vignesh.A.S
Department of CSE
Coimbatore Institute of Engineering and Technology
Coimbatore,India

ABSTRACT - Nowadays people using the internet for shopping, banking, mailing etc. Phishing is one of the major attacks on the website which people are facing in their day to day life. A phishing attack is one of cybercrime because it is the illegal attempt that gets sensitive information of the user such as username, password, and credit card detail. Too aware of such phishing attacks taken online so in this paper have to detect phishing Uniform Resource Locator (URL), that is, we loading the URL data from the Kaggle open source website which is an online community of data scientists and machine learning, owned by Google Limited Liability Company(LLC). In most of the phishing website, the attackers use a malicious URL which will display to the user like an authorized URL. Different algorithms like Naive Bayes, Random Forest, K nearest neighbor are performed in detection of the URL, by using algorithm their accuracy level will be different. So in this paper can adopt the best classification machine learning algorithm with SVM (Support Vector Machine), this predicts the phishing or non-phishing status of the given URL and it is the best algorithm in classification (based on the features of given data) and regression (is the continuous prediction of uniform data) from which we have to improve our accuracy level.

KEYWORDS - URL detection, Support Vector Machine, Phishing detection system, Machine Learning, Natural Language Processing

I. INTRODUCTION

In the real world, the internet plays a vital role in communication, where people create an online environment to manage with offline business and online business functionalities. The Internet offers many benefits to the user and it also has a negative side so the user must be aware of it because there are many risks when people are in an online environment. The user may be vulnerable to online fraud by phishing and their entry point where usually be a masqueraded URL. To detect phishing URL, phishing emails and phishing websites heuristic technique are used. When comparing with other models URL analysis is the most promising technique and this technique depends on the lexical analysis because the lexical feature is extracted from URL.

The user has many accounts on various websites like the social network, email, and banking. The unsafe target towards these attackers are the unimpeachable web users because they were unaware of their valuable information that helps to make this attack successfully. The spoofed links are placed on the popular webpage and the phishing links are send through emails to the user. Similar to legitimate web pages, the fake web pages are created, which directs the user to the attacker's server instead of a real web server.

Some solution like anti-virus, firewall, and software don't fully prevent the web spoofing attack. In a web spoofing attack, they implement a secure socket layer and digital certificate but it does not protect the web user against the attacker and the attacker divert the user to the fake web server. To protect the user from the attackers they use a secure browsing connection and they have knowledge about their secure connection. In phishing URL have some unique characteristics which are different from the legitimate web page.

II. RELATED WORK

Researches have worked on secure routing, intrusion detection, intrusion prevention, and smart grid security. For web phishing, they use graph mining techniques because URL analysis cannot detect some phishing, but it can be detected by graph mining techniques. The relationship between user and website can be utilized by it. After getting the dataset, the data can be cleaned and trained. Each cleaned data and trained dataset has eight fields: User node number, user source IP access time, visiting URL, reference URL, user agent, access server IP, user cookie. Each user assigned a unique user node number but a different IP address and therefore they create a relationship between user node number and visiting URL. Mutual behavior of the graph detects the phishing websites.

On the internet, due to the continuous growth of malicious activities, there is a need for identifying the malicious web pages. URL analysis is an efficient method for detecting phishing, malware, and other attacks. In the previous survey, URL classification using a combination of lexical features, network traffic, hosting information, and other strategies have been performed. Time-intensive lookups will introduce significant delay in real-time systems. In URL phishing data analysis and detecting phishing attacks, we describe a lightweight approach for classifying malicious web pages using URL lexical analysis alone. Our goal is to find the classification's accuracy of a purely lexical approach. It develops a flexible approach which is used in a real-time system. The Classification system is developed based on lexical analysis of URLs.

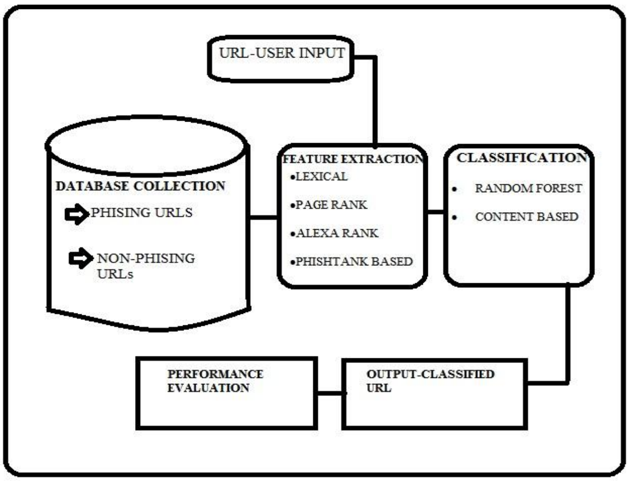


Fig 2.1: System Diagram of Existing System

To detect the phishing website the proposed different models, journals, conference and researches. Multiple classification algorithms used for detecting phishing website, which includes Adaboost and Naïve Bayes. In this algorithm, the tires are divided into three by using 21 fixed different features. With the help of another classification algorithm, a two-step procedure takes places. But the problem is time consumption and complexity and this is not an optimal method.

Pattern recognition was another method which is used for detecting the phishing websites. Phishing messages are distinguished from non-phishing one with the help of the model, which features extracted from emails. Many features are used to identify phishing without checking that they essentially need or not and therefore, it leads to computational cost.

In another technique, the phishing detection technique is divided into blacklist based and heuristic based approaches. In blacklist based approaches, it maintains a database which has the list of the URL address and they are classified as malicious. If the user request a site and the site is malicious then the connection will be blocked. The advantages of blacklist based approaches are easy implementation and low falls positive rate. But it can't detect the phishing site that is not in a database.

Data mining is also one of the methods for detecting phishing websites. The associative classification algorithm is used to detect a malicious website with accuracy. The data are collected from various source and rules are extracted from the trained dataset. The resulting roots are merged with a different class to produce multi-labeled rules so the redundant rules are eliminated. On the tested data, the testing classifier to measure the performance. The main problem in this technique is determining minimum confidence and minimum support in a large amount of data and it has to be replaced with more accuracy and less time complexity.

To classify URL they have used Random Forest Algorithm. It is a popular method which can be used for classification and regression problems. Random Forest uses multiple learning models for a better result. For the best possible answer using the Random Forest Algorithm to create and related decision tree. But in this algorithm, it generates the prediction slowly. If it makes a prediction all the tree in this algorithm have to make a prediction for the given input as the same for the performance. It is a time-consuming process. When this algorithm compared to the decision tree doesn't have an accurate result.

III. PROPOSED SYSTEM

Our proposal to find whether the URL is phishing or non-phishing using Support Vector Machine and Natural Language Processing techniques. Support Vector Machine is also called a Support Vector network which is a supervised learning model that associated with learning algorithm which is used to examine the data for classification and regression. We have a set of training exam and each of this exam belongs to any one of three categories. Support Vector Machine algorithm allocates a new example to one or other categories.

Support Vector Machine is used to find a hyper plane that divides a dataset into two classes. A Hyper plane is a line that linearly separates and classifies set of data. When the new testing set of data is added the other side of the hyper plane it lands will decide the class, that we assigned to it.

The data set is divided into two classes and the hyper plane gets divided into two different sets. Here we are going to find the right hyper plane, as the distance between the hyper plane and the nearest data point in a set is known as margin. In the good margin, the distance between the support vector and margin will be equal. In a bad margin, the distance between the margin and support vectors will not be the same.

The goal is to choose the hyper plane with the greatest possible margin between a hyper plane and any point within the training set. SVM is used for face detection, text and hypertext categorization, classification of the image. In addition to linear classification, kernel trick is used to perform a non-linear classification. It maps their input to their high dimensional feature space. When data is untagged, there will be no supervised learning only unsupervised learning takes place.

To detect the binary classification problem, using a linear classifier of labels y and feature x and to denote the class labels y ∈ {-1,1} is used with vector and parameter w,b

The notation is denoted with $h_{w,b}(x)$

$$\boxed{\boldsymbol{h_{w,b}(x) = g(w^T x + b)}}$$

Consider g(z)=1 if z > 0 otherwise g(z)=-1.
The notation w, b used to intercept term b separately from other parameters.

It is used to find the clustering of data into groups and map new data to form a group. The problem occurs infinite dimensional space. Original dimensional space is mapped into high dimensional space. SVM scheme is used to protect the dot product of the input data which are easily computed in terms of a variable.

In day-to-day life, NLP was used for speech recognition, speech translation, understanding complete sentence, understanding synonyms of matching words.

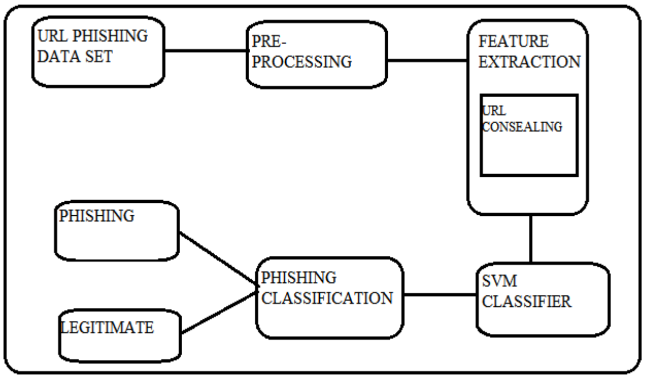


Fig 3.2 : Block diagram for proposed system

IV. IMPLEMENTATION

In the field of e-commerce, phishing attack gives annoying threads to web users. The system focus on the features between the legitimate and ill-legitimate URL. In training dataset, these features are utilized as a portion to train the system and it is used to detect the URL category based on their characteristics.

4.1. Feature Extraction

It will reduce the number of variables in the dataset which has the most discriminatory information and the feature are extracted from URL. Some of the parameters are Length of URL, IP Address, Subdomain, HTTPs Symbols, Website traffic, SVM, Dots, SSL, Feature Vectors.

4.2. **Length Of URL**

To hide the fraudulent domain, the lengthy URL is created to deceive the user by the attackers. The length of the URL is un-measurable but the research says the accepted URL should be less than 56 characters.

4.3. **IP Address**

It is a unique numerical label is assigned to each device which is connected to computer networks by using Inter Process Communication and it is the domain of URL.

4.4. **Subdomain**

To indicate doubtful URL domains are inserted into URL without the user knowledge. The dual sub domain has notification of a phishing website.

4.5. **HTTPs**

When the user doing the online transaction which reflects the security by linking HTTP protocol to a website. If the URL has the security certificate then it is a sign of authenticity.

4.6. **Symbols(@)**

The fake websites used the @ symbol within the URL address because the attackers lead the user to fake websites.

4.7. **Website Traffic**

When a website has high traffic then it is safe to browse otherwise it is a phishing website. By using Alexa-database it can verify the rank of a website.

4.8. **Dots**

If a URL has 5 dots then it is a legitimate URL or else it is a phishing URL.

4.9. SVM

SVM algorithm is a machine learning algorithm which is used to find a phishing website and it is support vector network. It is a supervised learning model used for classification and regression.

4.10. SSL

Secure Socket Layer is a security protocol which is used to link a web browser and a browser in online communication.

4.11. Feature Vectors

Feature Vector represents some object as an n-dimensional vector of numerical features.

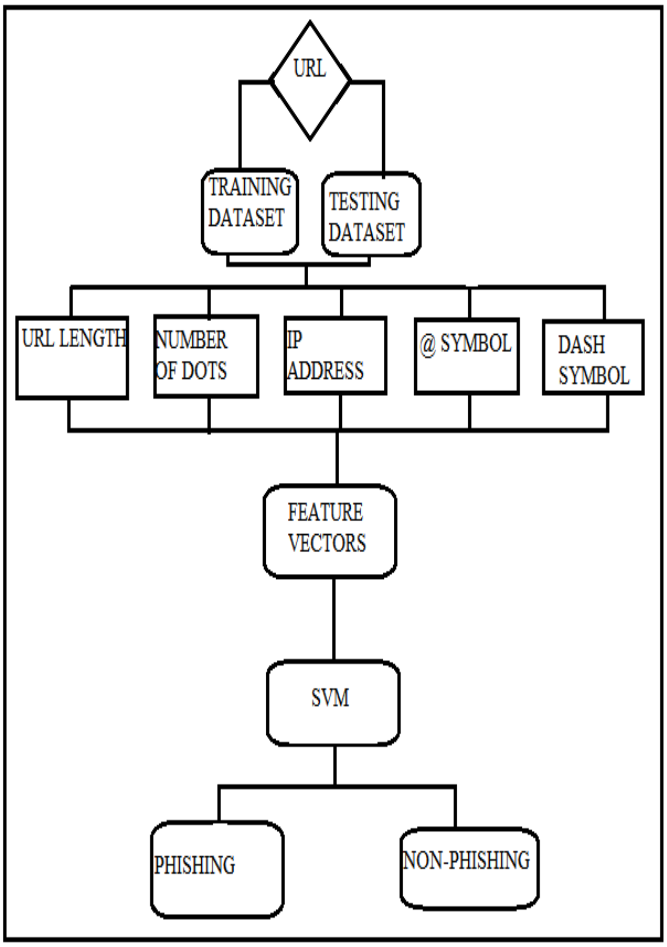


Fig 4.1: Workflow diagram of a complete system

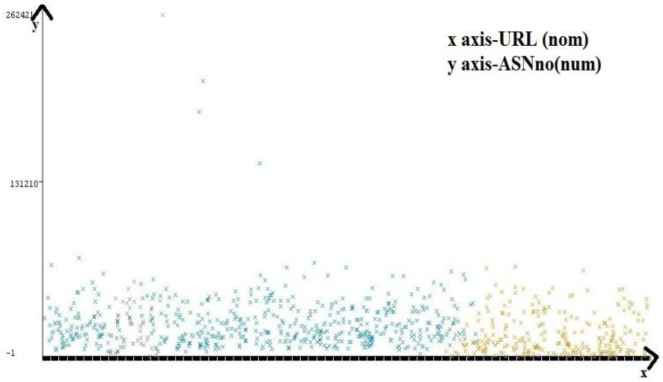


Fig 4.2: Scatter Diagram of attribute URL and ASN no.

By using URL and ASN (Autonomous System Number), Malware URL is detected based on the SVM algorithm.ASN is a unique number to identify the autonomous system which exchanges the routing information with other autonomous systems.

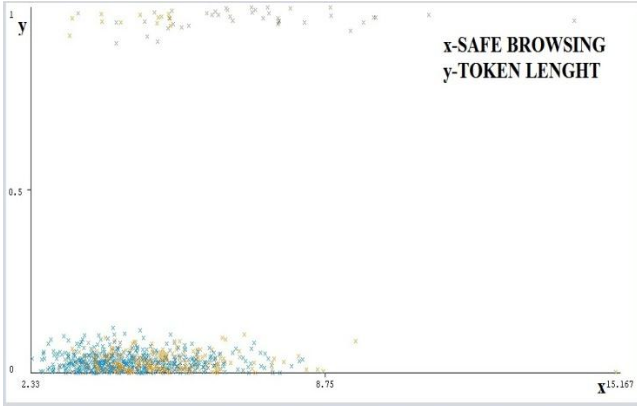


Fig 4.3: Scatter Diagram of attribute Safe Browsing and Token Length

By using safe browsing and token length, Malware URL is detected based on the SVM algorithm.

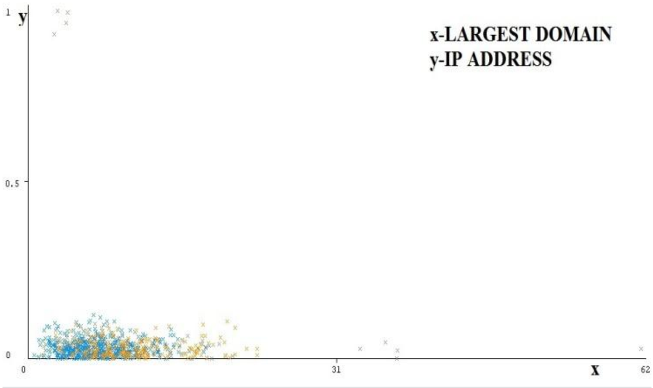


Fig 4.4: Scatter Diagram of attribute Largest Domain and IP Address

By using the largest domain and IP address, Malware URL is detected based on the SVM algorithm.

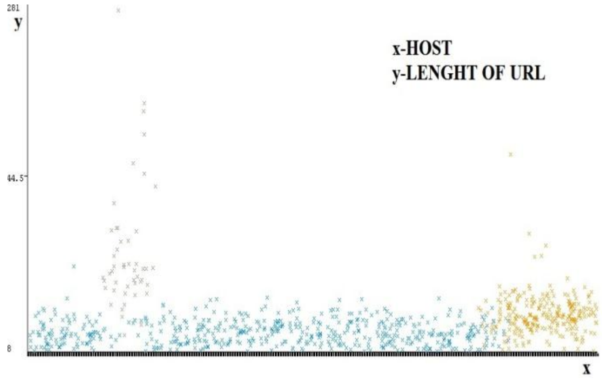


Fig 4.5: Scatter Diagram of attribute Largest Host and IP Address

By using host and length, Malware URL is detected based on the SVM algorithm.

V. DISCUSSION

In the phishing attack, the attack is to steal sensitive information about the user through the phishing URL. Malicious URL where using in phishing attacks and displayed like a legitimate URL.

In the kernel-based method, SVM is widely used for binary classification. In this, we used URL structure web-based content, web page traffic, and DNS information. This model is good at phishing detection malware sites and collective 100% of phishing sites.

VI. RESULT

While using the Random Forest algorithm, the time taken to detect the phishing is 7 millisecond and while using Support Vector Machine, the time taken to detect the phishing is 5 millisecond. When using the SVM algorithm for classification, time is reduced by a few milliseconds as compared to the existing system that uses clustering to make classification of the website. SVM is the most effective method used for classification of normal or phishing websites. The URL with label 0 is called begin and the URL with the label 1 is called malicious. The implementation of the proposed system was carried out successfully and it efficiently classified the URLs providing a satisfactory outcome. This experiment results the solution is powerful to catch phishing URLs and used as a plug-in in the browser to filter the phishing site.

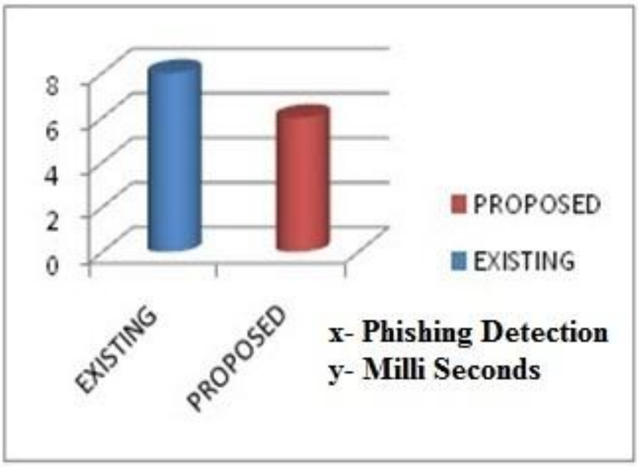
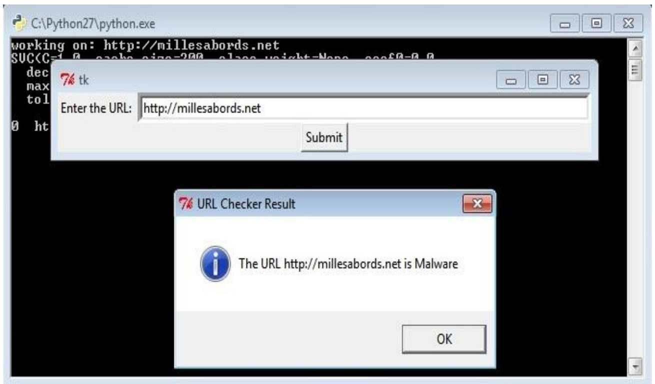


Fig 6.1: Existing Detection vs. proposed Detection results in milliseconds

In fig 6.1 the x-axis defines the time difference between existing and proposed system and y- axis defines milliseconds taken by existing and proposed a system.



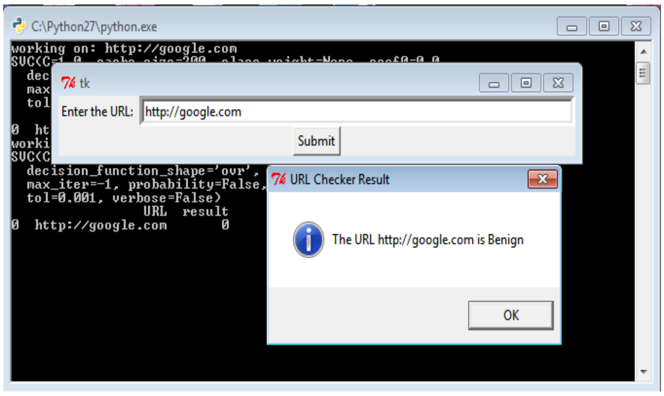
```
working on: http://millesabords.net
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  dec
  max
  tol

0  ht
```

Enter the URL: http://millesabords.net

The URL http://millesabords.net is Malware

Fig 6.2: The system specifying that the entered URL is Malware



```
working on: http://google.com
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  dec
  max
  tol

0  ht
worki
SVC(C
  decision_function_shape='ovr',
  max_iter=-1, probability=False,
  tol=0.001, verbose=False)
                 URL  result
0  http://google.com       0
```

Enter the URL: http://google.com

The URL http://google.com is Benign

Fig 6.3: The system specified that the entered URL is benign

VII. CONCLUSION

We discuss the phishing attack by taking a dataset that contains different URL and different URL features. In today’s world, people strongly believe in all the websites hence phishing is growing crime and we must be aware of it. The machine learning method is used to detect the given URL is phishing or non-phishing by using Support Vector Machine algorithm and NLP techniques to improve the accuracy level.

VIII. REFERENCE

[1] Adithya.M1, Adela Adhya Bhupesh1, Chitra Pallavi Hariharan1, Sakshi Sinha1 and Pandiaraj.S2, “URL Phishing Detection using Classification Techniques”

[2] Aggarwal, S., Kumar, V., and Sudarsan, S. D. Identification and, detection of phishing emails using natural language processing techniques. In Proceedings of the 7th International Conference on Security of Information and Networks (2014), SIN ’14.

[3] Basnet. R. B, Sung. A. H, "Mining web to detect phishing URLs", Proceedings of the International Conference on Machine Learning and Applications, vol. 1, pp. 568-573, Dec 2012.

[4] Choon Lin Tan, Kang LengChiew, San Nah Sze, “Phishing Website Detection Using URL-Assisted Brand Name Weighting System”, 2014 IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) December 1-4, 2014.

[5] Cohen W. (1995) Fast effective rule induction. In machine learning: Proceedings of the 12th International conference, pp. 115-123. LakeTahoe, California. Morgan Kaufmann.

[6] International Conference on Computing for Sustainable Global Development (INDIACom), (pp. 2125-2130).

[7] Jain, A. K., & Gupta, B. B. (2016). Comparative Analysis of Features Based Machine Learning Approaches for Phishing Detection.

[8] Luong Anh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen1 and Minh Hoang Nguyen, “A Novel Approach for Phishing Detection Using URL-Based Heuristic”, 2014 International Conference on Computing, Management and Telecommunications (ComManTel), IEEE 2014.

[9] Marco Cova, Christopher Kruegel, Giovanni Vigna, "Detection and analysis of drive-by-download attacks and malicious javascript code", Proceedings of the 19th International Conference on World Wide Web, pp. 281-290, 2010.

[10] Mohammad RM, Thabtah F, McCluskey L. Tutorial and critical analysis of phishing websites methods. Computer Science Review 2015;17:1–24

[11] Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu, "A survey of network anomaly detection techniques", J. Netw. Comput. Appl., vol.60, no. C, pp. 19-31, 2016.

[12] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, “Learning To Detect Phishing URLs”, IJRET: International Journal of Research in Engineering and Technology, Volume: 03 Issue: 06 | Jun-2014.

[13] Usha Narra, Corrado Aaron Visaggio, Mark Stamp, Thomas H. Austin, “Clustering versus SVM for malware detection”, Springer, Journal of Computer Virology and Hacking Techniques 10/2015

[14] Uzun E., Agun H. V., and Yerlikaya T. A. (2013) A hybrid approach for extracting informative content from web pages. Information Processing & Management, (49), 928-944, 2013

[15] Zuochao Dou; Issa Khalil; AbdallahKhreishah; Ala Al-Fuqaha; Mohsen Guizani, “Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection”, IEEE Communications Surveys & Tutorials, 2017