Taseer Suleman *et al.LGU (IJECI)* 2021

LGU (IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

LGU International Journal for
Electronic Crime Investigation

Research Article

Vol. 5 issue 2 Year 2021

A Survey On Web Phishing Detection Techniques: A Taxonomy-based Approach

Taseer Suleman
University of Management and Technology

Abstract:

The primary goal of website phishing is to obtain secret information i.e. passwords, account numbers, credit card details, etc. Web-phishing is used to deceive users, normally carried out through sending links using spoofed emails, instant messages etc. However, web-phishing detection is a challenging task. A number of techniques and mechanisms has been proposed for the detection of web-phishing. The aim of this study is to analyze different web-phishing detection techniques. Web-phishing techniques are characterized into machine-learning (ML) based, Heuristics-based, Blacklist/whitelist based and visual-based. A comparative analysis of these aforementioned categories has been done in this research based on their detection accuracy, performance, usability, and scalability. The research also identifies the advantages and limitations of web-phishing detection techniques.

Keywords: Web-phishing, Machine-Learning based, Heuristics-based, Blacklist-based

1. Introduction

Web-phishing is an online crime for obtaining personal information like banking details, credit card numbers, and social security numbers. Phishing was actually started in 1995 with America Online (AOL) users [1]. Attackers lure users by sending spoofed emails to them. Rogue links can also be sent through online social media and other messaging services [2]. Victims are redirected to the illegitimate websites when they click on those rogue links. These websites are usually the clone of a legitimate website. A careless user can give personal details on these rogue websites without checking the webpage legitimacy or Uniform Resource Locator (URL). Hackers then use these details for malicious purpose. The choice of victim and the amount of benefit are important parameters in the web-phishing attack. Web-phishing strategies include SQL injection, Tab-nabbing, Typo-Squatting, content-injection, malware-based and DNS-based attack [3]. Statistics from Anti Phishing Working Group (APWG) 2018 report shows the increase of web-phishing attacks in the previous year 2018. The report has also shown that most of these fake websites are using Hyper Text Transfer Protocol – Secure (HTTPS) services. The use of HTTPs hosted websites is to gain the trust of victims.

Many web-phishing detection solutions are proposed that can be categorized into Heuristics-based, Blacklist/Whitelist based, Machine learning based and Visuality-similarity based [4]. In this article, various techniques and mechanisms on web-phishing detection solutions are discussed. These techniques and mechanism are generally revolved around the aforementioned categories. In addition, a comparison analysis has been done to evaluate more about web-phishing detection techniques. It would help researchers in understanding the advantages and limitation of these anti-phishing techniques. In the end, we concluded our research based on the detailed analysis of the web-phishing detection techniques. In Figure 1, the taxonomy of web phishing and its detection is given.

The rest of the paper is organized as follows: In the next section, the web-phishing life cycle is discussed. In section III, phishing statistics are shown according to most recent research reports. Section IV highlights web-phishing strategies. In Section V, a detailed analysis of web-phishing detection techniques is given. Section VI concludes our research.

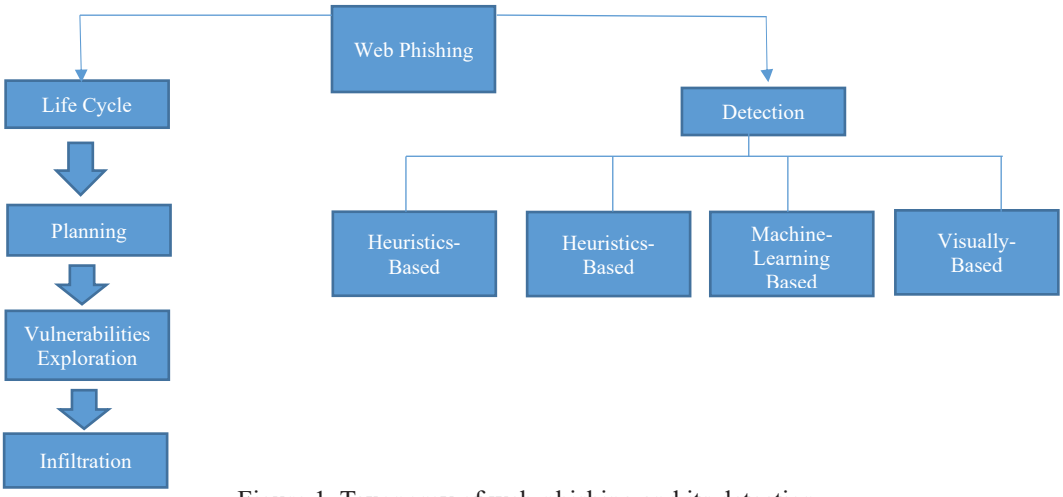

Figure 1. Taxonomy of web-phishing and its detection

2. Web-Phishing Life Cycle

A typical web-phishing life cycle comprises of few stages like planning and setup, vulnerabilities identification, infiltration and information accumulation [5]. We go through these stages in detail.

A. Planning and Setup

In the first stage, the phisher determines the objective association, an individual or a country to be targeted for malicious purpose. They uncover sensitive information with respect to their objective and its system. Normally phishing starts by sending spoofed emails or messages to the victims [6]. Victims are supposed to send required information via replying to the email. However, most of the users do not reveal their information through email. Another phishing technique can be adopted through the creation of phishing websites.

B. Finding system vulnerabilities

The target, purpose, and motivation of web-phishing is well defined. Web-phishing carried out by utilizing browsers vulnerabilities, web link manipulation, malicious use of scripting languages, spoofing website text and images.

C. Break-In or Infiltration

At this stage, the attacker penetrates into the system, takes control of the system, and perform malicious activities. This penetration may be caused through a vulnerability in the victim's system.

D. Information Accumulation

After the successful infiltration, the attacker does the information collection. Information may contain passwords, user identity number, contact lists, private images, and credit card information. The whole web-phishing life cycle is shown in Figure 2. An active attacker sends a link of the fake webpage via email to the victim. The victim is redirected towards the fakewebsite when click on this link. This fake website seems to be original to the victim. In this way, the victim gets compromised.

3. Web-phishing Statistics

According to the Anti-Phishing Working Group (APWG) 2018 report, there is an increase in web-phishing attacks. In September, 53,546 unique phishing websites detected which show a drastic increase than the month of July and August [7]. The APWG report also states the increase in the usage of phishing websites hosted on Hyper Text Transfer Protocol –Secure (HTTPS). Figure 3 shows an increase in websites using Secure Socket Layer (SSL) services. APWG 3rd quarter 2018 report reveals the most targeted industry is the online payment service. Webmail services and cloud service were also remain affected by phishing

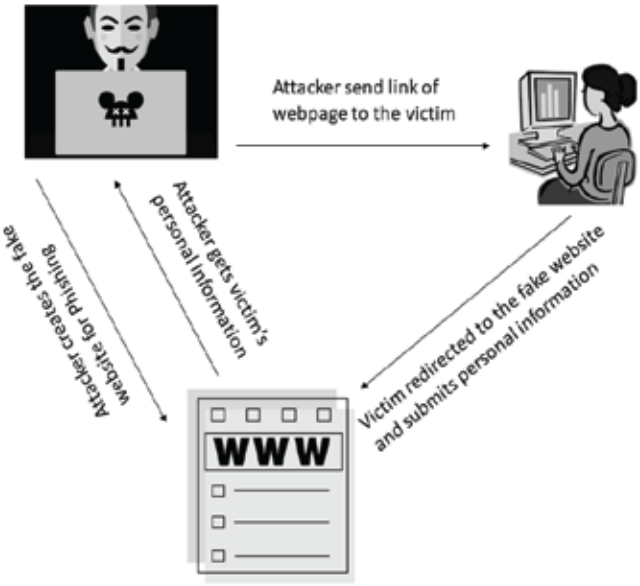

Figure 2. A typical Web-phishing Life Cycle

attacks in 2018. Kaspersky 3rd quarter report [8] figured out that web-phishing involved the compromise of personal data, malicious attacks against the banking sector, universities and job searching platforms. The report also revealed the phishing attacks has been increased against cryptocurrency.

4. Web-phishing Strategies

Attackers used many ways to carry out the website phishing attack. Few are the most common web-phishing strategies.

Figure 3. Phishing websites hosted on HTTPS [APWG report]

A. Spear Phishing:

It is a kind of phishing, which targets specific individual or a specific company. This technique is carried out by attackers usually by sending spoofed email or messages to the victims [9].

B. Tabnabbing:

The inactive tab of the browser replaced with the malicious webpage in this website phishing variant [10]. When the user switches back to the tab, it looks legitimate to the user. These tactics used for getting sensitive credentials from the users.

C. URL manipulation:

Website phishing is successfully achieved through the manipulating of URL [11]. Changing, adding, deleting, editing in the URL parts are very common tactics adopted by the attackers. Other techniques include exploiting browsers vulnerabilities, rogue scripts, spoofing websites etc. Figure 4 highlights some of the most common strategies used by the attackers.

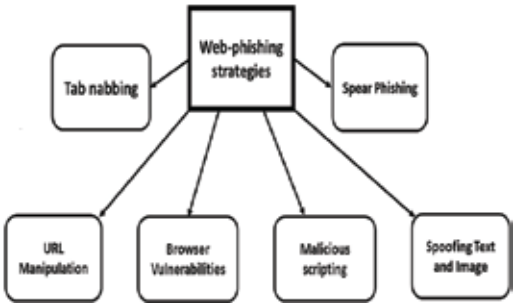

Figure 4. Web-phishing strategies

5. Web-Phishing Detection Techniques

There is a number of mechanisms developed for the detection of phishing websites [12]. These can be categorized into Heuristics-based approaches, Blacklist/Whitelist based mechanisms, Machine-Learning (ML) techniques, and visual similarity based mechanisms [13]. Heuristics-based mechanisms use unique features for the detection of illegitimate websites [14]. Based on these features, algorithms trained up to some threshold for possible detection of wrong websites. Blacklist contains the list of illegal websites as reported by the anti-phishing groups. Google had introduced this blacklist feature in its google chrome browser that checks each URL against the google blacklist [15]. The visual similarity based techniques used to compare the two web pages in terms of their appearance and layout. If it seems to be similar then the system check for URL authenticity and then the fake webpage is marked as phish webpage [16]. Different features of URL are marked for possible detection of phished webpages. Machine-learning algorithms trained on these features, in order to, automate the detection process [17].

In this section, we present a major detection mechanism for phishing websites. A detailed analysis of these mechanisms is discussed, in order to; create deep understanding for the researchers.

A. Heuristics-Based Detection Mechanism:

The heuristic-based approach used different features of the website, in order, to differentiate between phishing and non-phishing webpage. Jaydeep et al. [18] have used some features related to a webpage i.e. URL, source code etc. for phishing webpage detection. Lee et al. [19] proposed a heuristic-based approach for phishing detection using 3000 phishing websites and 3000 non-phishing websites. Their proposed method shows a good response to web-phishing detection.

Gastellier-prevost et al. [20] developed an anti-phishing toolbar named "Phishshark". This tool uses 20 heuristics for the identification of legal and illegal web pages.

Nguyen [21] applied a heuristics-based mechanism using URL features. They used a dataset of 11,660 phishing websites and 5,000 true websites. The proposed technique success rate is 97%.

1) Limitation in using Heuristics based detection mechanism:

Heuristics based mechanisms improved detection accuracy. However, implementation is difficult due to the complexity in the implementation and cost overhead.

B. Blacklist/Whitelist based mechanism:

Li et al. [22] developed an anti-phishing tool for the browser. This tool manages two lists named as white list and blacklist of web pages. When the user clicks on a link containing

URL, it then checked against these two lists. If the URL saved in blacklist the browser prevented from redirecting to that specific webpage.

Krishnamurthy et al. [23] also adopted the mechanism of blacklist/whitelist for possible phishing website detection. At first, the URL is searched in the white list. If no match found, the same URL is compared in the blacklist.

1) Limitation in Blacklist/White list based mechanisms:
Both lists should be updated, in order to, detect new URLs for phishing websites. Moreover, with the regular update on the client side it can create storage issues as well. On the server side, storage can create a delay in accessing the updated list for possible detection of phishing websites.

C. Machine Learning based mechanism:
Abu-Nimeh et al. [24] applied machine-learning algorithms to detect phishing emails. These algorithms are then compared in terms of accuracy in detecting phishing emails.

Le et al. [25] used ML techniques for the classification of websites. The algorithms used URL-based features such as URL length, a special character in URL and domain name etc. This technique improved accuracy but also increased the overhead for processing.

Tan et al. [26] developed 'PhishWho', an anti-phishing system, for the detection of possible phishing websites. This system works in three stages starting from the identification of keywords from a website to the decision of website legitimacy. Websites features also play a pivotal role in the identification of phishing websites.

Mohammad et al. [27] developed an anti-phishing system using neural networks for classification. The system used 17 features for classification.

Moghimi and Varjini [28] used Support Vector Machine (SVM) along with Levenshtein Distance for phishing detection. They used 25 features for classification. However, sophisticated website design by phishers remains undetected by the system.

Mohammad et al. [29] used data mining methods, in order to, detect phishing. They performed different data mining algorithms and proved C4.5 to be much better in terms of detection.

Tuan et al. used a dataset of almost 11660 phishing websites for the extraction of features for illegitimate websites [30]. They narrow down these features to six important features with a detection accuracy of 98% approximately.

Feng et al. [31] proposed a method for the detection of phishing web pages using neural networks based classification methods. Their proposed system shows 98% detection accuracy approximately.

Wewei et al. [32] used the results of trained classifiers along with the categorization of phishing websites using hierarchical clustering algorithms. Kausar et al. [33] used a combination of both heuristics mechanism and naïve Bayes classifier, in order to, improve accuracy for phishing detection.

Burber et al. [34] in their research used Natural Language Processing (NLP) for the extraction of URL-based features. They used three

machine-learning algorithms for possible detection of phishing websites. The proposed methodology improved detection accuracy.

Jain et al. [35] proposed a client-side solution for the detection of website phishing. They used 2141 phishing websites from PhishTank [36] and applied machine-learning approaches. Rao and Pais [37] used a hybrid methodology for the possible detection of phishing websites. Hybrid approach includes machine-learning approaches and image checking as well.

James et al. [38] connected distinctive sorts of machine learning based arrangement calculations, including Naive Bayes (NB), Support Vector Machine (SVM), Neural Net (NN), Random Forest (RF), IBK relaxed classifier and Decision Tree (J48). Performance of all these aforementioned algorithms is compared and accuracy was determined against each algorithm.

1) Limitation in using machine-learning based detection mechanism:
Machine-learning based detection mechanism contains computational overheads. The slow processing of datasets for algorithms learning increase the latency of website phishing detection. These kinds of techniques are difficult to apply on the client side in terms of browsers extensions or add-ons due to computational cost. In order to improve detection results, a lightweight solution is required. Moreover, a hybrid approach can also be used to make detection accuracy better.

D. Visual similarity based mechanism:
This technique based on the visual features extracted from the websites. These features are later used in the comparison of legitimate website visuals with illegal website visuals.

Chiew et al. [39] proposed a method of extracting a website logo for the detection of phishing websites. They used machine-learning algorithms for possible detection.

Philippe et al. [40] proposed "tab shows", a mechanism that takes the screenshot of the tabs. Whenever a tab is opened again, the screenshot is again saved. Match analysis is performed with the current screenshot and the previous one. It alerts the user in case of any difference in both screenshots.

Lam et al. [41], the author performed a similarity analysis of layout instead of webpage content analysis. In this scheme, image processing techniques are highly involved, in order to, carry out detection.

1) Limitation in visual similarity based mechanism:
These techniques require huge computational resources for processing of images. Complexity computational overhead is always involved in such a mechanism. A lightweight solution might be helping in such a case if that is implemented on the client side.

In Table 1, we have given a comparison of the most common used web-phishing detection techniques, in detail. It would help in the deep understanding of the detection of phishing websites techniques in a comparative analysis. Four major mechanisms are targeted in the analysis that includes Heuristics-based, ML-based, Blacklist/Whitelist based and Visual-based.

Table 1. Comparative analysis of web-phishing detection mechanisms

Detectaion Mechanism	Technique Used	Pros	Cons
Heuristics-Based [18]	Collecting URL Features	The good approach towards detection	Minimal features were used
Heuristics-Based [19]	Using dataset of Phishing and Non-Phishing websites	Improved detection accuracy	The dataset contains fewer samples
Heuristics-Based [20]	Anti-phish toolbar developed using 20 Heuristics	Much heuristics for differentiation	Client-side requires many computational resources.
Blacklist/Whitelist Based [22]	Maintain lists in the browser for anti-phishing	Check both lists for the legal or illegal webpage	Regular list updating issue, client-side list storage issues
Blacklist/Whitelist Based [23]	Improved scheme than in [22], First check whitelist for the legal webpage	Maintains both list i.e. Blacklist and Whitelist	Computational overhead, Processing slow
ML-Based [24]	Applied ML algorithms to detect Phishing emails	Novel approach as emails are a primary source for phishing	Applied more ML algorithms with feature-selection capability might improve results
ML-Based [25]	Applied ML algorithms using URL-based features	Improvement in detection accuracy	Processing overhead, Need more URL-based features to get better results
ML-Based [26]	Worked in 3 stages using website features	Use keywords for matching	Can incorporate more features to improve results
ML-Based [28]	Used SVM along with Levenshtein distance	Used 25 unique feature to detect fake website	A careful-designed website might remain undetected
ML-Based [29]	Used Data-Mining approach	Proved C4.5 to give better accuracy	The small dataset used, the Hybrid approach might produce a better result
ML-Based [30]	Used Dataset of 11660 phishing websites	Applied feature selection (up to 6 features)	Dataset can be increased for a better result.
ML-Based [31]	Neural networks applied	Accuracy detection up to 98%	Much beneficial if applied on a lightweight technique on client-side
ML-Based [33]	A combined approach for detection using heuristics and Naïve Bayes	Improve detection accuracy	More ML algorithms can be applied for performance checking
ML-Based [37]	A hybrid approach of ML algorithms along image-check	Improved detection accuracy	Makes detection processing slow, Computational overhead
ML-Based [38]	Applied six ML algorithms for training	Improved mechanism than in [33]	Can be improved if feature selection algorithms also used
Visuals-Based [39]	Used website logo for detection of phishing webpage	The logo is compared to the real website logo stored in the database	A spoofed website detection is difficult, can be improved incorporating more features
Visuals-Based [41]	Used website layout for detection of websites	Much improved method than [39]	Highly image processing required

6. Conclusions

In this research, we have focused on the web-phishing problem. The aim of this study is to conduct a deep analysis of web-phishing detection techniques. The research focused on these detection techniques in terms of accuracy, performance, scalability, usability, and applicability. A comparative analysis of these detection techniques is discussed. It has been concluded that there is a need for a lightweight approach for web-phishing detection. A hybrid mechanism can also be helpful that can use different web-phishing detection techniques for better detection accuracy. Along with the improvement of these techniques, end user awareness is an important parameter to avoid web-phishing attacks.

7. References

[1] James, L., 2006. Banking on phishing. In: Phishing Exposed. Elsevier Inc., Ch. 1, pp. 1–35

[2] Shraddha, P., Dhwanil, P., Srushti, K., Smita, S., "A new method for Detection of Phishing Websites: URL Detection," Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)

[3] V. Suganya, "A Review on Phishing Attacks and Various Anti PhishingTechniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.1, April 2016

[4] R. Gotham., I, Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages," Computers and Security, Elsevier, 2014.

[5] Anjum N. Sheikh, Antesar M. Shabut, M.A. Hossain, "A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps", 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)

[6] C.E Drake, J.J. Oliver, E.J. Koontz," Anatomy of a Phishing Email," CEAS, 2004.

[7] Anti Phishing Working Group (AWPG). Phishing Activity Trends Report, 3rd Quarter 2018. https://www.antiphishing.org/resources/apwg-reports/

[8] Kaspersky, https://securelist.com/spam-and-phishing-in-q3-2018/88686/, Retrieved February 08, 2019.

[9] Kaspersky report spear Phishing, https://www.kaspersky.com/resource-center/definitions/spear-phishing, Retrieved December 28, 2018.

[10] Rableen, Kaur, S., Deepak, Singh, T., and Divya, Rishi, S., "An Approach to Perceive Tabnabbing Attack," International Journal of Scientific & Technology Research Vol 1, Issue 6, July 2012.

[11] Jain A.K., Gupta B.B. (2018) PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning. In: Bokhari M., Agrawal N., Saini D. (eds) Cyber Security. Advances in Intelligent Systems and Computing, vol 729. Springer, Singapore

[12] Gaurav, V., Manoj, M., and Pardeep, K.A., "A survey and classification of web phishing detection schemes," Security Comm. Networks 2016; 9:6266–6284

[13] Ozgur, koray, S., Ebubekir, B., Onder, D., Banu, D.," Machine learning based phishing detection from URLs", Elsevier, September 2018.

[14] Luong Anh Tuan N, Ba Lam T, Huu Khuong N, Minh Hoang N. A novel approach for phishing detection using URL-based heuristic. In Computing, Management and Telecommunications (ComManTel), 2014 International Conference on, 2014; 298–303

[15] Developers G. Safe browsing API-developer guide V3. 2014; https://developers.google.com/safe browsing/developers_guide_v3, Retrieved on December 28, 2018

[16] Jian M, Pei L, Kun L, Tao W, Zhenkai L. BaitAlarm: detecting phishing sites using similarity in fundamental visual features. In Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference on, 2013; 790–795

[17] Nirmala Suryavanshi, Anurag Jain , "A Review of Various Techniques for Detection and Prevention for Phishing Attack", International Journal of Advanced Computer Technology (IJACT). Vol 4 No.03.

[18] Jaydeep, S., Rupesh, G.V., "Website Phishing Detection using Heuristic Based Approach," International Research Journal of Engineering and Technology (IRJET), Vol 03, Issue 05, May 2016.

[19] Jin-Lee, L., Dong-Hyun, K., Chang-Hoon and Lee" Heuristic-based Approach for Phishing Site, Detection Using URL Features," Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015

[20] Gastellier-Prevost Sophie, Granadillo Gustavo Gonzalez, Laurent Maryline. Decisive heuristics to differentiate legitimate from phishing sites. La Rochelle, France. In: Proc. Of conference on network and information systems security (SAR-SSI); May 2011. p. 1e9.

[21] Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[22] Li L, Berki E, Helenius M, Ovaska S. Towards a contingency approach with whitelist- and blacklist-based anti-phishing applications: what do usability tests indicate? Behaviour & Information Technology 2014; 33(11):1136–1147.

[23] Krishnamurthy B, Spatscheck O, Van Der Merwe J, Ramachandran A. Method and apparatus for identifying phishing websites in network traffic using generated regular expressions, to Google Patents, 2009.

[24] Abu-Nimeh S, Nappa D, Wang X, Nai S. A comparison of machine learning techniques for phishing detection. In: APWG ecrime researchers summit (eCRS), Pittsburgh, PA; October 2007.

[25] Le, A., Markopoulou, A., & Faloutsos, M. (2011). Phishdef: URL names say it all. In 2011 Proceedings IEEE INFOCOM, 2011 (pp. 191–195).

[26] Tan, C. L., Chiew, K. L., Wong, K., & Sze, S. N. (2016). Phishwho: Phishing webpage detection via identity keywords extraction and target domain name finder. Decision Support Systems, 88, 18–27

[27] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications, 25(2), 443–458.

[28] Moghimi M, Varjani AY. New rule-based phishing detection method. Expert Systems with Applications 2016; 53:231–242.

[29] Mohammad RM, Thabtah F, McCluskey L. Intelligent rule-based phishing websites classification. IET Information Security 2014; 8(3):153–160.

[30] Luong Anh Tuan N, Ba Lam T, Huu Khuong N, Minh Hoang N. A novel approach for phishing detection using URL-based heuristic. In Computing, Management and Telecommunications (ComManTel), 2014 International Conference on, 2014; 298–303

[31] Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. (2018). The application of a novel neural network in the detection of phishing websites. Journal of Ambient Intelligence and Humanized Computing.

[32] Weiwei Z, Qingshan J, Tengke X. An intelligent antiphishing strategy model for phishing website detection. In Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on, 2012; 51–56.

[33] Kausar F, Al-Otaibi B, Al-Qadi A, Al-Dossari N. Hybrid client side phishing websites detection approach.International Journal of Advanced Computer Science and Applications (IJACSA) 2014; 5(7):132–140.

[34] Buber, E., Diri, B., & Sahingoz, O. K. (2017). NLP based phishing attack detection from URLs. In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (Eds.), Intelligent systems design and Applications, springer international Publishing, cham (pp. 608–618)

[35] Jain, A. K., & Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using autoupdated white-list. EURASIP Journal on Information Security.

[36] https://www.phishtank.com, Retrieved January 02, 2019.

[37] Rao, R. S. &, & Pais, A. R. (2018). Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and Applications.

[38] James, Joby & L, Sandhya & Thomas, Ciza, (2013) ”Detection of phishing URLs using machine learning techniques,” 304-309. 10.1109/IC-CC.2013.6731669.

[39] Chiew KL, Chang EH, Sze SN, Tiong WK. Utilisation of website logo for phishing detection. Computers & Security 2015; 54:16–26.

[40] Philippe De R, Nick N, Lieven D, Wouter J. TabShots: client-side detection of tabnabbing attacks. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, Hangzhou, China, 2013

[41] Lam I-F, Xiao W-C, Wang S-C, Chen K-T. Counteracting phishing page polymorphism: an image layout analysis approach. Advances in Information Security and Assurance. Springer: Seoul, Korea, 2009; 270–279.