2020 16th IEEE International Colloquium on Signal Processing & its Applications (CSPA 2020), 28-29 Feb. 2020, Langkawi, Malaysia

Detecting Phishing Website Using Machine Learning

Mohammed Hazim Alkawaz
Faculty of Information Sciences & Engineering
Management and Science University
Shah Alam, Selangor, Malaysia
mohammed_hazim@msu.edu.my

Stephanie Joanne Steven
School of Graduates Studies
Management and Science University
Shah Alam, Selangor, Malaysia
stephaniejoanne_my@yahoo.com

Asif Iqbal Hajamydeen
Faculty of Information Sciences & Engineering
Management and Science University
Shah Alam, Selangor, Malaysia
asif@msu.edu.my

***Abstract*— Trying to gather personal information through deceptive ways is becoming more common nowadays. In order to assist the user to be aware of the access to such websites, the implemented system notifies the user through email and also pop-up, when trying to access a phishing site. This paper proposes an approach of phishing detection system to detect blacklisted URL also known as phishing websites, so that individual can be alerted while browsing or accessing a particular website. Therefore, it can be utilized for identification and authentication and become a legitimate tool to prevent an individual from getting tricked.**

Keywords: **Blacklisted, phishing, Agile Unified Process (AUP), alert, pop-up notification, Email notification, Machine Learning**

I. INTRODUCTION

Phishing can be defined as impersonating a valid site to trick users by stealing their personal data comprising usernames, passwords, accounts numbers, national insurance numbers, etc. Phishing frauds might be the most widespread cybercrime used today. There are countless domains where phishing attack can occur like online payment sector, webmail, and financial institution, file hosting or cloud storage and many others. The webmail and online payment sector was embattled by phishing more than in any other industry sector. Phishing can be done through email phishing scams and spear phishing hence user should be aware of the consequences and should not give their 100 percent trust on common security application. Machine Learning is one of the efficient techniques to detect phishing as it removes drawback of existing approach.

The objectives which is the most vital thing in proposed project is to verity the validity of the website by capturing blacklisted URLs. To notify the user on blacklisted website through pop-up while they are trying to access and to notify the user on blacklisted website through email while they are trying to access. This proposed project will allow administrator to add blacklisted URL's in order to alert user during their inquiry.

The two scope of project, which is well known as user scope and system scope. User has some responsibility towards the system. The system includes a few standards and policies that requires to be obliged in order to comply the system. The user can be notified if blacklisted website is being accessed. The admin can capture the blacklisted URL's to alert user. The system involves features like capturing blacklisted website, viewing blacklisted website, displaying pop-up notification and also displaying email notification.

II. RELATED WORK

In emerging technology industry which deeply influence today's security problems has given a non-ease of mind to some employer and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. [1] In the dimension of new era there are many security systems being developed to ensure security is given the utmost priority and prevention to be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in organization to ensure network security is not being breached. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. [2] Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, and, etc. Customer face numerous security threats like cybercrime. There are many cybercrime that are extensively executed for example spam, fraud, cyber terrorisms and phishing. Among this phishing is known as the popular cybercrime today. [3] Phishing has become one amongst the highest 3 most current forms of law-breaking in line with recent reports, and both frequency of events and user susceptibleness has enlarged in recent years, more combination the danger of economic damage. [4]

Phishing is a type of practice done on the Internet where individual data are obtained by illegal approaches.[5] It supply of obtaining sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence. Phishing attack will be enforced in varied kind like Email phishing, web site phishing, spear phishing, Whaling, Tab off his guard, Evil twin phishing etc. [6] Phishing is known as webpage violence. [7] Phishing is often done by email spoofing or texting, and it typically guides user to enter points of interest at a fake web site which look and feel the same. It tries to handle the increasing range of phishing got to be met by clients in awareness and alternative efforts to ascertain protection numerous anti-phishing tools. A number of

sites have currently created optional instruments for applications, like maps for redirection but clients ought to not utilize similar passwords anywhere on the net. [8] The primary key feature is to allow user to inquire whether visited websites is original or fake. This paper proposes a security tool called as Detecting Phishing Website Using Machine Learning.

III. LITERATURE REVIEW

The current situation that is majority of the population has been fooled into giving their personal details to hackers without noticing it. Many blacklisted website has been publish to appear as an original site in order to trap user by asking them to input their personal details. For example, password, bank account, email address and etc. Phishing activity in early 2016 was the highest ever recorded since it began monitoring in 2004.The total number of phishing attacks in 2016 was 1,220,523. Tis was a 65 percent increase over 2015. In the fourth quarter of 2004, there were 1,609 phishing attacks per month. In the fourth quarter of 2016, there was an average of 92,564 phishing attacks per month, an increase of 5,753% over twelve years. [9] According to the Anti-Phishing Working Group (APWG), there are at least 47, 324 phishing attacks and a top-ten American bank estimates that at least US$300 is lost for every hour that a phishing site remains up. [10] Machine learning is that the science of obtaining computers to act while not being expressly programmed. [11] Machine Learning was implement to develop this proposed system. Machine learning techniques identifies phishing URLs typically assess a URL based on some feature or set of features extracted from it. [12].Thus, before coming to conclusion that this was the major problem, related products were examined and compared view their libation before progressing to the proposed project.

Phishtank was proposed to carry out the inspection once a link has been pasted on the section given. This allow user to keep on track of faked website. They can copy and paste the link in order to identify whether the site that they are going to access is safe or not safe. User can use the website search feature directly or they can use information from PhishTank through its API. A search engine displayed on PhishTank website is to be used as the first method. Using its API will be the second method. API service can be avail by software builder after registering themselves on PhishTank website. Both methods mentioned above do not cost a single penny. The purpose of API's usage is for user who has basis information on software development. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [13]

PhishZoo was proposed to evaluate a new method for web phishing detection based on profiles of complex sites' appearance and content. PhishZoo makes profiles of sites comprising of the website contents and images displayed. These profiles are kept in a local folder and are either synchronized against the newly loaded sites at the time of loading or against risky sites for instance, links in email offline. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [14]

GoldPhish was proposed to perceive and report phishing sites. This was done by using optical character recognition (OCR) to recite the text from an image of the page precisely from the company logo, grasping the top hierarchical areas from a search engine, and comparing them with the current web site. The forte of the tool lies in the user's capability to recognize famous company logos. A phishing site cannot change a familiar company logo without the phishing target perceiving. Limitation of this project is there was no facility of displaying pop-up and email notification once user had access blacklisted website. [15]

IV. PROPOSED METHOD

In the effort of developing the proposed system, a project methodology has to be selected and defined, as to generate a practicable development environment and realistic schedule. Thus, this project will be done using the Agile Unified Process (AUP) Lifecycle for its abridged development period and flexible process as referred in **Figure 1**.

A baseline of hardware and software requirements are set. This is to ensure the operation system platform is capable to handle and perform the development of the system. The software that is used to develop system is using Mircrosoft Visual Studio 2010 Ultimate. Mircrosoft Visual Studio 2010 Ultimate generates the system C# was the most appropriate language to run the program. mySQL stocks up data and to implement database in this system. mySQL builds in database in the Mircrosoft Visual Studio 2010 Ultimate. In hardware, atleast 4GB RAM is required in laptop/PC to build the system .This ensures a smooth process during the development.

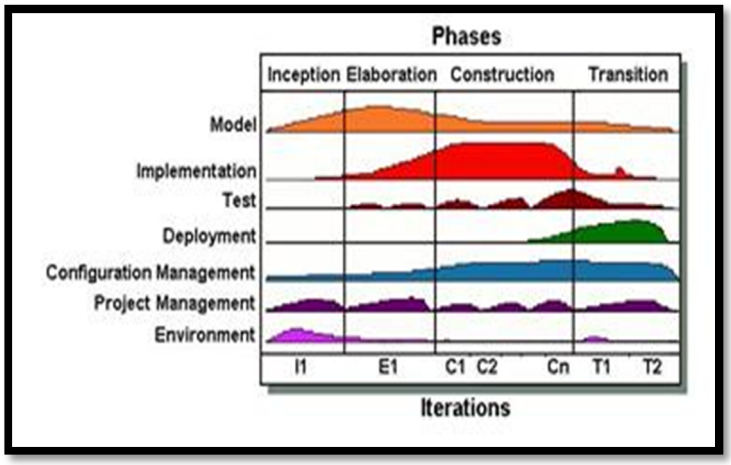

FIGURE 1: AGILE UNIFIED PROCESS MODEL

Based on **Figure 2,** admin can filter which URLs are blacklisted and which are not blacklisted by copy and pasting the URLs at "Site" row. Admin can classify blacklisted URLs as 1 and not blacklisted URL as 0. Admin can also edit, update and delete the sites once the URLs have been added.

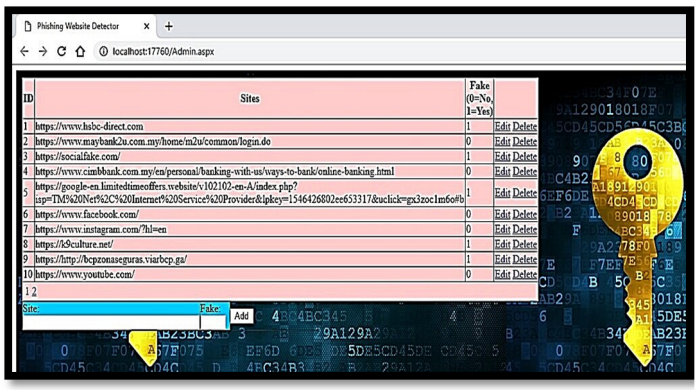

FIGURE 2: ADMIN PAGE

Based on **Figure 3,** this is a main page where user can classify which is blacklisted URLs and which is not blacklisted URLs by the color. Blacklisted URLs are in red colored row and non- blacklisted URLs are in white colored row. User can also inspect them by clicking on the URLs.

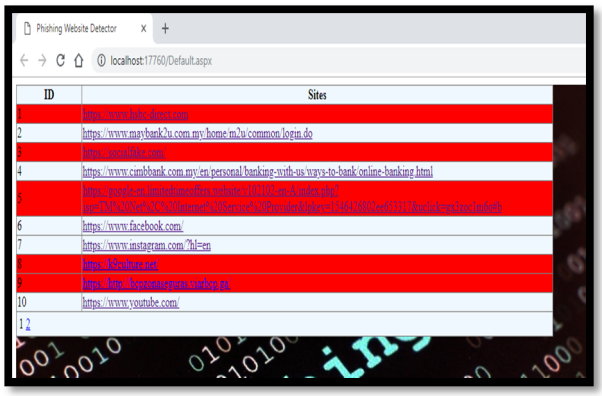

FIGURE 3: MAIN PAGE

URLs which are not blacklisted will redirect to the actual website once user clicks on it. **Figure 4** shows that a pop up notification a when user clicks on the blacklisted URL. The pop-up notification is an alert box to apprise the user by questioning whether they wish to continue knowing that it may be a phishing site.

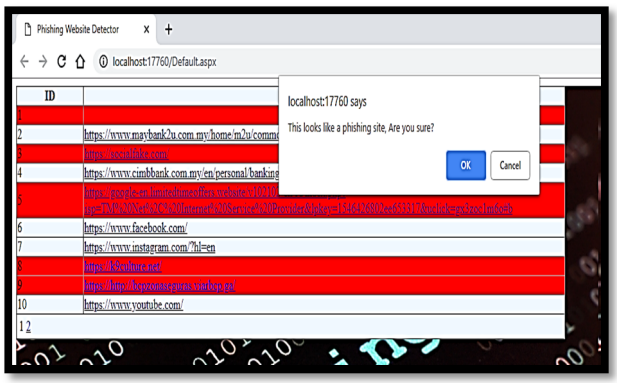

FIGURE 4: POP-UP NOTIFICATION

Based on **Figure 5**, a message from admin will be displayed once user clicks "OK" from the pop –up notification. This message will notifiy the user that the site accessed is confirmed a phishing site.

FIGURE 5: MESSAGE FROM ADMIN

Figure 6 shows a screenshot of Gmail interface. A notification from Gmail will be apprised once user clicks "OK" from the pop-up notification.

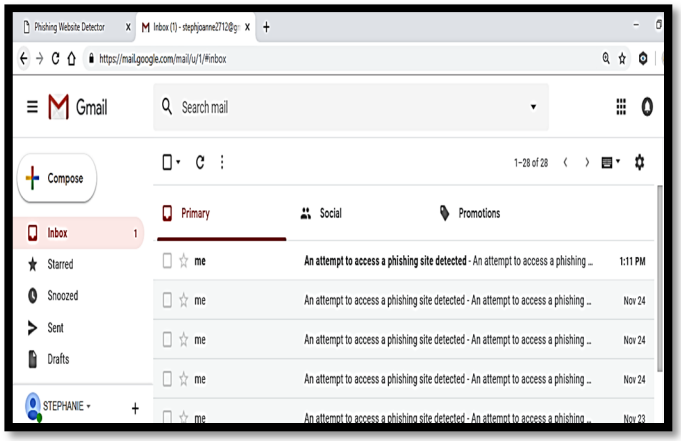

FIGURE 6: EMAIL NOTIFICATION

Based on **Figure 7**, a message from admin will be displayed as user received the Gmail notification that the site visited is confirmed a phising site.

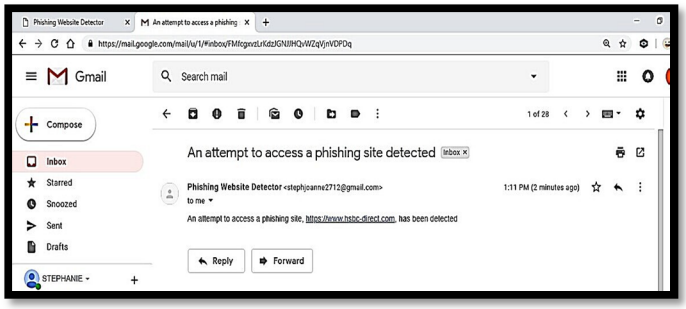

FIGURE 7: MESSAGE FROM ADMIN THROUGH EMAIL

As the overall discussion, the proposed project has been successfully coded and developed, and has met the expectations made during the project proposal phase. From the various results aggregated during the various tests, the system can be concluded that it has been successfully developed to its specifications. The system interface has been successfully designed, the functions are coded into function and the different requirements are met.

V. CONCLUSION

After reviewing and researching for appropriate monitoring tools, proposed system has been identified and chosen to address the complexity of monitoring requirement for current situation. This software is designed to show awareness of the extensive level of its functionality, features that can be displayed in the monitoring era. The system fosters many features in comparison of other software. Its unique features such as capturing blacklisted URL's from the browser directly to verify the validity of the website, notifying user on blacklisted websites while they are trying to access through pop-up, and also notifying through email. This system will assist user to be alert when they are trying to access a blacklisted website.

In conclusion, this system is designed for resources are used as intended, prevents from valuable information from leaks out, produce better control mechanism and alerts the user to keep their private information safe. Like any other programs, there are improvements which could be made into this system. Based on the capabilities which the current system processes, text message integration would a great recommendation that could be made to improve the program in the future. The future version of the application could also implement an option to directly notify the blacklisted website with a text message. The program could be made to access the list as an attachment. This text message integration function would further the usability of the application.

ACKNOWLEDGEMENT

Authors are grateful to the School of Graduate Studies and the Faculty of Information Sciences and Engineering, Management and Science University, Malaysia for their support.

REFERENCES

[1] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis"

[2] Rishikesh Mahajan (2018) "Phishing Website Detection using Machine Learning Algorithms"

[3] Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[4] David G. Dobolyi, Ahmed Abbasi (2016) "PhishMonger: A Free and Open Source Public Archive of Real-World Phishing Websites"

[5] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpage"

[6] Purvi Pujara, M. B.Chaudhari (2018) "Phishing Website Detection using Machine Learning : A Review"

[7] Satish.S, Suresh Babu.K (2013) "Phishing Websites Detection Based On Web Source Code And Url In The Webpage"

[8] Tenzin Dakpa, Peter Augustine (2017) "Study of Phishing Attacks and Preventions"

[9] Ping Yi (2018) "Web Phishing Detection Using a Deep Learning Framework"

[10] Jalil Nourmohammadi Khiarak (2017) "What is Machine Learning"

[11] Sadia Afroz, Rachel Greenstadt (2018) "PhishZoo: An Automated Web Phishing Detection Approach Based on Profiling and Fuzzy Matching"

[12] Arun Kulkarni, Leonard L. Brown (2019) "Phishing Websites Detection using Machine Learning"

[13] Rohan Saraf , Mayur Khatri , Mona Mulchandani (2014) "Phish Tank-A Phishing Detection Tool"

[14] Sadia Afroz, Rachel Greenstadt (2017) "PhishZoo: Detecting Phishing Websites By Looking at Them"

[15] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis"