Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

No Fish; Total Anti-Phishing Protection System

R.A.H.Devdunie Pabasara[1], Dhanushka Niroshan Atimorathanna[2], Tharindu Shehan Ranaweera[3] and Jayani Rukshila Perera[4]

[1] Sri Lanka Institute of Information Technology

Received: 10 December 2019 Accepted: 1 January 2020 Published: 15 January 2020

Abstract

Phishing attacks have been identified by researchers as one of the major cyber-attack vectors which the general public has to face today. Although software companies launch new anti-phishing products, these products cannot prevent all the phishing attacks. The proposed solution, ?No Fish? is a total anti-phishing protection system created especially for end-users as well as for organizations. In this paper, a realtime anti-phishing system, which has been implemented using four main phishing detection mechanisms, is proposed. The system has the following distinguishing properties from related studies in the literature: language independence, use of a considerable amount of phishing and legitimate data,

Index terms— cyber-attack, anti-phishing, information security, machine learning,

1 Introduction

owadays, with advances in technology, internetrelated crimes have increased at an alarming rate [1]. Among these crimes, phishing is one of the most popular cyber-attack vectors, which is a serious threat to information security and especially to the global economy. In phishing attacks, attacker develops web pages mimicking original websites and sends out fake emails, impersonating as a trusted entity such as popular brands or organizations, asking for sensitive information such as username, password, phone number, credit card details and other personal information. Internet users should be aware of phishing attacks as it has been in the cyber domain for years. However, many people still tend to fall victim and leak confidential information through suspicious web pages.

There are common ways of fighting phishing attacks. One way is to train employees to recognize the gravity of phishing attacks and their consequences. Awareness plays a crucial role in phishing prevention [2]. However, it is not practical to train employees or users on every possible phishing scenario. It is only human nature to be distracted and deceived. The other way is to block domain URLs and IPs, which are known from previous phishing attacks. However, hackers constantly create new domains to hunt fresh IPs [3]. The proposed "NoFish" identifies the website which the user is about to visit. It identifies logos and important features of a website using machine learning to detect the website which is being visited by the user. The visual similarity between the legitimate website and the current website is compared to get more accurate results. "NoFish" has an email client plugin for the Microsoft Outlook email client, which is implemented using content-based approaches and client-based programming languages. It should be downloaded to the Microsoft Outlook email client and it detects spam emails and extract URLs from the email body for further analysis. "NoFish" uses different classification algorithms, machine learning (ML), and natural language processing (NLP) based features [4]. NLP is proposed for URL analysis. It detects phishing URLs that users are about to visit. When using untrusted internet connections such as public WiFi services, DNS based anti-phishing approach, and HTTPS certificate transparency checking system are used to protect against DNS related phishing attacks [5] [6]. The system provides a feedback mechanism to enhance user experience through a dashboard. 'NoFish' innovative solution detects all kinds of phishing attacks, including future ones after a super simple deployment next to the user's email client and web browser. It implements a simple email client plug in and browser extension for users.

2 II.

3 Related Work

Phishing is a major security issue that needs to be addressed. Internet users should be aware of phishing attacks because this has been around for years. However, many domestic users still tend to get tricked by these phishing attempts. Therefore, everyone needed a good software-based solution to overcome this human error. In recent years, industry and academia have proposed several anti-phishing solutions to counter the phishing threat. Some of the important methods are discussed below.

4 a) Document Object Format

Document Object Format (DOM) is a languageindependent and cross-platform programming interface for XML, XHTML, and HTML documents [7]. The DOM is an object-oriented representation of the web page. The DOM-based phishing detection solutions use the similarity of a DOM tree on a suspicious web page and a legitimate web page to detect phishing. Since attackers always imitate a legitimate web page and create phishing web pages, the layout of the page is expected to be the same. Rosiello et al. have proposed a solution that alerts users when they use the same information on different websites, such as the same username and password [7].

5 b) Content-based comparison

Content-based comparison often attempts to compare the text of a web page through machine learning. Using the TF-IDF, the most used algorithm for extracting text and information from the web page, al. Zhang developed a content-based system to identify phishing websites [8]. Basnett et al. Evaluate their performance using various machine learning techniques, including neural networks, SVM (Support Vector Machine), and SOM (Self-Organizing feature Map) **??**8] [9].

6 c) Signature-based technique

Huang proposed a unique signature-based method to identify legitimate websites using text keywords and images on the website [10]. The system compares the signature of the currently open website with the signature database when a user tries to log in to a new website. If the domain name is changed but the signature matches, the web page will be declared as phishing. When a user visits a website for the first time, the system generates the signature and saves it to the database. Therefore this detection only works for the previously visited website sites, and it cannot detect zero-hour phishing attacks.

7 d) Phish Zoo

Afros and Greens tad have proposed a phishing detection solution called "Phish Zoo" that creates a unique profile for a website using URL, images, text content, secure connection layer (SSL) certification, and script [11]. When a user visits a website, Phish Zoo matches the current site profile with a list of legitimate sites and profiles stored in the database. As a first step, the URL and SSL certificate is compared with the stored profile. If it matches, the website is considered legitimate by Phish Zoo. Otherwise, the site's contents will be matched against appearance profiles to detect phishing attempts.

8 III.

9 Proposed Solution

In this section, the proposed phishing detection approach is explained. Phishing attacks have evolved a lot in past years such that even experienced users sometimes cannot be able to distinguish between phishing and legitimate pages. The proposed solution uses a level-based detection mechanism to identify phishing attacks in order to reduce the computational power and time consumption. Therefore it increases the performance and accuracy of the overall product than existing systems. Further, it provides protection against phishing attacks on trusted and untrusted internet connection. If the user is using an untrusted internet connection such as public WiFi, then the system checks the trustworthiness of the DNS servers [5] [12] [13]. Otherwise, it will be forwarded to the usual phishing detection mechanism. The system architecture is proposed under six main components. They are namely: NoFish uses a level-based security mechanism to detect phishing attacks. Researches have designed it in such a manner to reduce the computational power, reduce the time consumption of the NoFish clients, and to increase the performance and accuracy of the overall product. Figure 1 depicts the flow chart of the proposed system. As a first step URL will be matched with the white-list and black-list databases. The system uses this approach to identify known phishing sites, and if it is a white-listed URL, the system allows the user to visit the website. If the URL does not exist in those databases, it will be sent to the URL analyzer. Then the URL analyzer will be predicted as a legitimate or phishing URL. Based on the prediction, the system will deny the website or moves to the webpage similarity comparison stage.? Browser Extension ? URL Analysis ? Image Processing ? Email Phishing Detection ? DNS Phishing ? Feedback Mechanism

IV.

10 Browser Extention

The system has a browser extension for the Chrome web browser that must be downloaded by the user. This plays a major role in system performance, which is explained below a) Customized Whitelist and Blacklist Users can categorize websites into a white list or black-list through the extension, and it will be saved in the extension. When the user is bookmarking a website, it will automatically be added to the user-customized white-list within the extension once the phishing detection is completed. Consequently, the extension itself can allow or deny accessing a website without check with the server-side.

11 b) Extracting the URL

Extracting URL from the website the user is trying to visit is done by the extension and then it is forwarded to the NoFish server for further analysis.

12 Capturing Image of the Current Website

The extension takes a screenshot of the current web site and redirects it to the NoFish server. The current web page image is required for log detection and web page similarity comparison; hence the screenshot is forwarded to the analysis.

V.

13 Url Analysis

Many systems have been implemented to detect URL phishing attacks, and some of them have been focused mainly on email-based URL attacks only. However, phishing URLs can reach the victim in various ways. Nowadays, social media has become a major vector for phishing links. Very few of the existing solutions are still based on the old method, which is based on black-listing, and there are only a few existing systems that can-do real-time URL analysis to detect phishing attacks [14] [15]. However, they depend on language and algorithms that have been used to implement the system. The main purpose of implementing a URL analysis system such as NoFish is to detect any kind of phishing URL and secure the endusers as well as the organizations from phishing attacks better than prevailing solutions.

14 a) NoFish URL Analyzer

NoFish users can manage their own customized URL database in the extension. Therefore NoFish URL analyzer will not check URLs, which are in customized white-list and black-list available in the user browser extension, and it gives direct access to those sites. When a user browses a URL, which is not in customized data storage, the system request from the server to check it with a white-listed and a black-listed database. NoFish is not storing these databases, and it directly connects with the "Alexa" database and "Phish-Tank" database, and it uses their APIs to check the status of the URL. Alexa (Legitimate URLs) and Phish Tank (Phishing URLs) already maintain large databases orderly and authors believe it gives a better result and reduces the time to check compared to maintaining our own databases. However, according to user feedback, NoFish is maintaining its own white-listed and black-listed database to personalize the service. The database is automatically updated according to user feedback. If that URL is not belonging to one of them, that means it is a newly identified URL from the analyzer. That URL goes through the trained machine learning model and give predictions whether it is phishing or legitimate. The system shows a warning to the user if the URL is phishing. Users can acknowledge and not continue or ignore the warning. If the model gives it as a legitimate URL, it is then immediately moved to the image classification and computer vision process.

15 b) Algorithms and Model

URL analysis is a common subject in the information security domain. There are so many existing projects on phishing detection on URL analysis and have used deep neural networks. However, NoFish has simply created its analyzer using Machine Learning (ML) approach after extensive research on several existing URL analyzers. It consists of Machine Learning algorithms and Natural Language Processing (NLP) **??**4] [14]. For measuring the performance of the system, a new dataset of phishing and legitimate URLs was constructed, and the experimental results were tested on them. NoFish have used Random Forest Classifier, Decision Tree Classifier, Logistic Regression Classifier, Support Vector Machine (SVM), and Naive Bayes algorithm with NLP feature and have done modifications and fine-tuning to create a higher accuracy model [16]. NoFish uses 13 features of URLs for identifying phishing patterns of a URL such as protocol, domain, path, having IP, long URL, short URL, redirection, prefix_suffixseparation, sub domain, google index, DNS records, and https token. Test results are discussed in the test results section.

16 VI. Computer Vision for Phishing Detection

This is one of the most important stages in the system, and the goal is to categorize websites to make it easier to compare with the legitimate website layouts [17]. Figure 2 depicts how the system uses computer vision to identify the current website.

17 a) Logo Detection

For this prototype, the logo detector can identify 20 image classes, including the most popular banks in Sri Lanka, and mostly used international websites. The logo detection model was trained using the Tensor Flow software library on Google Colab. NoFish team trained several Tensor Flow object detection models [16] with our own dataset, and in every case, it returned the same accuracy levels. Those models are mentioned in the test results section, along with the accuracy rates obtained.

Since the website login pages are not very complex images, the model can classify the logos with high accuracy. Therefore, we selected NoFish has developed this algorithm using the OpenCV python library to identify the similarity between the current website and the legitimate website. First, the algorithm identifies key points in both images and compares them to identify matching key points. Then defines a rating of similarity from 0 to 10, where 0 means they are completely different and 10 means they are perfectly matched. Based on the score, the system defines security levels. If the score is greater than 5 it defines as a high possibility, and if the score is greater than 3 and lower than 5 it will define as low possibility.

18 Global

Then the system returns a warning to the user accordingly, as depicted in Figure 3.

19 Email Phishing Detection

Email phishing is a type of online scam where criminals ask users to provide sensitive information. This is mostly done by including a link that will appear to take you to the website that appears to be from a legitimate company; however, the website is bogus. About 70% of phishing scandals involve national-state or stateaffiliated actors, according to the Verizon 2018 Data Breach Investigations Report **??**18]. Phishing continues to be effective, more sophisticated, targeted, and difficult to identify. 4% of targeted people will click on the attachment, 94% of the time when the attachment is malicious. Only 17% of attacks are reported, and it usually takes 30 minutes to report. The cost of phishing for American businesses continues to grow, to more than half a billion dollars last year [1]. The email plugin detects spam emails for preventing the spams which are used by the phishers for attacks, and the plugin detects all the URLs in the email body to redirect to the existing URL analysis component to detect any phishing URLs. Detected spam emails are sent to the junk email folder, and phishing emails are blocked for users to view. A warning message will be notified for users about the phishing threat. The Yeoman generator, which is built with node.js, is used to create outlook add-in.

20 a) Proposed Model for Email Phishing Detection

21 Detect spam emails

NoFish system detects for spam emails because phishing emails are also received as spams. It uses an algorithm called Naïve Bayes Classifier to detect if the email is spam or not [16]. Naïve Bayes is part of a large Natural Language Processing toolset and can be trained better when fed with many and complete spam emails [16] [19]. They usually use a word bag to identify spam emails, a common approach to text sorting. Naive Bayes classification works by associating tokens (usually words, or perhaps other things), spam and non-spam emails, and using Bayes' theorem to calculate the probability that an email is not spam [20]. This spam filter accesses the email account using the IMAP protocol. We experienced that most of the time, spam mostly comes from Chinese email hosts. Therefore, as a special feature, we use a function to scan all the characters in the subject text. It triggers on any character that falls into the Han Ideographs Unicode Range. It simply scans the complete range for Chinese characters in Unicode and detect if it is spam or not.

Bayes classifier sets up two categories to choose from. It contains possible spam sentences, phrases, and word-lists, which are weighted against a white list. This returns its verdict as either "spam" or "mail". It is implemented to open the folder named spam on the email account and delete all emails older than ten days.

Our team has tested both the Naïve Bayes Algorithm (20) and Support Vector Machine (SVM) algorithms to detect spam emails. According to the test results, the Naïve Bayes Algorithm was used to detect spam emails. Test results are discussed in the section test results.

22 Detect phishing URLs in emails

JavaScript libraries are used to detect URLs in the email body. URLs may hide in emails in several ways as attachments, texts, images etc. These URLs are detected and redirect to the existing system called URL Analysis to determine the URL is phishing or not. If the URL is phishing, the user is notified by a popup message and blocked the phishing email for viewing. Since the NoFish has an existing system to analyze phishing URLs in advance, the accuracy of detect phishing URLs in high. It protects users from zero-day phishing attacks [2].

23 VIII.

24 Interactive Dashboard

NoFish system provides a user interactive dashboard to enhance the user experience. Users may use the interactive dashboard through the official site. It provides features for the user to explore more services that are provided by NoFish systems, such as feedback mechanism. Users can vote for black-listed URLs to verify it as a phishing or malicious website. This may be used after installing NoFish extension to the web browser.

25 IX.

26 Detecting Dns Based Attacks

When the user is connecting to a WiFi network first, the system checks whether it was saved in the user's computer. If it is a saved WiFi system, assume that it is a trusted connection. When the user is connecting to a new WiFi network, then the system checks whether the WiFi connection requires a WPA or WPA2 password. If not it is probably not secure. Further, to identify accurately, the system will ask the user whether it is public WiFi or trusted WiFi. If the WiFi is identified as untrusted, then the system will check for DNS related phishing attacks [12]. To identifying a fake DNS author [6] [5], proposed a solution that gives the IP address of the domain name of the current website using the IP Lookup API. Then using that IP address, the system can do a reverse IP lookup from the server-side and get the domain name, and by that, the system will define the DNS server is malicious or not [12][5].

27 X.

28 Test Results

In order to choose a model for logo detection our team trained several pertained models chosen from Tensor Flow object detection API with our own data set. Those models are mentioned below, along with the accuracy rates obtained. According to past researches, SVM, and Naïve Bayes has more accuracy than other algorithms when detecting spam emails [16] [9]. Within our calculation, SVM got 91.67%, and Naïve Bayes got 91.47% of accuracies, which shows the same accuracies. However, our team has identified SVM might not fast as other classification algorithms. Naïve Bayes classifier simply applies Bayes' theorem on the context of each email, with a strong assumption that the words included in the email are independent of each other. Therefore, NoFish has used the Naïve Bayes algorithm for spam detection with more success.

29 XI. Conclusion and Recommendations

In order to prevent phishing, business and consumers need to educate themselves about phishing and anti-phishing techniques. They should use current protection methods and report suspicious activities. By doing so, they can reduce their exposure to fraud and identity theft and protect their privacy. The most effective solution for phishing is to train users not to blindly follow links to websites that need to include sensitive information such as passwords. The ultimate technological solution to phishing is the significant infrastructure changes on the Internet that exceed the ability of any organization to deploy. However, there are steps that can now be taken to reduce the consumer's risk of phishing attacks. Some of those steps are:

For Corporations ? Provide a way for the consumer to validate that the email is legitimate. ? Stronger authentication on websites and emails.

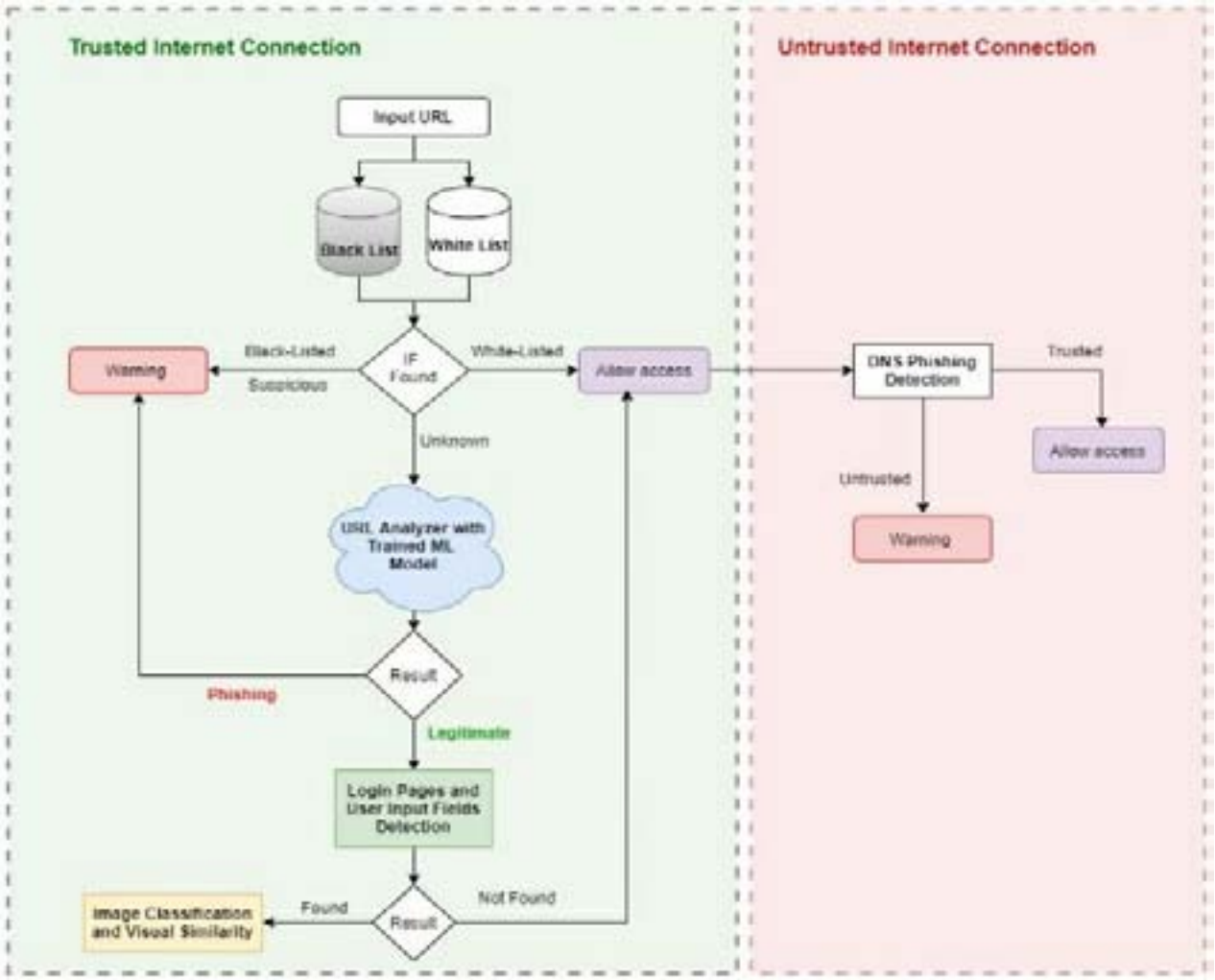
? Implement a good quality anti-virus, anti-spam, and content filtering solutions at the internet gateway.

For Consumers Be suspicious.

? Automatically detect and block malicious emails, websites, URLs, and DNS servers. ? Automatically block sensitive information from leaking to malicious parties. [1] [2]

[1] Year 2020 ( ) E © 2020 Global Journals No Fish; Total Anti-Phishing Protection System

[2] © 2020 Global JournalsNo Fish; Total Anti-Phishing Protection System



1

Figure 1: Figure 1 :

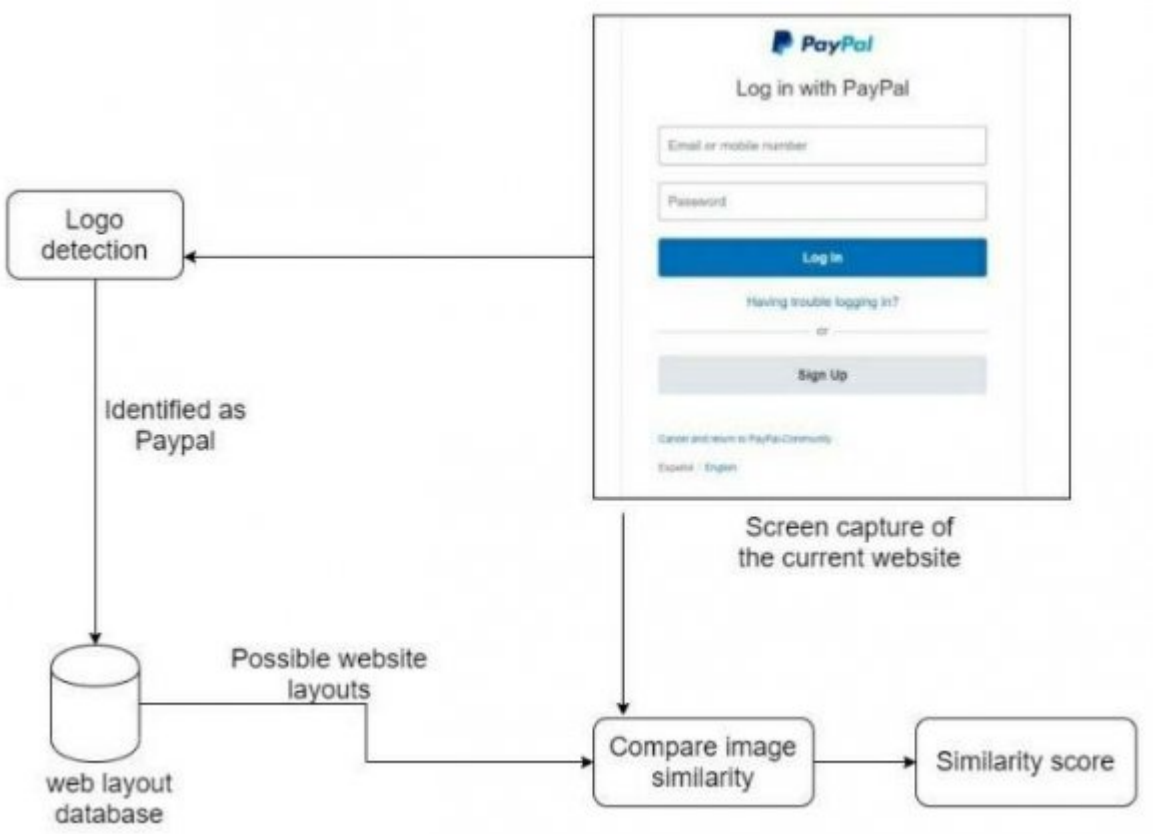
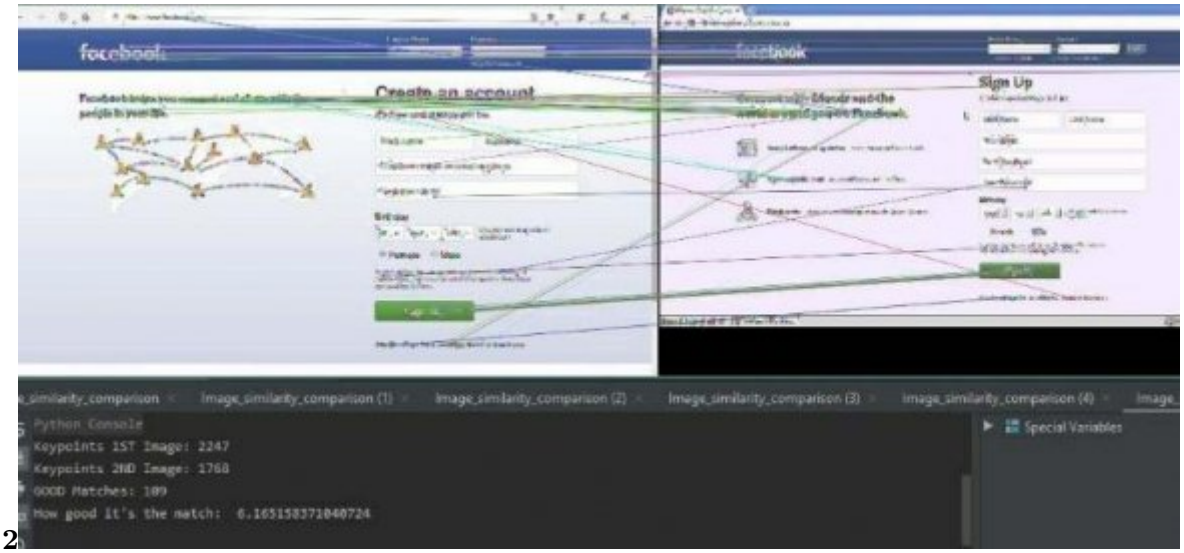
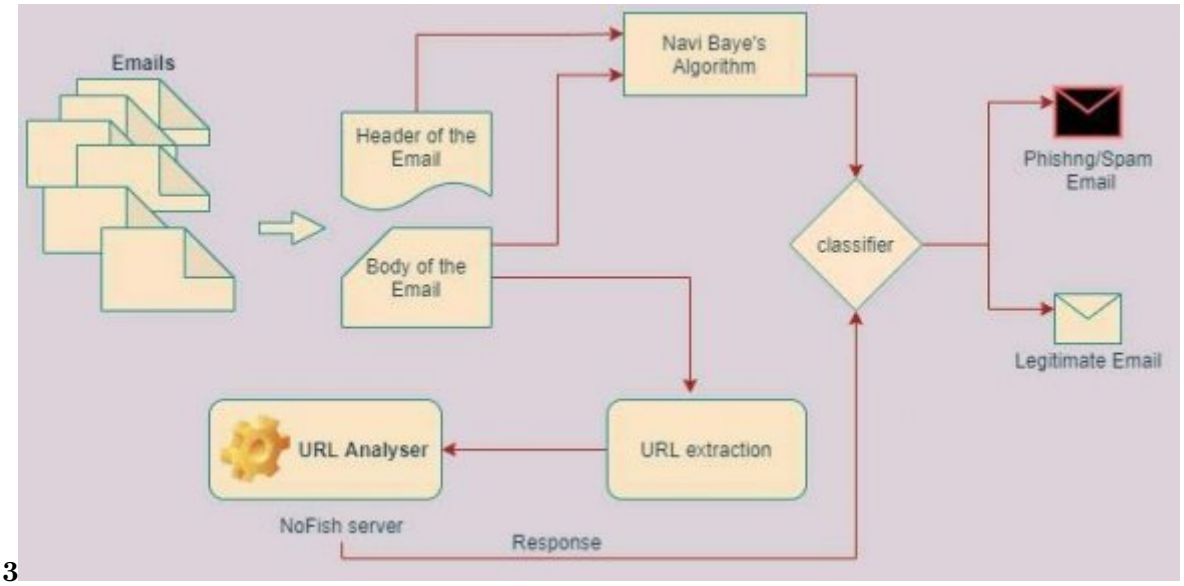


Figure 2:



2

Figure 3: Figure 2 :



3

Figure 4: Figure 3 :

? Faster_rcnn_inception_v2_coco model has a running time of 58ms per 600x600 image with mAP [^1] measure of 28 -over 95% accuracy.
? Ssd_mobilenet_v2_coco model has a running time of 31ms per 600x600 image with mAP [^1] measure of 22 -over 95% accuracy.
? Faster_rcnn_inception_resnet_v2_atrous_coco model has a running time of 620ms per 600x600 image with mAP [^1] measure of 37 -over 95% accuracy.
When evaluating the URL analyzer, all the algorithms were tested separately with large phishing and legitimate data sets and Random Forest Classifier [21][22] returned 96.257%, Decision Tree Classifier returned 84.119%, Logistic Regression Classifier returned 91.037%, Support Vector Machine returned 91.002%, and Navy Bayes returned 94.128% accuracies respectively. Consequently, in order to obtain a better accuracy level,

Figure 5:

[Berger and Merkl ()] 'A comparison of support vector machines and self-organizing maps for email categorization'. H Berger , D Merkl . *Aus DM 2005 Proc. -4th Australas. Data Min. Conf. -Collocated with 18th Aust. Jt. Conf. Artif. Intell. AI 2005 2nd Aust. Conf. Artifical Life, ACAL 2005*, 2005. p. .

[Rosiello et al. ()] 'A layout-similarity-based approach for detecting phishing pages'. P E Rosiello , E Kirda , C Kruegel , F Ferrandi . 10.1109/SECCOM.2007.4550367. *Proceedings of the 3rd International Conference on Security and Privacy in Communication Networks*, (the 3rd International Conference on Security and Privacy in Communication Networks) 2007. p. .

[Islam and Abawajy (2013)] 'A multi-tier phishing detection and filtering approach'. R Islam , J Abawajy . 10.1016/j.jnca.2012.05.009. *J. Netw. Comput. Appl* Jan. 2013. 36 (1) p. .

[Gajera et al. ()] 'A Novel Approach to Detect Phishing Attack Using Artificial Neural Networks Combined with Pharming Detection'. K Gajera , M Jangid , P Mehta , J Mittal . 10.1109/ICECA.2019.8822053. *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology*, (the 3rd International Conference on Electronics and Communication and Aerospace Technology) 2019. p. .

[Lakhita et al. ()] 'A review on recent phishing attacks in Internet'. S Lakhita , B Yadav , Pooja Bohra . 10.1109/ICGCIoT.2015.7380669. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things*, (the 2015 International Conference on Green Computing and Internet of Things) ICGCIoT 2015, 2016. p. .

[An Image-based Feature Extraction Approach for Phishing Website Detection] *An Image-based Feature Extraction Approach for Phishing Website Detection*, (Usenix-security-submission)

[Zhu et al. (2015)] 'Connection-oriented DNS to improve privacy and security'. L Zhu , Z Hu , J Heidemann , D Wessels , A Mankin , N Somaiya . 10.1109/SP.2015.18. *Proceedings -IEEE Symposium on Security and Privacy*, (-IEEE Symposium on Security and Privacy) 2015. 2015-July. p. .

[Kim and Huh ()] 'Detecting DNS-poisoningbased phishing attacks from their network performance characteristics'. H Kim , J H Huh . 10.1049/el.2011.0399. *Electronics Letters* 2011. 47 (11) p. .

[Hsu et al. ()] 'Identify fixed-path phishing attack by STC'. H Hsu , P Wang , S Pu . 10.1145/2030376.2030396. *ACM International Conference Proceeding Series*, 2011. p. .

[Carella et al. ()] 'Impact of security awareness training on phishing clickthrough rates'. M Carella , T M Kotsoev , Truta . 10.1109/BigData.2017.8258485. *Proceedings -2017 IEEE International Conference on Big Data*, (-2017 IEEE International Conference on Big Data) 2017, 2017. 2018-January. p. .

[Subasi et al. ()] 'Intelligent phishing website detection using random forest classifier'. A Subasi , E Molah , F Almkallawi , T J Chaudhery . 10.1109/ICECTA.2017.8252051. *2017 International Conference on Electrical and Computing Technologies and Applications*, 2017, 2017. 2018-January. p. .

[Sahingoz et al. (2019)] 'Machine learning based phishing detection from URLs'. O K Sahingoz , E Buber , O Demir , B Diri . 10.1016/j.eswa.2018.09.029. *Expert Syst. Appl* Mar. 2019. 117 p. .

[Huang et al.] 'Mitigate web phishing using site signatures'. Y Huang , S P Ma , W L Yeh , C Y Lin , C T Liu . 10.1109/TENCON.2010.5686582. *IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2010*, p. .

[Buber et al. ()] 'NLP Based Phishing Attack Detection from URLs'. E Buber , B Diri , O K Sahingoz . 10.1007/978-3-319-76348-4_59. *Advances in Intelligent Systems and Computing*, 2018. 736 p. .

[Afroz and Greenstadt ()] 'Phish Zoo: Detecting phishing websites by looking at them'. S Afroz , R Greenstadt . 10.1109/ICSC.2011.52. *Proceedings -5th IEEE International Conference on Semantic Computing*, (-5th IEEE International Conference on Semantic Computing) 2011, 2011. p. .

[Fang et al. ()] 'Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism'. Y Fang , C Zhang , C Huang , L Liu , Y Yang . 10.1109/ACCESS.2019.2913705. *IEEE Access* 2019. 7 p. .

[Kiruthiga and Akila ()] 'Phishing websites detection using machine learning'. R Kiruthiga , D Akila . 10.35940/ijrte.B1018.0982S1119. *Int. J. Recent Technol. Eng* 2019. 8 (2) p. . (Special Issue 11)

[Tewari et al. (2016)] 'Recent survey of various defense mechanisms against phishing attacks'. A K Tewari , B B Jain , Gupta . 10.1080/15536548.2016.1139423. *J. Inf. Priv. Secur* Jan. 2016. 12 (1) p. .

[Yang et al. ()] 'Spam filtering using Association Rules and Naïve Bayes Classifier'. T Yang , K Qian , D C T Lo , K Al Nasr , Y Qian . 10.1109/PIC.2015.7489926. *Proceedings of 2015 IEEE International Conference on Progress in Informatics and Computing*, (2015 IEEE International Conference on Progress in Informatics and Computing) 2015. 2016. p. .

[Hara et al. (2009)] 'Visual similarity-based phishing detection without victim site information'. M Hara , A Yamada , Y Miyake . 10.1109/CICYBS.2009.4925087.18. https://www.darkreading.com/threat-intelligence/14-million-new-phishing-sites-launched-each-month/d/d-id/1329955 *2009 IEEE Symposium on Computational Intelligence in Cyber Security*, 2009. Feb-2020. p. 23. (1.4 Million New Phishing Sites Launched Each Month)

[Vulnerabilities ()] "dns Vulnerabilities . *DNS Security Management*, 2017. John Wiley & Sons, Inc. p. .