Analysis Of Phishing Websites UsingAn Competent Feature-Based Framework

Bennila Thangammal.C[1], Ilamathi.K[1], Santhoshini.P[2], Nithya Shree G.P[3], Pavithra R[3] and Pooja V[3]

1 Associate professor,R.M.D Engineering college,Thiruvallur,India
{cbt.it,ilamathi.ece}@rmd.ac.in
2 Assistant professor,R.M.D Engineering college,Thiruvallur,India
santhoshini.ece@rmd.ac.in
3 UG Student , R.M.D Engineering college,Thiruvallur,India
{uec17302,uec17312,uec17315}@rmd.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

ABSTRACT: Phishing attacks take place through various forms such as email, websites and malware. To perform email phishing, attackers design fake emails which claim to be arriving from a trusted company. They send fake emails to millions of online users assuming that at least thousands of legitimate users would fall for it. Phishing attacks are one of the most common and least defended security threats today. Objective of study to identify phishing attacks using five machine learning algorithms. The proposed system handles feature selection through learning algorithm, after feature selection, training and prediction is done. The objective of our study to find an efficient algorithm, which achieves highest accuracy.

Introduction

In this cyber world, most of the people communicate with each other either through a computer or a digital device connected over the Internet. The number of people using e-banking, online shopping and other online services has been swelling due to the availability of convenience, comfort, and assistance. An attacker takes this situation as an opportunity to gain money or fame and steals sensitive information needed to access the online service websites. Phishing is one of the ways to steal sensitive information from the users. It is carried out with a mimicked page of a legitimate site, directing online user into providing sensitive information. The term phishing is derived from the concept of 'fishing' for victims sensitive information. The attacker sends a bait as mimicked webpage and waits for the outcome of sensitive information. The replacement of 'f' with 'ph' phoneme is influenced from phone phreaking, a common technique to unlawfully explore telephone systems. The attacker is successful when he makes a victim to trust the fake page and gains his/her credentials related to that mimicked legitimate site. Anti-Phishing Working Group (APWG) is a non-profit organization which examines phishing attacks reported by its member companies such as iThreat Cyber Group, Internet Identity (IID), Mark Monitor, Panda Security and Forcepoint. It analyzes the attacks and publishes the reports periodically. It also provides statistical information of malicious domains and phishing attacks taking place in the world.

Online users fall for phishing due to various factors such as:

1. Inadequate knowledge of computer systems.
2. Inadequate knowledge on security and security indicators. (In the current scenario, even the indicators are being spoofed by the phishers.)
3. Inadequate attention to warnings and proceeding further by undermining the strength of existing tools. (abnormal behavior of toolbars)
4. Inadequate attention to the visual deceptive text in URL and Website content.

Phishing attacks are one of the most common and least defended security threats today. Objective of study to identify phishing attacks using five machine learning algorithms. The proposed system handles feature selection through learning algorithm, after feature selection, training and prediction is done. The objective of our study to find an efficient algorithm, which achieves highest accuracy. Phishing attacks take place through various forms such as email, websites and malware. To perform email phishing, attackers design fake emails which claim to be arriving from a trusted company. They send fake emails to millions of online users assuming that at least thousands of legitimate users would fall for it. Phishing is the process whereby someone attempts to obtain your confidential information, such as your passwords, your credit card number, your bank account details or other information protected by the Data Protection Act. Such attempts, often referred to as Phishing attacks, are usually primitive and obvious; however, please be aware that they are becoming more sophisticated

Fig. 1 Overview of Phishing Attack

In website phishing, attacker builds a website which looks like a replica of legitimate site and draws the online user to the website either through advertisements in other websites or social networks such as Facebook and Twitter etc. Some of the attackers are able to manage phishing websites along with security indicators such as green padlock, HTTPS connection etc. In Malware phishing, attacker inserts a malicious software such as Trojan horse into a compromised legitimate site without the knowledge of a victim. According to APWG report, 20 million new malware samples were captured in the first quarter of 2016. The vast majority of the late malwares are multifunctional, i.e., they steal the information, make the victims system as a part of botnet or download and install different malicious software without client's notification. Spear Phishing target a specific group of people or community belonging to an organization or a company. They send emails which pretend to be sent by aIn website phishing, attacker builds a website which looks like a replica of legitimate site and draws the online user to the website either through advertisements in other websites or social networks such as Facebook and Twitter etc. Some of the attackers are able to manage phishing websites along with security indicators such as green padlock, HTTPS connection etc. In Malware phishing, attacker inserts a malicious software such as Trojan horse into a compromised legitimate site without the knowledge of a victim. According to APWG report, 20 million new malware samples were captured in the first quarter of 2016. The vast majority of the late malwares are multifunctional, i.e., they steal the information, make the victims system as a part of botnet or download and install different malicious software without client's notification. Spear Phishing target a specific group of people or community belonging to an organization or a company. They send emails which pretend to be sent by a colleague, manager or a higher official of the company requesting sensitive data related to the company. The main intention of general phishing is financial fraud, whereas spear phishing is a collection of sensitive information. Whaling is a type of spear phishing where attackers target bigger fish like executive officers or high profile targets of private business, government agencies or other organizations. There are many anti-phishing techniques proposed in the literature to detect and prevent phishing attacks. We have categorized these anti-phishing techniques into 4 categories. List-based techniques: Most of the modern browsers such as Chrome, Firefox and Explorer etc. follow list based techniques to block phishing sites. There are two types of list-based techniques such as whitelist and blacklist. The whitelist contains a list of legitimate URLs which can be accessed by the browsers. The browser downloads the website, only if the URL is present in the whitelist. Due to this behavior, even the legitimate websites which are not whitelisted are also blocked resulting in high false positives. The blacklist contains phishing or malicious URLs which are blocked by the browsers in downloading the webpages. Due to this behavior, the phishing sites which are not blacklisted are also downloaded by the browser resulting in high false negatives. These nonblacklisted phishing sites are also called as Zero-day phishing sites. A small change in the URL is sufficient to bypass the list-based techniques. Frequent update of these lists is mandatory to counter the new phishing sites.

Literature Survey

Zhongliu Zhuo et. al, proposed a website modeling method based on profile hidden Markov model (PHMM) which is widely used in bioinformatics for DNA sequencing analysis. Their technique explicitly accounts for possible hyperlink transitions made by users when fingerprinting a target website, and therefore can work in a more realistic environment than existing methods. Using SSH and Shadowsocks, they collected various data sets and conduct extensive evaluations. They also showed that their approach could work both in webpage and website identification in a closed world setting.

Paulo Jorge Costa Nunes et. al., proposed a benchmark for assessing and comparing static analysis tools in terms of their capability to detect security vulnerabilities. The benchmark considers four real-world development scenarios, including workloads composed of real web applications with different goals and constraints, ranging from low budget to high-end applications.

Penetration testing is a crucial defense against common Web application security threats such as SQL injection and cross-site scripting attacks. Hsiu-Chuan Huang et. al., proposed Web vulnerability scanner which automatically generates test data with combinative evasion techniques, significantly expanding test coverage and revealing more vulnerabilities.

Francesco Mercaldo et. al., proposed BehaveYourself!, an Android application able to discriminate a trusted application by a malicious one extracting opcode-based features. Our application is open and flexible: it can be used as a starting point to define, and experiment with, additional features.

In this article Rafal Kozik et. al., addressed the problem of automated Hypertext Transfer Protocol (HTTP) request structure analysis applied to web layer cyber attacks detection. In this method, they proposed a multiple HTTP sequences clustering algorithm combined with the machine-learnt classifier. The main goal behind this approach is the fact that they used the request structure and the statistical measurements of its content in order to detect anomalous behaviour of 17 connections established between client and server.

Gonzalo De La Torre Parra et al. (2020) developed two security mechanism namely a Distributed Convolutional Neural Network (CNN) to detect phishing and Distributed Denial of Service (DDoS) attack and a cloud-based temporal Long-Short Term Memory for detecting botnet attacks. Their distributed CNN model was embedded with machine learning engine in the users IoT device.

Patrick Lawson et al. (2020) explored the interaction between targeted user and persuasion principle was used in the domain of email phishing attack. They forecast vulnerabilities in phishing emails by using signal detection framework.

Justinas Rastenis et.al. (2020) formulated e-mail based phishing attack as six stages of attack. Each stage has at least one measure to categorize the attacks. Each stage have sub-section to explain the all variety of phishing attacks. They compared their proposed taxonomy with other similar taxonomies and identified their taxonomy performs well in terms of number of stages, measures and distinguished sections.

PROPOSED SYSTEM

Methodology

In this project we are going to analysis the data of phishing from multiple website through different algorithm

- First we have extract the required data from imported data. Data is to filter
- Secondly we have to classify the data base on the text through Support Vector Machine.
- Then we are going to check the data of a phishing website through their URL, Prefix and suffix.
- Then we are going to implement the other algorithm to classify the feature of data.
- Finally Compare accuracy of the data based on Different Algorithm performed

❖ Add new heuristic features with machine learning algorithms to reduce the false positives in detecting new phishing sites.

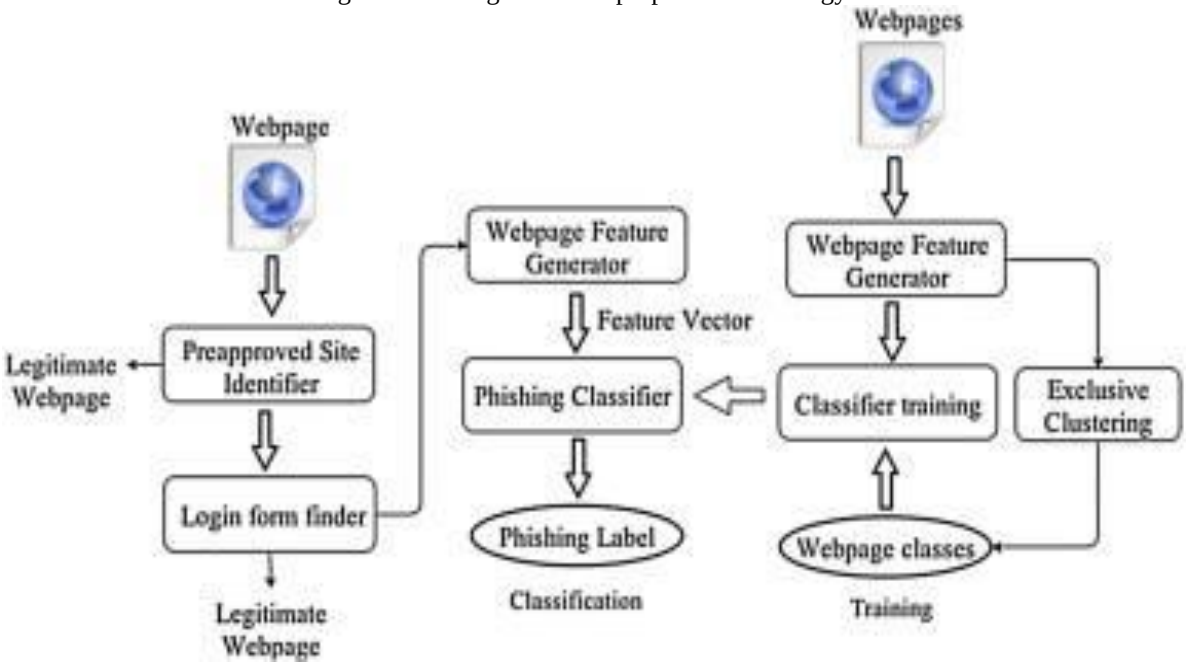
❖ Made an attempt to identify the best machine learning algorithm to detect phishing sites with high accuracy than the existing techniques.

❖ Used five machine learning algorithms (Logistic regression (LR), KNN, Random Forest (RF), support vector machine (SVM) and Decision Tree) to classify the websites as legitimate and phishing.

❖ Based on the experimental observations, Random Forest outperformed the others.

❖ The choice of considering these machine learning algorithms is based on the classifiers used in the recent literature.

Fig 2 Block Diagram of the proposed methodolgy



Modules

✓ Data Extraction

✓ Feature selection by SVM

✓ Data Analysis using URL, Prefix, Suffix, SSL etc

✓ Feature Selection by Logistic Regression

✓ Feature Selection by KNN, Decision Tree

✓ Result Analysis Modules

Data Extraction

• User Training Approaches - end-users can be educated to better understand the nature of phishing attacks, phishing and non-phishing messages.

- Software classification approaches - these mitigation approaches aim at classifying phishing and legitimate messages on behalf of the user in an attempt to bridge the gap that is left due to the human error or ignorance.

Feature selection by SVM

• SVM has been used successfully in many real-world problems Text (and hypertext) categorization

• The classification of natural text (or hypertext) documents into a fixed number of predefined categories based on their content.

• A document can be assigned to more than one category, so this can be viewed as a series of binary classification problems, one for each category

Data Analysis using URL, Prefix, Suffix, SSL etc.

• Helping to identify legitimate websites.

• Browsers alerting users to fraudulent websites.

• Eliminating Phishing Data.

• Monitoring and takedown

Feature Selection by Logistic Regression

The logistic function Interpretation of coefficients

• Continuous predictor (X)

• Dichotomous categorical predictor (X)

• Categorical predictor with three or more levels (X)

Feature Selection by KNN,

Decision Tree Hierarchical tree (Decision Tree) structure for classification

• Each internal node specifies a test of some feature

• Each branch corresponds to a value for the tested feature

• Each leaf node provides a classification for the instance KNN

• Given new test instance x,

• Compare it to all stored instances
• Compute a distance between x and each stored instance xt
• Keep track of the k closest instances (nearest neighbours)
• Assign to x the majority class of the k nearest neighbours

Result Analysis

• Preventing a phishing attack before it begins
• Detecting a phishing attack
• Preventing the delivery of phishing messages
• Preventing deception in phishing messages and sites
• Counter measures
• Interfering with the use of compromised information

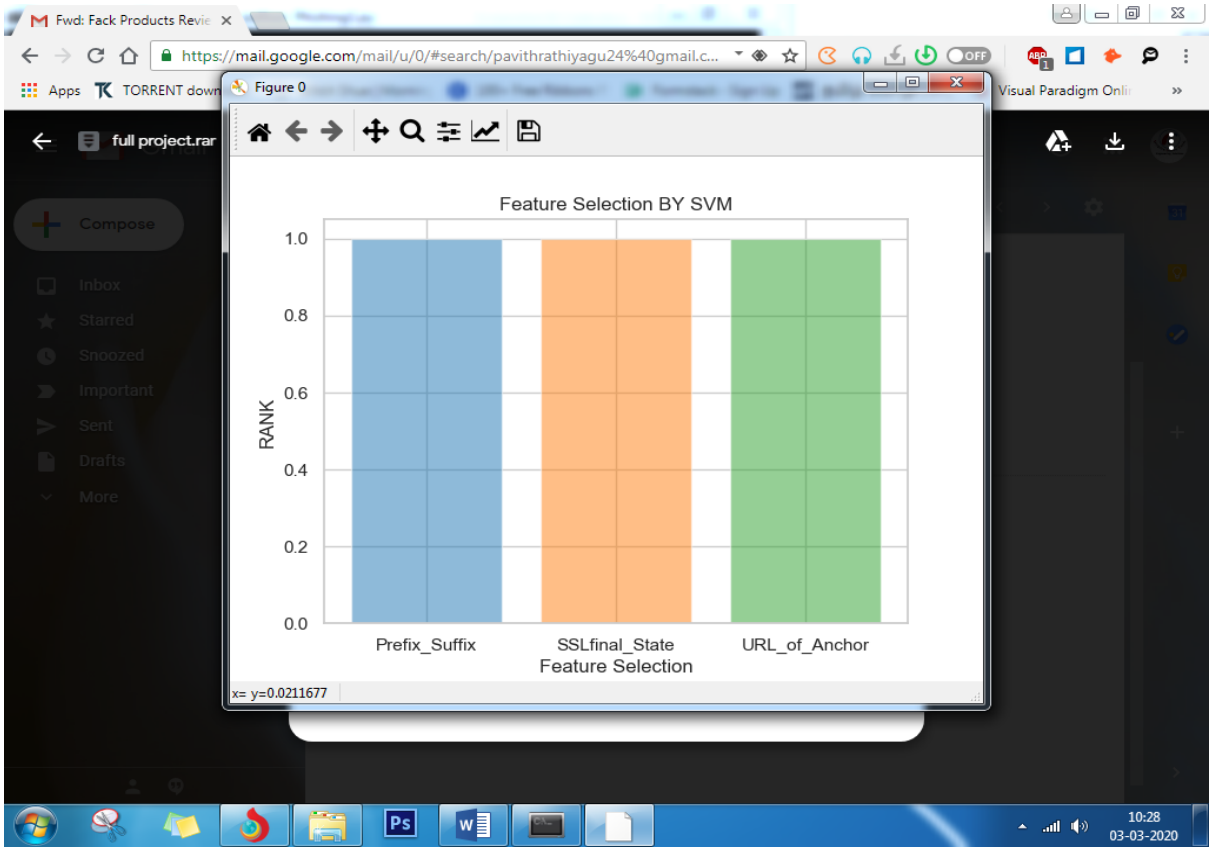


Fig 3 Feature selection by SVM

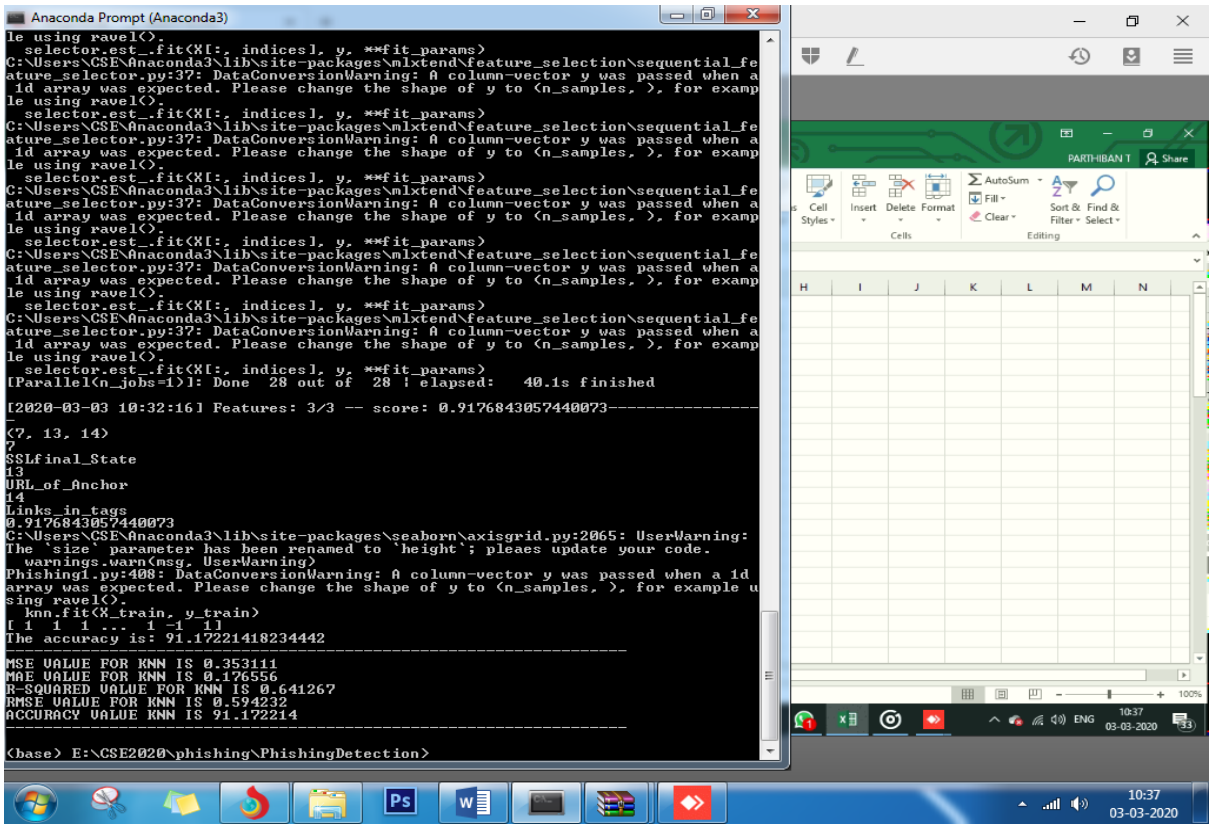


Fig 4 Delivery of phishing messages

CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure. The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that Random forest based classifiers are the best classifier with great classification accuracy of 91.42% for the given dataset of phishing site. As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

References :

1. Zhongliu Zhuo, Yang Zhang, Zhi-Li Zhang, Xiaosong Zhang, and Jingzhong Zhang, "Website Fingerprinting Attack on Anonymity Networks Based on Profile Hidden Markov Model", IEEE Transactions on Information Forensics and Security, Vol:15, No:5, May 2018
2. Paulo Jorge Costa Nunes, Iberia Medeiros, José Fonseca, Nuno Neves, Miguel Correia, and Marco Vieira, "Benchmarking Static Analysis Tools for Web Security", IEEE Transactions on Reliability , Volume: 67, Issue: 3, Sept. 2018.
3. Hsiu-Chuan Huang, Zhi-Kai Zhang, Hao-Wen Cheng, and Shiuhpyng Winston ,"Web Application Security: Threats, Countermeasures, and Pitfalls", in Computer, vol. 50, no. 06, pp. 81-85, 2017
4. Francesco Mercaldo, Corrado Aaron Visaggio, Gerardo Canfora, and Aniello Cimitile ,"Mobile malware detection in the real world", ICSE '16: Proceedings of the 38th International Conference on Software Engineering Companion , Pages 744–746,May 2016
5. Rafal Kozik, Michal Choras, and Witold Holubowicz,"Packets tokenization methods for web layer cyber security", *Logic Journal of the IGPL*, Volume 25, Issue 1, Pages 103–113, February 2017.
6. Patrick Lawson, Carl J. Pearson, Aaron Crowson, Christopher B. Mayhorn, Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy, Applied Ergonomics, Elsevier, vol. 86, pp. 1-10, 2020.
7. Gonzalo De La Torre Parra , Paul Rad, Kim-Kwang Raymond Choo, Nicole Beebe, Detecting Internet of Things attacks using distributed deep learning, Journal of Network and Computer Applications, Elsevier, vol. 163, pp. 1-13, 2020.

8. Justinas Rastenis, Simona Ramanauskaite, Justinas Janulevicius , Antanas Cenys , Asta Slotkiene and Kestutis Pakrijauskas, E-mail- Based Phishing Attack Taxonomy, Applied Sciences, MDPI, vol. 10, pp.1-15, 2020.