Feature Selections for the Machine Learning based Detection of Phishing Websites

Ebubekir Buber, Önder Demir
Marmara University Institute of Pure and Applied Sciences
Computer Engineering Department
Istanbul, Turkey
ebubekirbbr@gmail.com, odemir@marmara.edu.tr

Ozgur Koray Sahingoz
Turkish Air Force Academy
Computer Engineering Department
Istanbul, Turkey
sahingoz@hho.edu.tr

***Abstract*— Phishing websites are malicious sites which impersonate as legitimate web pages and they aim to reveal users important information such as user id, password, and credit card information. Detection of these phishing sites is a very challenging problem because phishing is mainly a semantics-based attack, which especially abuses human vulnerabilities, however not network or system vulnerabilities. As a software detection scheme, two main approaches are widely used: blacklists/whitelists and machine learning approaches. Machine learning solutions are able to detect zero-hour phishing attacks and they have superior adaption for new types of phishing attacks, therefore they are mainly preferred. To use this type of solution features of input must be selected carefully. The whole performance of the solution depends on these features. Therefore, in this paper, it is aimed to list and identify the important features for machine learning-based detection of phishing websites.**

***Index Terms*—phishing, features, machine learning, URL, domain names.**

I. INTRODUCTION

In the last few decades, Internet technology has been growing so pervasive from online social networking to online e-commerce and banking technologies to make people's lives more comfortable. Due to this uncontrollable growth, many security threats to networks systems are emerged: the mostly encountered one is "phishing". Phishing is a web-based attack in which attackers try to reveal some sensitive information such as user id/passwords or account information by sending an email from a reputable entity or person or communicate with other channels.

Many users unwittingly click phishing domains every day and every hour. The attackers are targeting not only the users but also the companies. According to the 3rd Microsoft Safer Index Report [1], released in February 2014, the annual impact of phishing attacks could be very high as $5 billion. The main reason of this huge cost is the lack of awareness of users. But security defenders must take precautions for users to not confront from these harmful sites. To prevent phishing damages these defenders mainly try to increase the consciousness of the company and build strong security mechanisms which can detect and prevent phishing attacks before they cause too much damage.

In a phishing attack, typically, a victim receives an e-mail message that appears to be sent by a known contact or organization. This message contains some web links which are targeting some malicious software targeting for the user's computer or has links to direct victims to malicious websites to trick them into divulging personal and financial information, such as passwords, account of Information Systems (IS) or credit card details.

According to [15] there are five main reasons for people to fall for phishing;

- They do not aware of the URLs (Uniform Resource Locator) and their usage
- They do not know which displayed URLs can be trusted.
- Due to the redirection or hidden URLs they do not access/see the target URL
- They can accidentally click some URLs, or they have no much time for consulting the URL
- They cannot distinguish legitimate URLs from phishing ones

All countries are attacked by the phishers, however, especially developing countries are the main target of the attackers. Phishing activity trends report of Anti-Phishing Working Group (APWG) in the last quarter 2016 emphasized that the world's most-infected country was China. It is followed by *Turkey* and Taiwan as depicted in Table 1[16].

TABLE I. MOST INFECTED COUNTRIES

Ranking	***Country***	***Infection Rate (%)***
1	China	47,09
2	Turkey	42,88
3	Taiwan	38,98
4	Guatemala	38,56
5	Ecuador	36,54
6	Russia	36,02
7	Peru	35,75
8	Mexico	35,13
9	Venezuela	34,77
10	Brazil	33,13

If we look from the attackers side, primary motives behind phishing attacks are focused in three main categories [2];

978-1-5386-1880-6/17/$31.00 ©2017 IEEE

- *Financial Gain*: attackers can utilize the stolen credentials for their financial advantages
- *Identity Hiding:* attackers may sell these stolen identities (such as user id/passwords), instead of using them directly, to any other criminals who are seeking a way to hide their activities and identities.
- *Fame and Notoriety:* attackers might attack victims for the sake of peer recognition.

Phishing is popular between attackers, since it is easier to trick someone by motivating to click a malicious link which seems legitimate than trying to break through a computer's defense mechanisms. The malicious links within the body of the message are designed to make it appear that they go to spoofed organization using that organization's logos and other legitimate materials.

Phishing attacks can be considered as a layered problem; technical and human layers. An effective mitigation would require addressing issues at both layers.

Since phishing attacks target at exploiting weaknesses found in human (i.e. end-users), it is difficult to mitigate them effectively. According to [3], end-users failed to detect nearly 30% of phishing attacks even when trained with good user awareness educations. It is also important to design powerful phishing detection systems as well as to raise awareness of users.

Once the phishing attack is detected, a number of actions could be applied to the attacks and it is easy to take down detected phishing attacks. Our focus in this study is the techniques of detecting phishing campaign. Phishing detections approaches are categorized into two main groups, because it is a two-layered structure; human layer and technical layer. An overview of phishing detection approaches is given in Fig. 1.

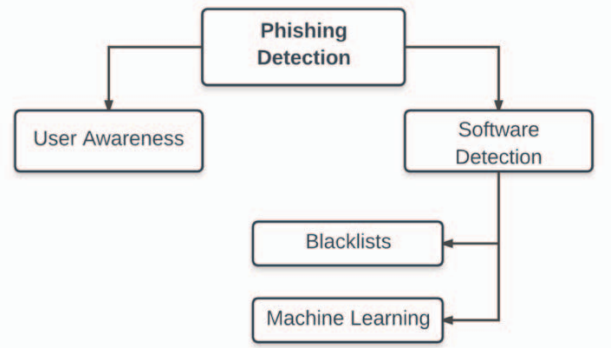

Figure 1. An Overview of Phishing Detection Approaches

The primary focus of this study is covering the software detection approaches of phishing attack detection. It is explained blacklist-based phishing detection approaches in Section II, it is analyzed phishing domains and extracted its distinguishing features from legitimate domains, and how they can be detected using machine learning and natural language processing techniques are explained in Section III.

II. Phishing Detection by Blacklists

Blacklists are frequently updated lists of previously detected phishing URLs or IP addresses. Blacklists generally have lower False Positive (FP) rates than machine learning based phishing detection systems [4]. However, blacklists do not provide protection against zero-hour phishing attacks. As mentioned in [5] blacklists were able to detect only 20% of phishing attacks. Therefore, blacklists are found to be ineffective against detecting zero-hour phishing attacks.

The study [4] shows that 47% to 83% of phishing URL were blacklisted after 12 hours. This time interval is very significant because 63% phishing campaigns end within the first 2 hours.

There are some blacklist providers such as Google Safe Browsing API [6], DNS-Based Blacklist, PhishNet [7], Automated Individual White-List [8].

- *Google Safe Browsing API* enables client applications to validate whether a given URL exist in blacklists that are constantly updated by Google [6]. This service consists of 2 blacklists which are named goog-phish-shavar and goog-malware-shavar for phishing and malware respectively. Google Safe Browsing Service is not fully completed yet. However, it is used by Google Chrome and Mozilla Firefox.
- *DNS-Based Blacklist (DNSBL)* providers use the standard DNS protocol. When an SMTP connection is established, the system can verify whether the connection source is listed in phishing blacklists [9] by sending a DNS A RR query to a DNSBL server. Due to its use of standard DNS specification, any configured DNS server could act as a DNSBL.
- For detecting phishing URL, it is required to exact match between analyzed URL and any URL in blacklists. If any changes are made to a Phishing URL, it would result in no match. This means, even a URL highly similar to phishing URL cannot be detected as Phish. To solve exact match limitation, *PhishNet* [7] process blacklisted URLs (Parents) and produce multiple variations of the parent URL via 5 different methods.
- *Automated Individual White-List (AIWL)* maintains a whitelist of features describing trusted Login User Interfaces (LUIs) where the user submitted his/her credentials. Every LUI will cause a warning except if trusted. If a LUI labeled as trusted, its features will be stored locally in a whitelist. The structure of AIWL is consist of two main components; Whitelist, Automated Whitelist Maintainer. Whitelist is a list of trusted LUIs. Automated Whitelist Maintainer track user login activity and decide a URL is whether suspicious or not. Classification metric of the maintainer is that an end-user logs in successfully for enough amount times via target LUI, then that LUI is trusted. When a LUI is detected as trusted, it is added to the whitelist.

III. Phishing Detection by Machine Learning

Phishing web page detection can be considered as a document classification or clustering problem, where models are created by taking advantage of Machine Learning algorithms such as k-Nearest Neighbor (k-NN), Decision Tree,

Random Forest, Support Vector Machines (SVM), k-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), etc.

Clustering algorithms (e.g. k-means, DBSCAN) partition instances in an unsupervised manner. To construct the clusters, it is not required to know the class labels of instances. These algorithms cluster instances according to similarities between instances. The main purpose is to cluster similar instances is a single group.

Generally, clustering algorithms based on some assumptions. For example, there are two separate classes for the phishing detection problem. One class must represent phishing instances while the other represents legitimate ones. The assumption says that the cluster which has low instance count represents the phishing instances, and the cluster which has high instance count represents the legitimate ones. This is the main assumption of the phishing detection algorithms.

Detecting phishing domains is considered as a classification problem. Therefore labeled data which have samples as phishing domains and legitimate domains in the training phase are needed. The dataset which will be used in the training phase is one of the crucial point to build successful detection mechanism. Detection Systems should use samples whose classes are precisely known. So, the samples which are labeled as phishing must be absolutely detected as phish. Likewise, the samples which are labeled as legitimate must be absolutely detected as legitimate. Otherwise, the system cannot work correctly if we use samples that we are not sure about the class information. For this purpose, a number of public datasets are created for phishing. Some of the well-known ones are PhishTank [10] and TechHelpList [11]. These data sources are commonly used in academic researches.

Collecting legitimate domains is another problem. For this purpose, site reputation services are commonly used. Web site reputation services analyze and rank available websites. This ranking may be global or may be country-based. Ranking Mechanism depends on a wide variety of features. The websites which have high rank scores are identified as legitimate sites which are used very frequently. One of the well-known reputation ranking service is Alexa [12] Researchers are using top lists of Alexa for legitimate sites for studies of phishing detection in academic researches.

When we have raw data for phishing and legitimate sites, the next step should be processing these data and extract meaningful information from it to detect fraudulent domains. The dataset to be used for machine learning must consist these features. So, we must process the raw data which is collected from Alexa, PhishTank or other data resources, and create a new dataset to train our system with machine learning algorithms. The values should be selected according to our needs and purposes and should be calculated for every one of them.

In the phishing detection problem, we have two classes; “Phishing” and “Legitimate”.

To build detection mechanism, system calculate the features that we’ve selected according to our needs and purposes with labels, and use it for train. After that, instances which has no label are classified by system.

A. Characteristics of Phishing Domains

To understanding how attackers think when they create a phishing domain, it is needed to know URL structure. Uniform Resource Locator (URL) is created to address web pages. Figure 2 shows relevant parts in the structure of a typical URL.

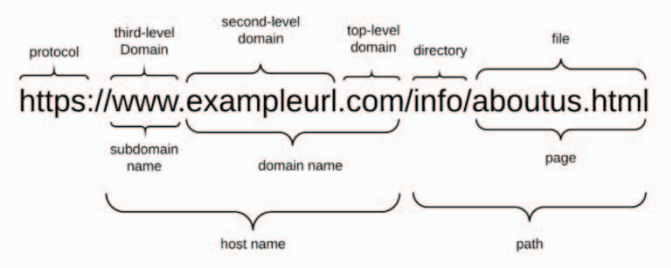

Figure 2. The Structure of URL

URL begins with a protocol used to access the page. The fully qualified domain name identifies the server who host the web page. It consists of a registered domain name (second-level domain) and suffix which we refer to as top-level domain (TLD). The domain name portion is constrained since it has to be registered with a domain name Registrar. A host name consists of a subdomain name and domain name. A phisher has full control over the subdomain portions and can set any value to it. The URL may also have a path and file components which, too, can be changed by the phisher at will. Subdomain name and path are fully controllable by the phisher. In this study, we use the term “Free URL” to refer to those parts of the URL in continuation of the article.

The attacker can register any “Domain Name” that has not been registered before. This part of URL can be set only once. The phisher can change Free URL at any time to create a new URL. Unique part of the web site is “Domain Name”, that’s why the security defenders struggle to detect phishing “Domain Names”. When a domain detected as a fraudulent, it is easy to prevent this domain before a user access to it.

Some threat intelligence companies [13, 14] detect and publish fraudulent web pages or IPs as blacklists, thus preventing these harmful assets by others is getting easier.

The attacker must intelligently choose the domain names because the aim should be convincing the users and then setting the Free URL to make detection difficult. An example of phishing URL is given in Fig.3.

http://paypal.com-webappsuserid29348325limited.active-userid.com/webapps/89980/

protocol	http://
Domain name	active-userid.com
path	/webapps/89980/
Subdomain item1	com-webappsuserid29348325limited
Subdomain item2	paypal
Words	paypal, webapp, user, id, limited, active, userid

Figure 3. An Example of Phishing URL

In the example which is given in Fig.3, the attacker tried to make the domain look like "paypal.com" by adding Free URL, although the real domain name is "active-userid.com". When the user sees "paypal.com" at the beginning of the URL, they can trust the site and connect it, then can share their sensitive information to this fraudulent site. This is a frequently used method by attackers.

Other methods that are often used by attackers are Cybersquatting and Typosquatting.

Cybersquatting (also known as domain squatting), is registering, trafficking in, or using a domain name with bad faith intent to profit from the goodwill of a trademark belonging to someone else. The cybersquatter may offer selling the domain to a person or company who owns a trademark contained within the name at an inflated price or may use it for fraudulent purposes such as phishing. For example, the name of your company is "abcompany" and you register a domain as "abcompany.com". Then phishers can register "abcompany.net", "abcompany.org", "abcompany.biz" and they can use it for fraudulent purposes.

Typosqutting, also called URL hijacking, is a form of cybersquatting which relies on mistakes such as typographical errors made by internet users when inputting a website address into a web browser or based on typographical errors that are hard to notice while quick reading. URLs which are created with Typosquatting looks like a trusted domain. A user may accidentally enter an incorrect website address or click a link which looks like a trusted domain, and in this way, they may visit an alternative website owned by a phisher. A famous example of Typosquatting is **goggle.com**, an extremely dangerous website. Another is **yutube.com** which is like **goggle.com** except it targets Youtube users. Similarly, **www.airfrance.com** has been typosquatted as **www.arifrance.com**, diverting users to a website peddling discount travel. Some other examples; **paywpal.com**, **microroft.com**, **applle.com**, **appie.com**. Another example of usage like this given in Fig.4.

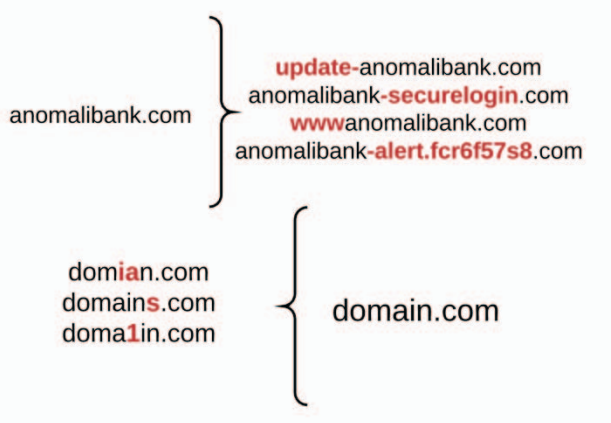

Figure 4. Typosquatting Examples

B. Features Used for Phishing URL Detection

There are a lot of algorithms and a wide variety of data types for phishing domain detection in the literature and commercial products. A phishing URL and corresponding page have several features which can be differentiated from a malicious URL. For example; an attacker can register a long and confusing domain name to hide the actual domain name (Cybersquatting, Typosquatting). In some cases, attackers can use direct IP addresses instead of using the domain name. This type of event is out of our scope, but it can be used for the same purpose. Attackers can also user short domain name which is irrelevant to any legitimate brand and does not have any Free URL addition. These types of websites are also out of our scope because they are more relevant to fraudulent domains instead of phishing domains.

Different kinds of features which are used in machine learning algorithms in the detection process of academic studies are used. Features collected from academic studies for the phishing domain detection with machine learning techniques are grouped as given below.

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

It is explained these feature groups in the continuation of this section.

1) URL-Based Features

URL is the first thing to analyze a website to decide whether it is a phishing URL or not. URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of the URL-Based Features are given below.

- Digit count in the URL
- Total length of URL
- Checking whether the URL is Typosquatted or not
- Checking whether it includes a legitimate brand name or not
- Number of subdomains in URL
- Is TLD one of the commonly used one?

2) Domain-Based Features

The purpose of phishing domain detection is detecting phishing domain names. Therefore, passive queries related to domain name which we want to classify as phishing or not provide useful information to us. Some useful Domain-Based Features are given below.

- Is domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?
- Is the registrant name hidden?

3) Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give us information about how much reliable a website is. Some of Page-Based Features are given below.

- Global Pagerank
- Country Pagerank

- Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There some paid services for obtaining these types of features.

- Estimated number of visit for the domain ın a daily, weekly, or monthly basis
- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Counting references from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

4) Content-Based Features

Obtaining these types of features requires active scan to target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

- Page Titles
- Meta Tags
- Hidden Text
- Text in the Body
- Images etc.
- Screen Shots of webpages

By analyzing these information, we can gather information such as;

- Is it required to login to the website?
- Website category
- Information about audience profile etc.

All of these explained features are useful for phishing domain detection. In some cases, it may not be logical that using some of these features, because of some limitations. For example, it may not be logical to use Content-Based Features for developing a fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000 in a day. Another example, if we want to analyze newly registered domains Page-Based Features are not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism. Which features are needed to use in the detection mechanism should be selected carefully.

IV. Conclusion

Phishing is a major problem, which uses both social engineering and technical deception to get users' important information such as financial data, emails, and other private information. Phishing exploits human vulnerabilities, therefore, most protection protocols cannot prevent the whole phishing attacks. Many of them use the blacklist/whitelist approach, however, this cannot detect zero-hour phishing attacks, and they are not able to detect new types of phishing attacks. Therefore, the use of machine learning for phishing detection is an effective and efficient tool in this domain. To use ML approach, lots of (labeled) data are needed, and additionally, the features of these data are very important. Therefore, in this paper, it is aimed to list and identify the important features for machine learning-based detection of phishing websites.

References

[1] Microsoft, "Microsoft Security Index Report, "https://news.microsoft.com/en-sg/2014/02/11/microsoft-consumer-safety-index-reveals-impact-of-poor-online-safety-behaviours-in-singapore/#sm.0000c8bivc14h3dyfvxak6545kbcz#4FXDf2H3VbYmD1b1.97 , accessed May 2017.

[2] W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web based systems." in Proceedings of the *13th IEEE Symposium on Computers and Communications (ISCC 2008)*. Marrakech, Morocco: IEEE, July 2008, pp. 326-331.

[3] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proceedings of the 28th international conference on Human factors in computing systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.

[4] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *Proceedings of the 6th Conference in Email and Anti-Spam*, ser. CEAS'09, Mountain view, CA, July 2009.

[5] Khonji, M., Iraqi, Y. and Jones, A., 2013. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials, 15(4), pp.2091-2121.

[6] Google, "Google safe browsing API," http://code.google.com/apis/safebrowsing/ , accessed May 2017.

[7] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in INFOCOM'10: *Proceedings of the 29th conference on Information communications*. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.

[8] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in *DIM '08: Proceedings of the 4th ACM workshop on Digital identity management*. New York, NY, USA: ACM, 2008, pp. 51–60.

[9] Rbldnsd, http://www.corpit.ru/mjt/rbldnsd.html, accessed on May 2017.

[10] PhishTank, https://www.phishtank.com/, accessed on May 2017.

[11] TechHelpList, https://techhelplist.com/pastes/, accessed on May 2017.

[12] Alexa, http://www.alexa.com/about, accessed on May 2017.

[13] Cymon, https://cymon.io/, accessed on May 2017.

[14] Firehol, http://iplists.firehol.org/, accessed on May 2017.

[15] M. Volkamer, K. Renaud, B. Reinheimer, A. Kunz, "User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn", Computers & Security (2017), doi: 10.1016/j.cose.2017.02.004

[16] Anti-Phishing Working Group (APWG), ''Phishing activity trends report — last quarter 2016. http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf, accessed on May 2017