Categorization of Phishing Detection Features

And Using the Feature Vectors to Classify Phishing Websites

by

Bhuvana Namasivayam

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved May 2017 by the
Graduate Supervisory Committee:

Rida Bazzi, Chair
Huan Liu
Ziming Zhao

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

Phishing is a form of online fraud where a spoofed website tries to gain access to user's sensitive information by tricking the user into believing that it is a benign website. There are several solutions to detect phishing attacks such as educating users, using blacklists or extracting phishing characteristics found to exist in phishing attacks. In this thesis, we analyze approaches that extract features from phishing websites and train classification models with extracted feature set to classify phishing websites. We create an exhaustive list of all features used in these approaches and categorize them into 6 broader categories and 33 finer categories. We extract 59 features from the URL, URL redirects, hosting domain (WHOIS and DNS records) and popularity of the website and analyze their robustness in classifying a phishing website. Our emphasis is on determining the predictive performance of robust features. We evaluate the classification accuracy when using the entire feature set and when URL features or site popularity features are excluded from the feature set and show how our approach can be used to effectively predict specific types of phishing attacks such as shortened URLs and randomized URLs. Using both decision table classifiers and neural network classifiers, our results indicate that robust features seem to have enough predictive power to be used in practice.

ACKNOWLEDGMENTS

I would like to thank Dr. Bazzi, Dr. Liu and Dr. Zhao who supported my work and helped me get results of better quality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Phishing is a major online security concern. According to latest Google Safe Browsing report, Google search blacklists over 50,000 malware sites and over 90,000 phishing sites monthly [30]. The APWG (Anti – Phishing Working Group) reported that the number of phishing attacks in 2016 was 65% more than 2015 [29]. In the last 12 years, the number of phishing attacks per month has increased 5753%. Kaspersky Labs reported that its anti-phishing system was triggered over 30.8 million phishing sites during the second quarter of 2015 [31].

The damage caused by phishing attacks is as extensive as it is diverse. It was reported that loss due to phishing attacks was $65 million in 2015 [29]. Kaspersky also reported that criminals in Eastern Europe had used phishing attacks to access more than 100 banks from 30 different countries over the past few years [31]. A total of 557 different brands or institutions were targeted by phishers in the first quarter of 2014 [42]. Almost any enterprise that takes in personal data via the Web is a potential target. The main objective of a phishing attack is usually financial in nature causing data breaches and leakage of confidential information pertaining to individual users and organizations and resulting in huge losses to individuals and companies [31]. MarkMonitor found that companies in the Retail and Financial services sectors remained the top targets. Hence, detecting and minimizing the impact of phishing attacks is extremely important [29].

We discuss here some of the ways in which phishers operate. Phishers work by compromising a legitimate domain to create phishing websites or by compromising an existing website to include scripts to redirect to a malicious server where user data can be downloaded or by luring users with domain names such as pay5al.com, pay.pal.com, or paypal.sign-in.online, which look like benign site Paypal.com [5] [6] [12].

Phishers also try to confine their attack to certain regions, by using IP filters. IP filters operate at TCP/IP stack to allow or deny traffic based on rules set by the administrator. Attackers make use of this filtering technique to deny people on IP addresses outside specific domain/country, to see

the fraud sites – only people inside of domain/country can see the fraud sites. The goal is to make it more difficult for response teams at hosting provider outside of country to view the active fraud, so they cannot confirm the problems and then eliminate them. This IP filtering technique is prevalent in Brazil and was used in 29 percent of phishing attacks [29].

There are different ways in which phishing attacks can be detected. Phishing blacklists are frequently updated based on user reports and used in browser plugins to check if the website entered by the user is present in the blacklist [46] [47] [48]. Visual similarity based detection techniques are also used, where the web page is captured as an image and compared with potential target sites to detect a phishing attack [49] [50]. Several approaches in phishing detection use features extracted from URL of the website, web page content, hosting domain, traffic ranking, and popularity of the website [1 - 28]. Each approach uses a unique combination of features to differentiate between a benign and a phishing site. Once features are extracted, data mining algorithms (classification algorithms in most cases) are used for training them to build a classification model that will help classify any new website as phishing or benign.

In the literature surveyed, an exhaustive listing of all website features and finer categories under which these features can be classified is not present [1-28]. Most of the available research works do not address the robustness of the features used in phishing detection techniques. A phishing detection feature or a set of features are robust if either an adversary cannot easily create a phishing website for which the features looks like that of a benign website or even if he/she creates such a phishing website, the success probability of the phishing attack would become less. For example, we consider an adversary who operates through email spams by creating new phishing websites to lure users. An example of a non-robust feature is the number of dots in the URL. Though we find that in many phishing website URLs, the number of dots is more than that would be in a benign site URL, the adversary can carry out a phishing attack by creating a URL with lesser number of dots in it. An example for robust feature would be WHOIS registration date (age of domain), as the adversary cannot create a new phishing website with an older registration date.

Thus, a discussion on feature robustness is necessary to create phishing detection systems that are hard to break or compromise. Moreover, each research work uses a different classification algorithm to train features and different prediction performance measure, and hence comparing the works with each other is difficult. Finally, many of the existing works do not justify the datasets used to extract the website features.

The contributions of our work are as follows.

We built a broader framework to describe various phishing detection research works and to categorize the various features such as URL features, web page features, hosting domain based features, website popularity based features and network level features used in these research works into finer categories. In chapter 2, we present the list of features gathered from literature surveyed.

We discuss the robustness of different feature categories and mention set of robust features for specific phishing attacks in chapter 3. Additionally, we also state how even if some features can be good predictors individually, they many not give proper predictions when used to predict certain attacks and we provide examples for this in chapter 4.

We extract 59 features from website URL, URL redirection, hosting domain and popularity, as mentioned in chapter 5 and train classifier with combinations of URL, hosting domain and site popularity feature categories, and set of all extracted features and determine the predictive performance of robust features. To show that robust features have enough predictive power to be used in practice, we train our entire feature set using decision table classifier and neural networks and obtain classification accuracy of 96.18% and 98.16% respectively. The classification results are mentioned in detail in chapter 6. We also mention how our work can be useful to create systems that are unbiased towards any feature category and can thus predict well for new types of phishing attacks with shortened URLs or newly compromised websites.

We also discuss how neural networks can provide better classification accuracy in detection phishing websites, in chapter 7.

The benefit of this work is that any future phishing detection system can effectively make use of this exhaustive list of features under various categories and features' robustness analysis to get information on what kind of features can be extracted and used effectively in their system. Our work is a good place to start, for a new phishing detection tool that aims to use features that can are easily collected, robust and provide more accurate predictions.

CHAPTER 2

SURVEY AND FEATURE CATEGORIZATION

1. Overview

We have surveyed 28 research works that use features from website URL, page content, hosting domain, popularity etc. to train a classification model and use the model to classify phishing websites. The papers selected are a good representative of the research works that have studied the various characteristics and features specific to phishing websites.

The survey includes a consolidated list of close to 120 features from the 28 research works categorized into broader and finer categories. Combining similar features together as a category, we get 6 broader and 32 finer categories of features.

The 6 broader categories discussed are URL Features, Page and JavaScript features, Security features, Site popularity features, Hosting Domain based features and Network layer features.

Each of these broader categories and features under these categories are described below.

In this survey, we also analyze the robustness of these feature categories, as mentioned in chapter 3.

2. URL Features

URL or Lexical features are obtained based on the properties of the URL of the website. The composition of words in the domain portion, part portion and TLD (Top Level Domain) of the URL and presence of certain special characters and their positions are significant URL features that contribute to detection of phishing sites.

Most of the phishing attacks are through email scams [29] and hence attackers need a way in which they can lure users by creating a phishing website that looks exactly like an existing benign website. Phishers normally use obfuscation techniques to deceive the user to click on the phishing URL. By

obfuscating the host name with an IP address or including the target brand's domain name in the URL path with or without spelling errors or using longer URLs to embed benign looking tokens in the sub-domain or path of the URL, phishers try to lure users into accessing the phishing site, for example, www.ebay.example.com. Usage of special characters such as '@' which cause the browser to ignore the string on the left of '@' and treat the string on the right as actual URL are few other ways to trick users, for example, http://ebay.com/personal_info@www.xyz.com [13]. Phishing pages generally have multiple redirects to redirect from the initial URL to the final site which is hosted by the attacker in any compromised machine [12] [24]. Attackers also infect benign sites with a heavily obfuscated malicious JavaScript code, that embeds an iframe with attacker's malicious domain URL and then throws an HTTP 302 redirection to load the phishing website exploit domain [12].

In several research works studied, [1] [2] [3] [4] [5] [7] [8] [15] [16] and [19], bag-of-words representation [43] on the entire URL is considered as a significant feature. In this representation, the URL is processed to extract each segment delimited by special characters ('/', '.', ',', '?', '=', ';', etc.) as a token and binary features are created for each of the tokens. Bag-of-words representation for each portion of the URL (domain and path portions) is also considered in [2] [4] [5]. Length of the different parts of the URL, number of tokens in each part and count of each special character in the URL are other common features that are extracted and trained to contribute to phishing website detection [2] [4] [5] [6] [9] [10] [12] [15] [16] [19] [20] [24]. Some research works also look for presence of brand names in the URL, and domain brand name distance [3]. Domain brand name distance is the edit-distance between domain name and brand name that can be potential phishing target. Redirection features such as Number of redirects between the pages and status of redirection can help detect phishing attacks where a legitimate hosting domain is compromised to redirect to another malicious domain where the attacker steals user's information [12] [24]. The different URL features used in the research works that contribute to phishing detection are categorized as follows as mentioned in Table 1. The different features under each of these categories are listed in Table 2 (Appendix 8.1)

Table 1

Finer categories of URL features

Category
Tokens in URL – Bag-of-words approach
Presence of security sensitive, client server keywords / brand name
Presence of character codes
Presence of @, port No, IP address
Length of URL/Domain/Path
Number of dots, hyphens, underscores (special characters)
Domain/Path Token features
TLD Organization
URL Redirects

3. Page content and JavaScript features

Many features extracted from the web page content such as the presence of login forms, the presence of password fields and presence of abnormal scripting content can help in detecting a phishing site. Most phishing web pages contain forms with input fields to obtain user payment card details or password. Apart from that, several techniques such as hidden elements, popups, and prompts to enter sensitive information are used to lure users into entering passwords and confidential information [19]. JavaScript abnormalities, the presence of shell code in the page and suspicious Active X controls can also denote a phishing page – where an attacker exploits any vulnerability in the web page to inject scripts/code that can download user sensitive information

from the page to the attacker's server [21]. Many research works extract these features to detect phishing websites.

Afroz et al [8], extract the entire web page along with HTML content and images and store it as a profile. Whenever user loads a new site, it is checked against all stored profiles. If a close match is found, there are high chances that the loaded site is a phishing site [8]. Prophiler system extracts many features from JavaScript, DOM and Active X controls of the web page along with URL and Host based features to train the classifier to predict phishing websites [21]. Some research work check for the presence of forms and password field in forms, source URL match with request URL and links on page [10] [11] [13] [14] [16] as these features are claimed as good indicators of phishing websites. This makes the overall system a light weight operation with less computational overhead, when compared to systems like Prophiler [21] and the one designed by Thomas, Kurt, et al [19] which extract many pages and JavaScript features to detect phishing. The different page features used in the research works that contribute to phishing detection are categorized as follows as mentioned in Table 3. The different features under each of these categories are listed in Table 4 (Appendix 8.1).

Table 3

Finer categories of Page features

Categories
HTTP header information (server, cache control)
Presence / Absence of specific page events
Number of abnormalities in Scripting Content
HTML code, Text and Images
Number of White space, unknown tags and Hidden elements, Small area elements
Abnormal features in Forms / Presence of specific form fields

Features of Iframes, Pop ups, User Prompts and plugins
Features of Links on page / Redirect pages
Request URL
Shell code and suspicious Active X controls

4. Hosting Domain features

Features from WHOIS and DNS records of the website provide insight into how and where phishing websites are hosted. WHOIS services are provided by registrars and registries for the domain names that they sponsor. WHOIS records provide information about the registrant of the website, registration creation and expiration dates and few other details [52]. DNS records are mapping files that tell the DNS (domain name system) server which IP address each domain is associated with. When someone visits a web site, a request is sent to the DNS server and then forwarded to the corresponding IP address where the website is hosted [53].

WHOIS features such as the age of domain, the life span of a domain, registrar details and few others and DNS record features such as autonomous system number, name server address, and location, time to live value of DNS record etc. are commonly used in the literature surveyed.

Phishers normally target compromised hosting domains to launch their attacks, so that obtaining user information through the phishing site is easy. Phishing sites are created for a short span to get the maximum out of it in a few days before the site is detected and blocked. Phishing campaigns are short-lived as phishers cannot afford to pay for a hosting domain for a long period. Phishers also use free web hosting services and domain tasting [5] to host their websites for a short span of time.

TTL (Time to Live) is a setting for each DNS record that specifies how long a DNS resolver is supposed to cache the IP address before it expires and a new one needs to be queried [54]. By using lower TTL values in DNS NS (name server) records, attackers easily redirect the webpage to different IP addresses when the DNS entry refreshes each time.

Most of the research works studied use hosting domain features along with other categories of features in-order to improve the accuracy in classifying phishing websites [4] [5] [7] [9] [10] [16] [19] [21]. Some of these works use only URL and hosting domain features to achieve good classification results – 96% to 99% accuracies [4] [5] [7]. WHOIS registration dates, especially age of domain is considered as an important predictor in many of the works surveyed [3] [13] [14]. Table 5 lists the hosting domain based features used in the research works surveyed.

Table 5

Hosting Domain features

Finer Categories of Host based features	**Features under each category**
Hosting Domain features	**1.** Domain's autonomous host number, **2**. IP Address of host, **3**. Primary domain name, **4.** Domain Reputation Score, **5.** Domain Confidence level, **6.** Connection speed of host, **7**. DNS server domain and IP address, **8**. Mail server domain and IP address.
WHOIS registration Information	**1**. Age of Domain (Creation Date – Today), **2**. Life span of Domain (Creation date – Expiration date), **3.** Last Update (Updation date – Today), **4**. Number of Registration Info Available, **5.** Presence of Creation, Updation and Expiration dates, **6**. WHOIS Registrar name **7**. WHOIS Registrant name, **8**. Whether domain name applicant is an individual or enterprise?, **9**. Whether registration dates are defined.
DNS / PTR Record	**1**. Presence of PTR record, **2**. Does the PTR record in turn resolve one of the host's IP addresses?, **3**. TTL value of DNS records of hostname, **4**. Presence of DNS A record, **5.** Presence of DNS NS record, **6**. Presence of DNS MX record.

DNS IP Address	**1**. Number of Resolved IP addresses, **2.** Are IPs of A, MX and NS records belonging to same autonomous system, **3**. IP Prefix, **4.** BGP Prefix, **5.** Is the IP address in a blacklist, **6.** Country where the IP address belongs to.

5. Security Features

Presence of SSL certificate for the website, presence of public key certificate and if the website cookie is abnormal are few security related features.

While most phishing attacks run over HTTP, a significant number run on sites for which SSL certificates have been issued as certificate authorities do not scrutinize who gets their SSL certificates [51]. Bonafide certificate owners sometimes unwittingly provide facilities for phishing because their site has been compromised by an attacker. In certain cases, where phishers host their own websites, it may be difficult to obtain a fake SSL certificate as some certificates also require validation by a certificate authority [51]. Table 6 lists the different security features extracted to detect phishing.

Aburrous, et al, [9] [20] use the security features listed in Table 6 to classify phishing websites. Some approaches check for presence of SSL certificate along with other feature categories [8] [10] [11].

Table 6

Security Features

Finer Categories of Security features	**Features under each category**
SSL features	**1**. Presence of SSL, **2**. SSL Certificate match with protected site, **3**.SSL Certificate authority name, 4. SSL Certificate age
Abnormal Cookie	**1**. Abnormal Cookie - If the cookie points to its own domain which is inconsistent with claimed identity or points to real site which is inconsistent with its own domain
Public Key Certificate features	**1**. Distinguished names in public key certificate whether inconsistent with claimed identities

6. Site Popularity features

Google rank [56], Alexa traffic rank [55] and number of links pointing to the website are good indicators of how popular the website is [3] [15] [16] [18]. Alexa traffic rank is calculated by Alexa.com and provides three months of aggregated data based on number of links within site viewed by users and number of users viewed the website [55].

Phishing web pages are short lived and thus either have a very low page rank [22] or their page rank does not exist in the Alexa database. Their social reputation score, which is the number of likes/shares on Facebook and Twitter would be less. We extract the features as listed in Table 7 to get details on the popularity and social reputation of a website. We use this feature category along with URL and hosting domain features for phishing website detection as these features are easier to collect and make the phishing detection system a light weight operation [3] [15] [16] [22].

Table 7

Site Popularity features

Finer Categories of Site popularity based features	**Features under each category**
Page Rank	**1**. Google Page Rank, **2**. Page rank of hosting site
Website Traffic	**1**. Real traffic rank of that site from Alexa.com, **2**. Number of visitors, **3**. Number of pages they visit
Number of links pointing to site from Search Engines	**1**. Domain Google links, **2**. Domain Baidu links, **3**. Domain Bing links, **4**. Domain Yahoo! Links, **5.** SLD (Second level domain) Google links, **6**. SLD Baidu links

7. Network Features

Malicious websites use rich web resources that can cause multiple HTTP requests sent to the web server, including multiple redirections, iframes and external links to other domain names [24]. Hence network layer information such as the number of packets sent and received to establish connection and number of ports opened on the web server can help in detecting phishing websites. Table 8 lists some of them mentioned by Li et al [24] to detect phishing websites.

Table 8

Network Features

Finer Categories of Network Layer features	**Features under each category**
TCP Information	**1**. Number of TCP packets sent to remote server by crawler, **2**. Number of distinct TCP ports of web server used, **3**. No of Remote IP addresses connected by crawler
Application layer communication	**1**. Number of bytes of data from/to web server, **2**. Number of UDP packets generated, **3**. Number of TCP urg (urgent flag set) packets, **4**.Number of data packets from / to crawler, **5**. Average local packet rate, **6**. Average remote packet rate
DNS queries	**1**. Number of DNS queries sent by crawler, **2**. Response time of DNS server
Traffic Flows	**1**. Inter-arrival time between consecutive flows, **2**. Number of flows generated during entire life cycle, **3**. Duration of each flow.

8. Comparison with similar surveys

Our survey includes a consolidated list of all features from 26 research works that detect phishing websites by using the features extracted from the webpages, whereas the survey by Khonji et al [4] only explains six such research works on website feature extraction and classification.

Comparing our work with Survey by Doyen et al [38], which was independently published in Jan 2017, we have categorized URL features from 27 research works and hosting domain features from 16 research works, as compared to their work that surveys URL features from 12 research works and hosting domain features from 10 research works. The other feature categories have a similar coverage in both the surveys. We have categorized similar features into finer categories and provide a better representation of the feature categorization, as compared to the other surveys [4] [38]. We also analyze the robustness of different feature categories and how the features can be effectively used to predict specific phishing attacks and new types of phishing attacks with shortened URLs or newly compromised websites. This analysis has not been done in the other surveys studied.

CHAPTER 3

ROBUST FEATURES

1. Overview

A phishing detection feature or a set of features are robust if either an adversary cannot easily create a phishing website for which the features looks like that of a benign website or if it is costly for the adversary to create a website that, with high probability, is indistinguishable from a benign site. Most URL features such as number of dots in URL, length of URL etc. can be modified by adversary to look like a benign website and so these features are not robust individually. But in certain cases, such as when long URL names are used by adversary to trick users, these features when considered together can be more robust than individual features as discussed in detail in this chapter.

Similarly, considering DNS features, the website owner has access to change the mail server, name server and PTR records for his/her website and hence these features are not robust. Attackers who host and own their phishing websites can modify these features so that the feature looks like that of a benign site. Thus, a discussion of the robustness of different categories of features is necessary to arrive at a set of robust features that can be used effectively to detect phishing websites.

In this chapter, we have considered the main types of phishing attacks such as compromising websites / servers, obfuscating URLs to lure users and creating new phishing campaigns and discuss the robustness of features we extracted under each category that can help detect these attacks. We justify robustness of these features by explaining why crafting features to make website appear as a benign website would reduce attacker's profitability. We also discuss few non-robust features from the literature surveyed.

2. Threat Model

We assume an adversary with capability to create a phishing attack by either of the following ways.

a) Adversary is limited to phishing through email spams and makes use of existing compromised hosting domains to host the new phishing website. Adversary crafts URLs (long URLs or shortened URLs) with legitimate looking tokens in-order to lure users into clicking it.

b) Adversary compromises security of legitimate hosting domains/website to embed malicious scripts that would redirect a benign website to a phishing website belonging to the adversary. So, when user visits the benign site, he/she is redirected to the phishing website that would steal user's information.

The following are the assumptions on adversary's ability to create phishing website with features crafted in such a way that would make the website to look like a benign website.

a) We assume adversary has limited affordability to create multiple hosting servers for load balancing and hence mail servers and name servers of the phishing website are hosted in the same hosting domain.

b) We assume adversary cannot compromise WHOIS registry databases and is external to Alexa.com, WOT, SEOquake [44] and Google servers. We also assume that these servers are secure and cannot be compromised by adversary.

c) We assume adversary can modify DNS mail server (MX), name server (NS) and PTR records for a website that he/she has hosted and owns.

d) Adversary can create shortened URL names to lure users, or hyperlinks that hide the actual destination domain.

e) Some attack types such as configuring IP filters in network to block detection systems, sending deceptive email attachments and impersonation of an executive to authorize fraudulent wire transfers are beyond the scope of this thesis work [57].

3. Robustness of URL Features:

The goal of the adversary in crafting a phishing URL is to deceive users to click on it and at the same time, evade detection by phishing detection systems.

We analyzed several phishing URLs from phishtank and we find that most of them have the characteristics such a long URLs, more number of dots, more number of tokens in the URL and presence of words such as online, verify, secure etc. to deceive users into thinking that is a benign site.

Phishing detection systems that work on URL features, commonly check for length of domain name, length of path, number of dots/ special characters in the URL, presence of brand name etc. Features such as brand name presence and brand name distance if changed by attacker in such way that makes the phishing website appear benign, can reduce the probability of user being victimized by the phishing attack. For example, URLs with brand names embedded in it, such as pay.pal.com and pay5al.com can be detected using these brand name features. If attacker crafts a malicious URL with domain name - funpal.com, the purpose of phishing is lost if the URL cannot lure users into clicking on it. Hence brand name features are robust.

Similarly, assume adversary creates a phishing URL, for example, http://slindau.ch/STD/Standardbank.co.za/index.php, to appear as a webpage from standard bank website. This website can be detected using features such length of hostname, length of URL and number of forward slash characters in URL. Alternately, if the phishing website created is www.slin.ch, the chances of user clicking on it is lesser. So, to in-order to make a successful phishing attack, the URL created by the adversary should be long enough to include deceptive names (online, verify, account etc) or brand names to deceive users into clicking on it.

Though we assume that long URLs with above characteristics can lure users better than short URLs, there are lot of attacks in recent times that use shortened URL names or hyperlinks that hide the actual URL.

Hence, robustness of URL features varies depending on the threat model considered. As mentioned above, for phishing attacks involving long URLs in the email to lure users, URL features are robust. But, in other types of email spams such as those that include shortened URLs, randomized URLs and hyperlinks to hide URLs, these features are not robust.

4. Robustness of WHOIS features

We use WHOIS records obtained from website registrars and ICANN WHOIS database, which cannot be forged or compromised easily. In many phishing campaigns, phishing websites are newly created and hosted for a short span to steal user information, as we have mentioned in our first threat model. If adversary creates a phishing site with a longer life span, he can evade detection by a system that uses these features, but he might have to pay for the domain name until it expires, even if is captured and blocked before its expiry by phishing detection systems. Hence, for attacks of this type, WHOIS registration dates, age of domain and life span of domain are robust [5] [6].

5. Robustness of Site popularity features

Adversary will not be able to forge or modify Alexa traffic rank, SEMRush from SEOquake [44], Google +1 count and WOT scores. Any newly created phishing website either has a lower Alexa rank or there is no entry in Alexa database. So, for detecting attacks created using new phishing websites, site popularity features are robust.

6. Robustness of DNS / IP features

As many phishing websites are hosted in a single compromised hosting domain, it is easier to find such domains using features such as autonomous system number, IP address prefix of

hosting domain etc [21]. Due to limited affordability, attacker usually creates the name server, mail server and other services in the same hosting domain's infrastructure. Otherwise, cost of additional hosting domains can be higher than what the attacker can get out the phishing campaign and hence can result in loss to the attacker. This makes the features, DNS ASN, IP Prefix, 'are IP addresses of DNS A, MX and NS records present in the same autonomous system', Mail server and name server records, robust.

7. Robustness of URL redirection features.

Phishing attack created by redirecting a benign website to a phishing website, can be detected by checking for number of redirects between the initial page and final landing page and redirect status for each redirect. These features cannot be forged and with fewer redirects attacker would not be able to make a successful redirect to his/her malicious server. We find that one of the 4 ways in which malicious web pages are used is attacks is through redirection [12]. Hence we consider these features as robust.

8. Analysis of Non-robust features.

We identified the following non-robust features from the literature surveyed. We argue that an adversary can create sites for which these features are indistinguishable from those of benign sites while not affecting the adversary's ability to launch a phishing attack.

Table 9

Non-robust features

Feature category	**Features**
URL features	1. Presence of hexadecimal or Unicode characters in URL, 2. Presence of @ in URL, Presence of Port Number / IP address in domain portion, 3. Length of file name or directory portion

Page features	1. HTTP header tokens, fields and values, cache control
Hosting Domain features	1. Presence of DNS A, MX, NS records, 2. Mail Server address and Nameserver address

The presence of hexadecimal, Unicode or other special characters are not necessary for launching a phishing attack. The adversary can still create URLs that can deceive the user without using such characters. The presence of IP address is useful because it hides the domain of the adversary, but it is not necessary for launching a phishing attack. Similarly, the adversary should be able to register a URL to obtain DNS A, MX and NS records without divulging information about the adversary. Similarly the adversary can provide header tokens without divulging information about the adversary.

All the other features, including page features such as malicious scripting and form features, if modified to look like a benign site, the probability of a successful phishing attack would become less. Hence all other features are robust.

.

CHAPTER 4

FEATURES FOR SPECIFIC PHISHING ATTACKS

1. Overview

In the previous chapter, we have discussed the robustness of feature categories with respect to specific attacks. Though robust features make the system hard to break, including non-robust features together with robust features to create a detection system, would not negatively impact the performance of the system. But, some features even though they are good predictors individually, might reduce the performance of the system when used to predict certain types of phishing attacks. For example, we have features from the path portion of the URL in our feature set and when this feature set is used for predicting shortened phishing URLs, the URLs might be misclassified as benign. Hence, in this chapter we show that it is necessary to analyze the features that might create a bias in the results and select features that can predict well for specific phishing attacks.

2. Feature categories for shortened URLs crafted with phishkits.

Nowadays, phishkits are used by attackers to strategically craft shortened URLs that evade detection [29]. Hence URL features like number of dots in URL, length of URL etc. can prove ineffective in efficiently predicting such a phishing website.

Hence, a system that uses URL features in the feature set might misclassify shortened phishing URLs as benign. Hence, we train our classifier with a combination of hosting domain (WHOIS + DNS/IP) features, site popularity features, and URL redirection features and evaluate the results to see how we can use other feature categories, apart from URL features to effectively detect this kind of phishing attack.

3. Features used to detect phishing attack by compromising existing websites.

Existing websites have a good traffic ranking, confidence score and legitimate WHOIS features. We consider a use case where the attacker compromises the security of an existing website to embed scripts that would redirect to a malicious server at attacker's domain, through which attacker can get access to any data user enters in the website. If we use a classification model trained with all features including site popularity features and WHOIS features, this compromised website may fall into the set of false negatives, and evade detection. Hence, it is necessary to avoid site popularity features and WHOIS features and use features such as URL redirection, DNS TTL and page related features to effectively detect phishing attacks of this type.

CHAPTER 5

FEATURE EXTRACTION AND MACHINE LEARNING

1. Overview

We extract features from the website's URL, hosting domain and popularity and convert them to feature vectors (binary, numeric, nominal and text forms) to feed it to a classification algorithm. The algorithm learns a pattern from the feature vectors and defines a classification model that maps the input (feature vectors) to a target class. The target class values (phishing or benign) are also provided along with the input data for training the classifier. For building a classification model, we need feature vector inputs from both benign websites and phishing websites to predict both the classes accurately. Once a classification model is defined using the data, it can be used to predict the target class for new samples of data. The datasets we use to collect these features are described in the section 5.5. We use python libraries to extract the features and Weka [58] to define a classification model by training the feature set.

2. Feature Extraction with Python libraries.

Table 9 (Appendix 8.2) lists the various features (URL, hosting domain (WHOIS and DNS/IP) and site popularity features) we extract from both benign and phishing websites. These features are selected from literature surveyed with the intention to create a light –weight faced paced detection system.

We implemented a Python application to extract these features from the input set of URLs and store results as a csv (comma separated values) file. The input URL list is fed as a csv file to the application. URL is processed to split the domain name, path and TLD portions separately. The tokens in each part (tokens are words delimited by special characters) are obtained using python regular expression library using which number of tokens in hostname / path of URL, length of tokens and similar other features are created. These tokens together form the bag-of-words feature and are represented as word vectors [43], which is done automatically by Weka's filtered

classifier [59], when we feed the entire URL as one of the features. We also create numeric features such as length of the different parts of URL, count of tokens in each part of URL and count of each special character. Python library urlparse [39] is used to get the domain name and http scheme of the URL.

Python whois library is imported and used to get the WHOIS record information from ICANN WHOIS. WHOIS databases are run by domain registrars and registries. ICANN's WHOIS service is a publicly searchable tool that searches the databases of registries and registrars to detail the domain owner contact information across all contracted gTLDs [33]. WHOIS registrar and registrant details are represented as bag-of-words [43]. WHOIS registration dates objects are parsed to get the date values from which are the age of domain (WHOIS creation date – today's date), life span of domain (WHOIS expiration date – creation date) and last update value (WHOIS last update date – today's date) are calculated as numeric values. For some websites, the WHOIS object obtained in the form of a python dictionary of key value pairs has some missing information. The dictionary structure is parsed to get the count of WHOIS information that's available for the website – number of registration information available feature.

DNS record details for DNS A, MX and NS records are obtained from dnspython library through DNS resolver query [34]. IP address information such as IP prefix, ASN (autonomous system number) information and SOA record details including the time-to-live (TTL) value of DNS records are obtained as a json object from ipwhois python module, by querying the domain name of the website. The json object is parsed to get the necessary DNS information mentioned above [35].

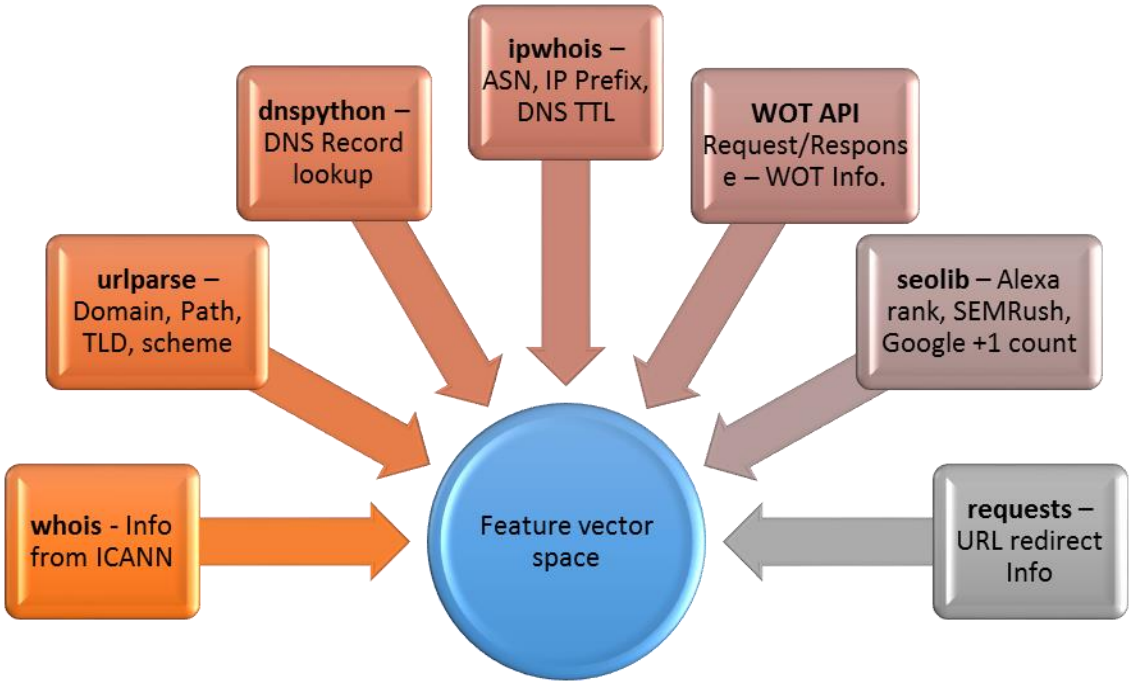
Site popularity features such as WOT (Web of Trust) features and page rank features are obtained from WOT API and seolib python modules respectively [36] [37]. WOT information is obtained using python's urllib2 request to WOT API using the API key provided in mywot.com. Web of Trust features are explained in detail in the next section.

The features extracted are in the form of binary values (presence of DNS A record, for example), numeric values (length of hostname and age of domain, for example) and text (URL, WHOIS registrar, for example) and are stored in csv format.

Figure 1 image depicts the different python libraries used to extract the features.

Figure 1

Python Modules for Feature Extraction



3. New Features used in our work

We use new features in our work that have not been used in the literature surveyed such as the Web of Trust (WOT) features (URL reputation and confidence, WOT category and confidence value), Google +1 count (user votes for the website), SEMRush rank (rank based on search engine keywords and traffic coming from search engine lookup of the website), presence of IP address in PTR record and presence of DNS SOA (State of authority) record.

Web of Trust Wiki is a website reputation and review service based on crowdsourcing approach that collects ratings and reviews from a global community of millions of users who rate and comment on websites based on their personal experiences. It helps people in making informed decisions about whether to trust a website or not [32].

4. Building a classification Model

The accuracy of the classification model in predicting a given website as phishing largely depends on the features trained to create the model and classification algorithm used. We choose a classifier that fits well with the data.

We use Decision table, Logistic Regression, J48 and Naïve Bayes classifiers in Weka [58] to train the model based on these features and present results for classifier that fits well with the data. The text features can be converted to bag-of-words using Weka's filtered classifier [59], where we select a StringToWordVector filter for that feature, which tokenizes the text and converts it into binary feature vector. We also use neural network fitting tool in MATLAB [60] to train a neural network with these features as inputs.

5. Datasets

The features 5999 phishing and 6158 benign as mentioned in Table 9 (Appendix 8.2) were extracted from URLs. We fetched a recent phishing URL list from Phishtank archive [40] and benign URL list from all categories of DMOZ directory [41]. From our lists, we randomly sampled a set of 12157 unique URLs (5999 phishing and 6158 benign) and used it for feature extraction.

CHAPTER 6

RESULTS

1. Overview

We train different classifiers and regression models as mentioned in chapter 5, with features mentioned in Table 9 (Appendix 8.2). Decision Table and Neural network provide better results for most feature vectors.

We start by analyzing individual predictive performance of all features considered and present the predictive power of features with better individual predictive power (above 70% (approx.) accuracy). We also present set of features selected by Weka along with their individual prediction performance. Finally, we present classification accuracy for the different combinations of robust features tested and show how they can be used in practice to detect specific phishing attacks.

2. Features with better predictive performance

We present the features which provide better prediction results over others on training individually with Decision Table classifier. Table 10 (Appendix 8.3) shows features with individual classification accuracy of 70% (approx.) or above in classifying a website as phishing or benign.

3. Features selected by Weka

Decision Table Classifier was trained with all features as mentioned in Table 9 (Appendix 8.2). We used Weka's feature selection [61] to select significant features from all these features.

The features selected by Weka's feature selection algorithm are number of redirects of URL, number of token in path portion of URL, number of dots in domain and path portions, number of special characters in path portion, length of path portion of URL, Number of tokens in the hostname, Google +1 count, WOT Reputation, Presence of DNS NS record, DNS ASN, Age of domain and HTTP scheme.

Out of these features selected by Weka, all of them, except number of redirects in URL, HTTP scheme and Google +1 count, provide more than 70% (approx.) accuracy individually in predicting a phishing website.

4. Results

The results obtained on training with features as mentioned in Table 9 (Appendix 8.2) is presented in Table 11. 75% of the dataset was used for training and 25% for testing.

Table 10

Results

Combination of features trained	**Decision Table Classifier**	**Neural Networks**
All features - URL + URL redirection + WHOIS + DNS/IP + Site popularity	96.18% ACC	98.16% ACC
URL redirection + WHOIS + DNS/IP + Site popularity	91.11% ACC	92.16% ACC
URL + URL redirection + WHOIS + DNS/IP	94.40% ACC	96.7% ACC
URL + URL redirection + DNS/IP	92.76% ACC	93.8% ACC

Note : ACC denotes prediction accuracy

5. Comparing our results with available research works.

Dataset Selection

Our datasets obtained from phishtank and DMOZ are a good representation of both phishing and benign URLs. DMOZ directory includes popular domains as well as regional non-popular benign

domains. Few research works use Alexa.com's list of top web sites for obtaining a benign URL set [18] [22]. As Alexa's top web sites have higher ranking compared to other benign URLs this might not be a good representation of benign sites.

Evaluation of results

As mentioned in Chapter 4, we analyze the results we get, for our different combinations of robust feature categories and state how it can be used in detecting specific phishing attacks. Our results indicate that robust features seem to have enough predictive power to be used in practice.

Few scams use directory generation to generate a different path for each user and randomized URLs that can evade detection by URL features. We know that URL features are not robust and may not predict well for phishing attacks that use shortened URL services, as mentioned in chapter 4. By using 32 features belonging to categories other than URL features category we get reasonable accuracy of 92.16% with neural networks. Most of the works we had surveyed do not present appropriate feature sets for different phishing use-cases.

Comparing our work with [1] [2] and [4] which mostly rely on URL features for classification of phishing sites, we extract more than 50% of our feature set from hosting domain of the website. We hence provide a more robust system that is tolerant to spams with crafted URLs that evade detection and that can detect attacks with shortened URLs effectively.

To detect phishing attacks executed by compromising existing popular websites, detection systems can make use of a combination of DNS/IP + URL features to create the classification model to detect phishing with 92% to 94% accuracy as mentioned in Table 11. Site popularity features and WHOIS features of the existing website are not used in this case to avoid feature set bias in defining the classification model which can result in incorrect predictions.

Change in Phishing Trends

Comparing our work with the approach by Ma, Justin, et al in 2009 [5], we use more categories of features to cater to the change in phishing trends between 2017 and 2009. Ma, Justin, et al, [29]

had shortlisted top 6 generic TLDs that hosted most phishing sites, whereas, in 2016, 220 new malicious TLD were found and more than half of phishing sites are hosted by .com TLDs. Comparing the dataset description mentioned in the paper [5] with our current phishtank dataset, we find that phishing URL lengths have shortened over the years and very few URLs have IP addresses in the hostname in the recent years.

Hence, our analysis on feature robustness and feature categories that would predict well for specific phishing attacks would prove useful to predict the present-day phishing attacks.

Light Weight operation

We provide a fast-paced light weight operation, as the features used in our work can easily be collected within a short period, and hence advantageous over works that collect features from web page and scripting content [9] [10] [12] [19] [21].

CHAPTER 7

Neural Networks in Phishing detection.

Neural Networks are used in a few research works to detect phishing websites. In the research works surveyed, the features used to train neural networks are from URL of website, webpage features, links in email and features from body of email. The features used and methods of extraction of these features are similar to machine learning algorithm based phishing detection techniques. When conducting the experiments, the number of input layers, number of hidden layers and number of hidden neurons are specified and changed per the output and MSE (Mean Square Error) values after each test.

The advantages of using neural networks is that we can adjust weights as per the changes in the environment and we can retrain the network with new set of inputs and targets in-order to get better fitting for new data. There are continuous changes in the phishing trends and different ways in which attackers try to fool users. Training neural networks with new data each time, would provide better prediction as compared to machine learning algorithms that train of a single batch of data. This is more efficient for cases where we predict output class for inputs not in the training set.

Zhang et al [27], predict email spam using features from email body and links and test their neural network with two activation functions (Hyperbolic tangent and Sigmoid). The accuracy is 95.5% which is similar to results obtained from SVM and Naïve Bayes classifiers. But using decision table, they get a better accuracy of 96.5%.

Rami Mohammad et al [28], use features from URL, WHOIS and Webpage to model a neural network and they present the experimental results for various number of hidden layers and neurons. The best MSE rate is 0.00223 with 2 neurons and 1 hidden layer.

Online learning algorithms such as AROW and CW algorithms are used in few research works [4] [7], which also train on data in real time and hence are better at predicting new phishing trends as

compared to machine learning algorithms. But online learning algorithms require prior knowledge such as an expected behavior or a probability distribution in-order to obtain a certain level of accuracy on the data, whereas neural network can also be trained in an unsupervised manner where it self organizes data to detect patterns. The latter is called deep learning. Feed forward neural networks and kernels under the supervised learning paradigms can be adopted for unsupervised learning. Hence deep learning can be used in real time to predict phishing with better accuracy and better adaptability. Moreover, while training the network, higher weights can be assigned to robust features, thus increasing the adaptability of the neural network.

Some phishers create fake websites by recreating or copying Logos and images in the original website. The minor changes or deviations in these logos can be caught by using neural networks and convolutional matrix [45].

In our work, we trained a neural network with a set of 25 features (binary and numeric features) extracted from the URL, WHOIS records, DNS IP records of the website and features based on site popularity. We used MATLAB Neural network fitting application and Levenberg-Marquardt algorithm to train the network with 6 hidden neurons and 1 hidden layer. Trained with 12157 samples of data (75% training; 5% validation and 20% testing), the MSE value is 1.84 X e^{-2} (98.16% accuracy) for testing dataset.

REFERENCES

1. Blum, Aaron, et al. "Lexical feature based phishing URL detection using online learning." Proceedings of the 3rd ACM workshop on Artificial intelligence and security. ACM, 2010.

2. Su, Ke-Wei, et al. "Suspicious URL filtering based on logistic regression with multi-view analysis." Information Security (Asia JCIS), 2013 Eighth Asia Joint Conference on. IEEE, 2013.

3. Chu, Weibo, et al. "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs." 2013 IEEE International Conference on Communications (ICC). IEEE, 2013.

4. Le, Anh, Athina Markopoulou, and Michalis Faloutsos. "Phishdef: Url names say it all." INFOCOM, 2011 Proceedings IEEE. IEEE, 2011.

5. Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

6. McGrath, D. Kevin, and Minaxi Gupta. "Behind Phishing: An Examination of Phisher Modi Operandi." LEET 8 (2008): 4.

7. Ma, Justin, et al. "Identifying suspicious URLs: an application of large-scale online learning." Proceedings of the 26th annual international conference on machine learning. ACM, 2009.

8. Afroz, Sadia, and Rachel Greenstadt. "Phishzoo: Detecting phishing websites by looking at them." Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE, 2011.

9. Aburrous, Maher, et al. "Predicting phishing websites using classification mining techniques with experimental case studies." Information Technology: New Generations (ITNG), 2010 Seventh International Conference on. IEEE, 2010.

10. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." IET Information Security 8.3 (2014): 153-160.

11. Singh, Priyanka, Yogendra PS Maravi, and Sanjeev Sharma. "Phishing websites detection through supervised learning networks." Computing and Communications Technologies (ICCCT), 2015 International Conference on. IEEE, 2015.

12. Eshete, Birhanu, Adolfo Villafiorita, and Komminist Weldemariam. "Binspect: Holistic analysis and detection of malicious web pages." International Conference on Security and Privacy in Communication Systems. Springer Berlin Heidelberg, 2012.

13. Zhang, Yue, Jason I. Hong, and Lorrie F. Cranor. "Cantina: a content-based approach to detecting phishing web sites." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

14. Miyamoto, Daisuke, Hiroaki Hazeyama, and Youki Kadobayashi. "An evaluation of machine learning-based methods for detection of phishing sites." International Conference on Neural Information Processing. Springer Berlin Heidelberg, 2008.

15. Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

16. Whittaker, Colin, Brian Ryner, and Marria Nazif. "Large-scale automatic classification of phishing pages." (2010).

17. Jiang, Hansi, Dongsong Zhang, and Zhijun Yan. "A Classification Model for Detection of Chinese Phishing E-Business Websites." PACIS. 2013.

18. Xiang, Guang, et al. "Cantina+: A feature-rich machine learning framework for detecting phishing web sites." ACM Transactions on Information and System Security (TISSEC) 14.2 (2011): 21.

19. Thomas, Kurt, et al. "Design and evaluation of a real-time url spam filtering service." 2011 IEEE Symposium on Security and Privacy. IEEE, 2011.

20. Aburrous, Maher, et al. "Intelligent phishing detection system for e-banking using fuzzy data mining." Expert systems with applications 37.12 (2010): 7913-7921.

21. Canali, Davide, et al. "Prophiler: a fast filter for the large-scale detection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.

22. Garera, Sujata, et al. "A framework for detection and measurement of phishing attacks." Proceedings of the 2007 ACM workshop on Recurring malcode. ACM, 2007.

23. Khonji, Mahmoud, Andrew Jones, and Youssef Iraqi. "A novel Phishing classification based on URL features." 2011 IEEE GCC Conference and Exhibition (GCC). 2011.

24. Xu, Li, et al. "Cross-layer detection of malicious websites." Proceedings of the third ACM conference on Data and application security and privacy. ACM, 2013.

25. Chou, Neil, et al. "Client-Side Defense Against Web-Based Identity Theft." NDSS. 2004.

26. Cook, Debra L., Vijay K. Gurbani, and Michael Daniluk. "Phishwish: a stateless phishing filter using minimal rules." International Conference on Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2008.

27. Zhang, Ningxia, and Yongqing Yuan. "Phishing detection using neural network." CS229 lecture notes (2012).

28. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Predicting phishing websites using neural network trained with back-propagation." Proceedings on the International Conference on Artificial Intelligence (ICAI). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013.

29. http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

30. Google Safe Browsing Transparency Report (2015). www.google.com/transparencyreport/safebrowsing/

31. http://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attack-statistics/

32. https://www.mywot.com/en/aboutus

33. https://whois.icann.org/en/using-whois

34. http://www.dnspython.org/

35. https://pypi.python.org/pypi/ipwhois

36. https://www.mywot.com/en/reputation-api

37. https://pypi.python.org/pypi/seolib

38. Sahoo, Doyen, Chenghao Liu, and Steven CH Hoi. "Malicious URL Detection using Machine Learning: A Survey." arXiv preprint arXiv:1701.07179 (2017).

39. https://docs.python.org/3/library/urllib.parse.html

40. PhishTank: Phishtank developer information. http://www.phishtank.com/developer_info.php

41. DMOZ: Open directory project. http://www.dmoz.org/

42. http://www.lavasoft.com/mylavasoft/company/blog/number-and-diversity-of-phishing-targets-continues-to-increase

43. https://en.wikipedia.org/wiki/Bag-of-words_model

44. https://www.seoquake.com/guide/index.html

45. http://llcao.net/cu-deeplearning15/projects/phrish.pdf

46. “Google safe browsing API,” http://code.google.com/apis/safebrowsing/, accessed Oct 2011.

47. 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.

48. http://www.dnsbl.info/

49. K-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, “Fighting phishing with discriminative keypoint features,” Internet Computing, IEEE, vol. 13, no. 3, pp. 56 –63, may-june 2009.

50. M. Hara, A. Yamada, and Y. Miyake, “Visual similarity-based phishing detection without victim site information,” in IEEE Symposium on Computational Intelligence in Cyber Security, 2009. CICS ’09, 2009, pp. 30 – 36.

51. https://www.netcraft.com/anti-phishing/certificate-authority-phishing-alerts/

52. https://whois.icann.org/en/basics-whois

53. http://www.pcnames.com/articles/what-are-dns-records

54. http://dyn.com/blog/dyn-tech-everything-you-ever-wanted-to-know-about-ttls/

55. http://www.alexa.com/siteinfo

56. http://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-web masters-11068

57. https://www.tripwire.com/state-of-security/security-awareness/6-common-phishing-attacks-and-how-to-protect-against-them/

58. http://www.cs.waikato.ac.nz/ml/weka/

59. http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/FilteredClassifier.html

60. https://www.mathworks.com/products/neural-network.html

61. http://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

APPENDIX A

FEATURE CATEGORIZATION

Table 2

URL Features

Finer Categories of URL features	**Features under each category**
Tokens in URL – Bag-of-words approach	**1**. Tokens in entire URL, **2**. Tokens in Domain portion, **3**. Tokens in path portion.
Presence of security sensitive, client server keywords / brand name	**1**. Presence of security sensitive words, **2.** Presence of client/ server keywords in the path portion, **3.** Presence of brand names in domain and path portions, **4.** Brand name distance for brand names in Domain / Path portion.
Presence of character codes	**1**. Presence of hexadecimal or Unicode characters in URL
Presence of @, port No, IP address	**1**. Presence of @ in URL, **2**. Presence of Port Number / IP address in domain portion.
Length of URL/Domain/Path	**1**. Length of entire URL, **2.** Length of TLD, **3**. Length of primary domain, **4.** Length of secondary domain, **5**. Length of path portion, **6**. Length of file name or directory portion
Number of dots, hyphens, underscores (special characters)	**1**. Number of Dots in URL, **2**. Number of Dots in domain portion, **3**. Number of dots in path portion, **4**. Number of hyphens in URL, **5**. Number of underscore in URL, **6**. Number of special characters in URL, **7**. Number of forward slash in URL
Domain/Path Token features	**1**. Number of tokens in domain / path portion, **2**. Length of longest token in Domain/Path portion, **3**. Average length of token in Domain/Path portion, **4**. Average Domain / Path tokens count
TLD Organization	**1**. Tokens in the TLD, **2.** Length of TLD
URL Redirects	**1**. Number of Redirects between initial and final landing page, **2**. Status of Redirect, **3**. Cause of Redirect (Http response or JavaScript or flash).

Table 4

Page and Content Features

Finer Categories of Page features	**Features under each category**
HTTP header information (server, cache control)	**1.** HTTP header tokens, fields and values, cache control
Presence / Absence of specific page events	**1**. Presence of on mouse over to hide link, **2.** Presence of On before unload event**,** **3.** Whether Right click is disabled
Number of abnormalities in Scripting Content	**1**. Percentage of scripting content on page, **2**. Number or string assignments in script, **3**. Number of DOM modifying functions, **4.** Number of event attachments (event handler calls) in script, **5**. Number of suspicious JavaScript functions, **6**. Number of long strings in script
HTML code, Text and Images	**1**. HTML code matching with phishing sites, **2.** Images on page, **3**. tokens from text on page, **4**. Check for copying website and spelling errors
Number of White space, unknown tags and Hidden elements, Small area elements	**1**. Percentage of white space in page, **2**. Percentage of unknown tags, **3**. Number of elements such as div, iframe or object with small area, **4**. Number of Hidden elements
Abnormal features in Forms / Presence of specific form fields	**1**. Presence of empty string or about:blank in form action or pointing to a different domain, **2**. Presence of Data field that take user Input (credit card , password), **3**. Presence of forms with <input> tag, sensitive keywords, **4**. images in form, **5.** non-http scheme in URL in action field.
Features of Iframes, Pop ups, User Prompts and plugins	**1**. Number of Iframes, **2**. Embedded Iframe URL features, **3**. Number of Popup windows, **4**. Pop up window URL features, **5**. Behavior that caused pop up window, **6**.

	Number of User prompts, **7**. Text of User prompts, **8**. Number of plugins on page, **9**. Plugin URL features, **10**. Application type of plugin (Java, flash)
Features of Links on page / Redirect pages	**1**. Links on Page - Same URL Heuristics are checked, **2**. Number of Links on Page <a> tags, **3**. Number of links that point to target website, Abnormal Anchor: Whether different from domain of Page URL, Anchors With values such as "file:///E:/", “#”
Request URL	**1**. Request URL (Whether domain of URL in address bar and source code (<src>) are different)
Shell code and suspicious Active X controls	**1**. Number of elements with shellcode between the start tag and end tag, **2**. Number of suspicious objects / Active X controls.

APPENDIX B

FEATURES EXTRACTED

Table 11

Features Extracted under different categories

Category	**Features**
URL Features	**1**. Number of dots in domain name / path, **2**. Length of Hostname, **3**. Presence of IP address in domain name, **4**. Number of Tokens in domain name / path, **5**. Length of Longest token of domain name / path, **6**. Average Length of tokens in domain name / path, **7**. Number of '@' in URL, **8**. Number of hyphens in domain name / path, **9**. Number of underscore in domain name / path, **10**. Number of forward slash in domain name / path, **11**. Number of underscore in domain name / path, **12**. Number of forward slash in domain name / path, **13**. Number of special characters in path, **14**. Number of Client/Server keywords in URL, **15**. Number of Security keywords in URL, **16**. Bag of words representation, **17**. Scheme (http or https), **18**. No of redirects between initial and final page (URL Redirection feature)
Site Popularity	**1**. Alexa traffic rank, **2**. Google +1 count, **3**. SEMRush rank, **4**. URL Reputation from Web of trust, **5**. URL Confidence from Web of Trust, **6**. Web of Trust Category Identifier, **7**. Web of Trust confidence value for that category
WHOIS features	**1**. Age of Domain (Creation Date – Today), **2**. Life span of Domain (Creation date – Expiration date), **3.** Last Update (Updation date – Today), **4**. Number of Registration Info Available, **5.** Presence of Creation, Updation and Expiration dates, **6**. WHOIS Registrar **7**. WHOIS Registrant
DNS/ IP Features	**1**. DNS ASN Number, **2**. DSN MX and NS server name and IP address, **3**. No of resolved IP addresses, **4.** Presence of A, MX, NS and PTR and SOA Record, **5**. Are IPs of A, MX and NS records belonging to same autonomous system?, **6.** DNS TTL, **7**. IP Address in PTR Record, **8**. IP Prefix, **9**. DNS ASN Country

APPENDIX C

RESULTS

Table 12

Individual feature performance results

Category	**Feature**	**Individual Feature performance**
URL Features	Tokens in URL (Bag-of-words)	80% ACC
	Number of Tokens in path portion	87.1% ACC
	Length of entire URL	82.4% ACC
	Length of URL path portion	88.15% ACC
	Number of special characters in path portion	86.4% ACC
	Dots in Hostname	71% ACC
	Dots in Path portion	72.39% ACC
	Number of Tokens in hostname	77.85% ACC
Site Popularity	Alexa traffic Rank	69.6% ACC
	WOT Confidence	74.1% ACC
	WOT category identifier	83.7% ACC
	WOT Reputation	85.96% ACC
WHOIS features	Age of Domain (Creation Date – Today)	70.85% ACC
	Life span of Domain (Creation date – Expiration date)	69.8% ACC

DNS/ IP Features	DNS ASN Number	69.9% ACC
	DNS TTL	75.1% ACC
	IP Prefix	77.7% ACC
	Presence of DNS NS record	74.76% ACC

Note: ACC denotes prediction accuracy.