**CHAPTER 8**

**FEATURE EXTRACTION**

One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publicly, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

**8.1 Address Bar based Features**

In a web browser, the address bar (also location bar or URL bar) is a GUI widget that shows the current URL. The user can type a URL into the bar to navigate to a chosen website in most modern browsers, non-URLs are automatically sent to a search engine. The address bar is the familiar text field at the top of a web browser's graphical user interface (GUI) that displays the name or the URL (uniform resource locator) of the current web page. Users request websites and pages by typing either the name or the URL into the address bar.

**8.1.1 Using the IP Address**

If an IP address is used as an alternative of the domain name in the URL, such as “http://125.98.3.123/fake.html”, users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link “http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html”.

If the domain part of website has an ip address. So, the website is under phishing website. Otherwise, the website is safe and called as legitimate website.

**8.1.2 Long URL to Hide the Suspicious Part**

Phishers can use long URL to hide the doubtful part in the address bar. For example: <http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_hom&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html>. To ensure accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset, we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

If the URL length is less than 54 then the website is safe and called as legitimate website or URL length is between 54 and 75 then the website is suspicious. Otherwise, the website is phishing website.

We have been able to update this feature rule by using a method based on frequency and thus improving upon its accuracy

**8.1.3 Using URL Shortening Services**

URL shortening is a method on the “World Wide Web” in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL. For example, the URL “http://portal.hud.ac.uk/” can be shortened to “bit.ly/19DXSk4”.

If the URL is very short or tiny then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.4 URL’s having “@” Symbol**

Some websites having special character which causes of phishing. Using “@” symbol in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.

If the website has @ symbol then website may be phishing website. Otherwise, the website is safe and called as legitimate website.

**8.1.5 Redirecting using “//”**

The existence of “//” within the URL path means that the user will be redirected to another website. An example of such URL’s is: “http://www.legitimate.com//http://www.phishing.com”. We examine the location where the “//” appears. We find that if the URL starts with “HTTP”, that means the “//” should appear in the sixth position. However, if the URL employs “HTTPS” then the “//” should appear in seventh position.

The Position of the Last Occurrence of "//" in the URL is greater than 7 then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.6 Adding Prefix or Suffix Separated by (-) to the Domain**

Another special symbol is hyphen (-) or dash symbol which sometimes treat as phishing website symbol. The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example: <http://www.Confirme-paypal.com/>.

If the domain name part includes – symbol then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.7 Sub Domain and Multi Sub Domains**

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD) and “hud” is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as “Suspicious” since it has one sub domain. However, if the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.

If number of dots present on the website is one then the website is legitimate or if number of dots present in website are two then the website is suspicious. Otherwise, the website is under the category of phishing.

**8.1.8 HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)**

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. The authors in (Mohammad, Thabtah and McCluskey 2012) (Mohammad, Thabtah and McCluskey 2013) suggest checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names include: “GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster and VeriSign”. Furthermore, by testing out our datasets, we find that the minimum age of a reputable certificate is two years.

If the website age of certificate is less than one year then the website may be the phishing website. If the website age of certificate is greater than or equal to one year then the website is come under legitimate.

**8.1.9 Domain Registration Length**

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

If domain of website expires in less than or equal to one year then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.10 Favicon**

A favicon is a graphic image (icon) associated with a particular Web page and/or Web site. Many recent user agents (such as graphical browsers and newsreaders) display them as a visual reminder of the Web site identity in the address bar or in tabs. If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.

If favicon loaded from external domain then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.11 Using Non-Standard Port**

This feature is useful in validating if a particular service (e.g., HTTP) is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened.

If Port # is of the preferred Status then the website is phishing website. Otherwise, the website is legitimate website.

**8.1.12 The Existence of “HTTPS” Token in the Domain Part of the URL**

The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example: http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

If using http token in domain part of the url then the website is phishing website. Otherwise, the website is legitimate website.

**8.2 Abnormal Based Features**

**8.2.1. Request URL:**

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

Rule: if (i) % of Request URL < 22% → Legitimate

(ii) %of Request URL ≥ 22% and 61% → Suspicious

(iii) Otherwise → feature = Phishing

**8.2.2. URL of Anchor**

An anchor is an element defined by the tag. This feature is treated exactly as “Request URL”. However, for this feature we examine:

1. If the tags and the website have different domain names. This is similar to request URL feature.
2. If the anchor does not link to any webpage.
3. <a href = “#”>
4. <a href = “#content”>
5. <a href = “#skip”>
6. <a href = “Javascript::void(0)”>

**8.2.3. Links in <Meta>, <Script> and <Link> tags**

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script: and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

**8.2.4. Server Form Handler (SFH)**

SFHs that contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

**8.2.5. Submitting Information to Email**

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user’s information to his personal email. To that end, a server-side script language might be used such as “mail()” function in PHP. One more client-side function that might be used for this purpose is the “mailto:” function.

**8.2.6. Abnormal URL**

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

**8.3. HTML and JavaScript Based Features**

**8.3.1. Website Forwarding**

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

**8.3.2. Status Bar Customization**

Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “onMouseOver” event, and check if it makes any changes on the status bar.

**8.3.3. Disabling Right Click**

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as “Using onMouseOver to hide the Link”. Nonetheless, for this feature, we will search for event “event.button==2” in the webpage source code and check if the right click is disabled.

**8.3.4. Using Pop-up Window**

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

**8.3.5. IFrame Redirection**

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the “frameBorder” attribute which causes the browser to render a visual delineation.

**8.4. Domain based Features**

**8.4.1. Age of Domain**

This feature can be extracted from WHOIS database (Whois 2005). Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

**8.4.2. DNS Record**

For phishing websites, either the claimed identity is not recognized by the WHOIS database (Whois 2005) or no records founded for the hostname (Pan and Ding 2006). If the DNS record is empty or not found then the website is classified as “Phishing”, otherwise it is classified as “Legitimate”.

**8.4.3. Website Traffic**

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”. Otherwise, it is classified as “Suspicious”.

**8.4.4. PageRank**

PageRank is a value ranging from “0” to “1”. PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In our datasets, we find that about 95% of phishing webpages have no PageRank. Moreover, we find that the remaining 5% of phishing webpages may reach a PageRank value up to “0.2”