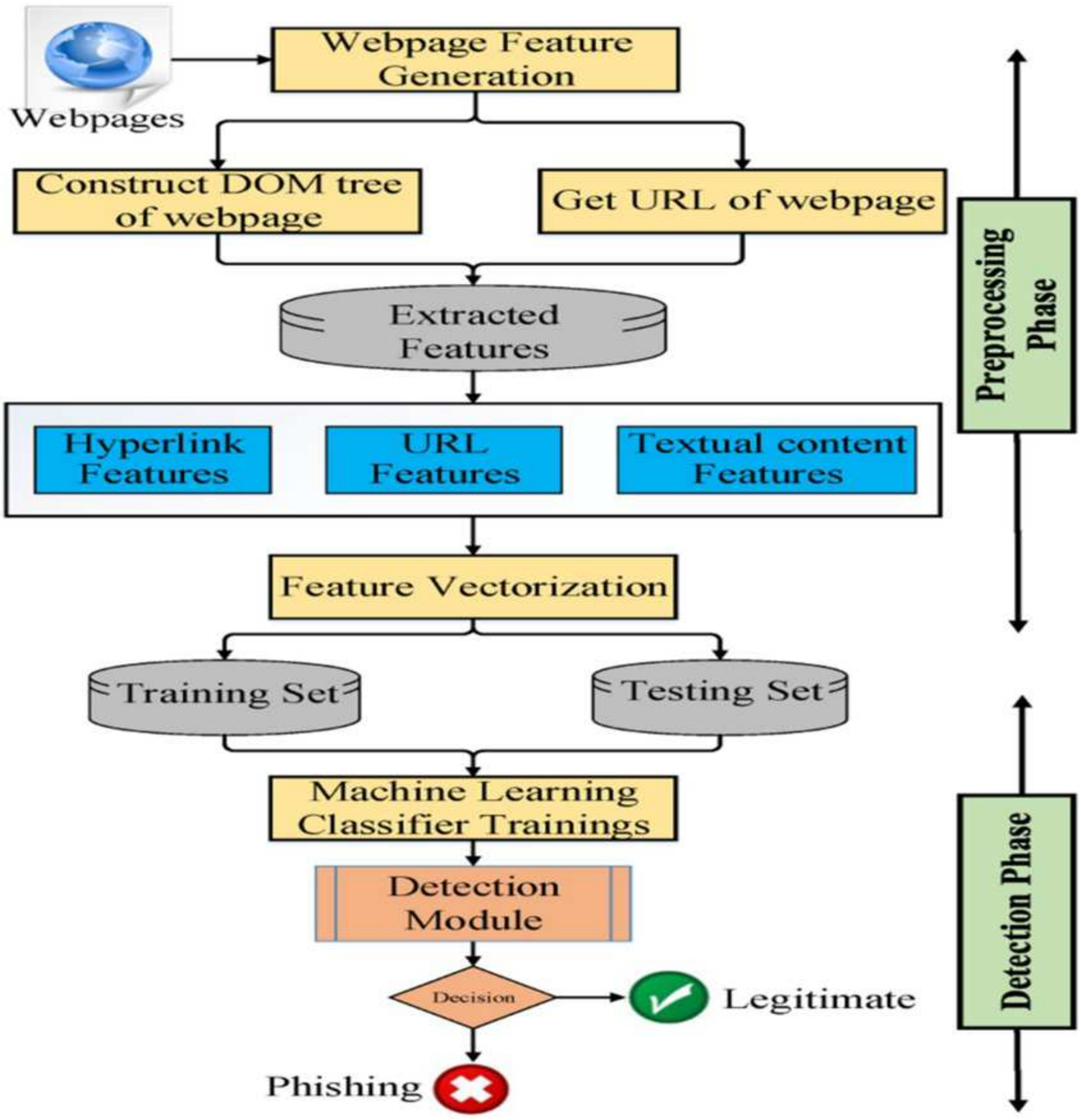
DYNAMIC PROGRAMMING

We propose a phishing detection approach, which extracts efficient features from the URL and HTML of the given webpage without relying on third-party services. Thus, it can be adaptable at the client side and specify better privacy. An efficient solution for phishing detection that extracts the features from website's URL and HTML source code proposed.



WORKING :

- We have collected unstructured data of URLs from Phishtank website, Kaggle website and Alexa website, etc.

- In pre-processing, feature generation is done where features are generated from unstructured data. These

  features are length of an URL, URL has HTTP, URL has suspicious character, prefix/suffix, number of dots, number of slashes, URL has phishing term, length of subdomain, URL contains IP address.

- After this, an organized dataset is made in which each detail incorporates the paired (0,1) which is then passed to the various classifiers.

- Next, we train the three unique classifiers and analyse their presentation based on exactness two classifiers utilized are Decision Tree and Random Forest algorithm.

- At that point, the classifier identifies the given URL dependent on the preparation information that is if the site is phishing it prompts the user that the website is phished and if genuine, it prompts the user that the website is legitimate.

- We look at the exactness of various classifiers and discovered Random Forest as the best classifiers which gives the most extreme precision.

However, if the URL entered by a user is found to be a phishing website, a small pop-up will appear on the screen to warn the user regarding this malicious website. There are times when a user needs to access some data on that website, so he/she can select a CONFIRM option to open the website, otherwise he/she will be sent back to the above webpage.

The components for detection and classification of phishing websites are as follows:

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript Based Features
4. Domain Based Features

1.Address Bar based Features

- Using the IP address

If IP address is used instead of domain name in the URL

e.g. 125.98.3.123 the user can almost be sure someone is trying to steal his personal information.

- Long URL to hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar.

- Using URL shortening services TinyURL

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.

- URLs having @ symbol

Using @ symbol in the URL leads the browser to ignore everything preceding the @ symbol and the real address often follows the @ symbol.

- Redirecting using //

The existence of // within the URL path means that the user will be redirected to another website.

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

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.

- 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).

- 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.

- 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.

- Favicon

A favicon is a graphic image (icon) associated with a specific webpage.

- Using Non-Standard Port

This feature is useful in validating if a particular service is up or down on a specific server.

- 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.

2.Abnormal Based Features

- Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain.

- URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as Request URL.

- 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.

- Server From Handler(SFH)

SFHs that contain an empty string or about:blank are considered doubtful because an action should be taken upon the submitted information.

- 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 users information to his personal email.

- Abnormal URL

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

3.HTML and JavaScript Based Features

- Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. Status Bar Customization

- 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.

- Using Pop-Up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window.

- IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown.

4.Domain Based Features

- Age of Domain

This feature can be extracted from WHOIS database. 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.

- DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records founded for the hostname. If the DNS record is empty or not found then the website is classified as Phishing, otherwise it is classified as Legitimate.

- Website Traffic

This feature measuresthe popularity of the website by determining the number of visitors and the number of pages they visit.

- Page Rank

PageRank is a value ranging from 0 to 1. PageRank aims to measure how important a webpage is on the Internet.

- Google Index

This feature examines whether a website is in Googles index or not. When a site is indexed by Google, it is displayed on search results.

- Number of Links Pointing to Page

The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain.