PHISHTICO: Phishing Website Detection

Project Report

5/20/2022

-SHUBHAM GARG

Abstract

In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users' sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing web pages look identical to the actual web pages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection.

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels.

Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

Table of Contents

Abbreviations……3

1. Introduction.....4

    1.1. Overview 6

    1.2. Objectives.....6

    1.3. Problem Statement.....6

    1.4. Project Goals.....6

2. Background Study.....7

    2.1. Classification of phishing attack techniques.....7

    2.2 Phishing detection approaches…….....8

3. Requirement Analysis.....9

    3.1. Tools & Technologies Used.....9

    3.2. Phishing Website Features 9

    3.3. Functional Requirement 11

    3.4. Non-Functional Requirement 11

4. Detailed Design.....12

    4.1. Flow Diagram 12

    4.2. Use Case Diagram 13

    4.3. S3quence Diagram 13

    4.4. Activity Diagram 14

    4.5. State Diagram 14

5. Conclusion & Future Scope……..15

    References……..……..16

Abbreviations

ML	Machine Learning
URL	Uniform Resource Locator
DNS	Domain Name Server
GPS	Global Positioning System
APWG	Anti-Phishing Working Group

1. Introduction

Phishing is a fraudulent technique that uses social and technological tricks to steal customer identification and financial credentials. Social media systems use spoofed e-mails from legitimate companies and agencies to enable users to use fake websites to divulge financial details like usernames and passwords. Hackers install malicious software on computers to steal credentials, often using systems to intercept username and passwords of consumers' online accounts. Phishers use multiple methods, including email, Uniform Resource Locators (URL), instant messages, forum postings, telephone calls, and text messages to steal user information. The structure of phishing content is similar to the original content and trick users to access the content in order to obtain their sensitive data. The primary objective of phishing is to gain certain personal information for financial gain or use of identity theft. Phishing attacks are causing severe economic damage around the world. Moreover, Most phishing attacks target financial/payment institutions and webmail, according to the Anti-Phishing Working Group (APWG) latest Phishing pattern studies .

In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data. This e-mail is rendered using a legitimate company's logos and slogans. The design and structure of HTML allow copying of images or an entire website. Also, it is one of the factors for the rapid growth of Internet as a communication medium, and enables the misuse of brands, trademarks and other company identifiers that customers rely on as authentication mechanisms. To trap users, Phisher sends "spooled" mails to as many people as possible. When these e-mails are opened, the customers tend to be diverted from the legitimate entity to a spoofed website.

There is a significant chance of exploitation of user information. For these reasons, phishing in modern society is highly urgent, challenging, and overly critical.There have been several recent studies against phishing based on the characteristics of a domain, such as website URLs, website content, incorporating both the website URLs and content, the source code of the website and the screenshot of the website. However, there is a lack of useful anti-phishing tools to detect malicious URL in an organization to protect its users. In the event of malicious code being implanted on the website, hackers may steal user information and install malware, which poses a

serious risk to cyber security and user privacy. Malicious URLs on the Internet can be easily identified by analyzing it through Machine Learning (ML) technique. The conventional URL detection approach is based on a blacklist (set of malicious URLs) obtained by user reports or manual opinions. On the one hand, the blacklist is used to verify an URL and on the other hand the URL in the blacklist is updated, frequently. However, the numbers of malicious URLs not on the blacklist are increasing significantly. For instance, cybercriminals can use a Domain Generation Algorithm (DGA) to circumvent the blacklist by creating new malicious URLs. Thus, an exhaustive blacklist of malicious URLs is almost impossible to identify the malicious URLs. Thus new malicious URLs cannot be identified with the existing approaches. Researchers suggested methods based on the learning of computer to identify malicious URLs to resolve the limitations of the system based on the blacklist. Malicious URL detection is considered a binary classification task with two-class predictions: malicious and benign. The training of the ML method consists of finding the best mapping between the d-dimensional vector space and the output variable. This strategy has a strong generalization capacity to find unknown malicious URLs compared to the blacklist approach.

1.1 Overview

Phishing is a form of identity theft that occurs when a malicious Website impersonates a legitimate one in order to acquire sensitive information such as passwords, account details, or credit card numbers. Phishing Websites Detection is a type of web-based application which will provide us a facility to find the original or fraud nature of the websites and informative as well.

1.2 Objectives

- The purpose of this System Requirement Specification document is describing the cyber security system which is called Phishing Websites Detection based on Machine Learning.
- Proposed a web-based system which provides a security through phishing detection.
- This system aims to provide a security a system which holds previous information and characteristics of websites.

1.3 Problem Statement

- The phishing attacker's trick users by employing different social engineering tactics such as threatening to suspend user accounts if they do not complete the account update process, provide other information to validate their accounts or some other reasons to get the users to visit their spoofed web pages.
- Phishing attacks affect millions of internet users and are a huge cost burden for businesses and victims of phishing. Phishing has become a significant threat to users and businesses alike.
- Over the past few years, much attention has been paid to the issue of security and privacy.

1.4 Project Goals

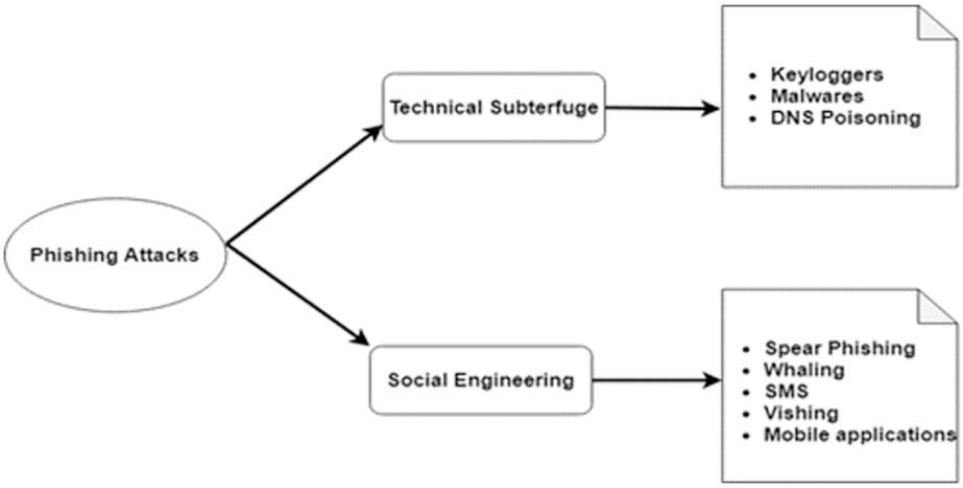
- To analyze datasets and identify the characteristics of WebPages.
- Data Mining and training of data through machine learning.
- Predict the nature of websites.

2. Background Study

Phishing attacks are categorized according to Phisher's mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs. Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites.

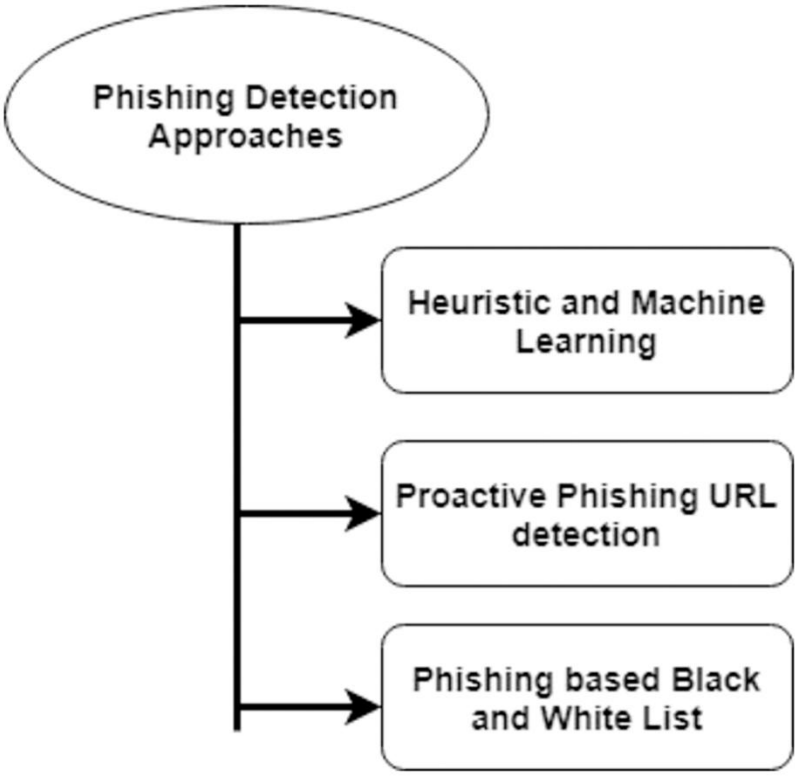
2.1 Classification of phishing attack techniques

Phishing websites are challenging to an organization and individual due to its similarities with the legitimate websites. Fig 1 presents the multiple forms of phishing attacks. Technical subterfuge refers to the attacks include Key logging, DNS poisoning, and Malwares. In these attacks, attacker intends to gain the access through a tool / technique. On the one hand, users believe the network and on the other hand, the network is compromised by the attackers. Social engineering attacks include Spear phishing, Whaling, SMS, Vishing, and mobile applications. In these attacks, attackers focus on the group of people or an organization and trick them to use the phishing URL [6, 7]. Apart from these attacks, many new attacks are emerging exponentially as the technology evolves constantly.



2.2 Phishing detection approaches

Phishing detection schemes which detect phishing on the server side are better than phishing prevention strategies and user training systems. These systems can be used either via a web browser on the client or through specific host-site software. Fig 2 presents the classification of Phishing detection approaches. Heuristic and ML based approach is based on supervised and unsupervised learning techniques. It requires features or labels for learning an environment to make a prediction. Proactive phishing URL detection is similar to ML approach. However, URLs are processed and support a system to predict a URL as a legitimate or malicious. Blacklist and Whitelist approaches are the traditional methods to identify the phishing sites. The exponential growth of web domains reduces the performance of the traditional method.



3. Requirement Analysis

3.1 Tools & Technologies Used

Python: We used the most common programming language that is used for building the machine learning model i.e., Python.

Jupyter Notebook: The IDE (Integrated Development Environments) that is used to build the unsupervised model is Jupyter Notebook.

Anaconda Command Prompt: It is used to launch Jupyter Notebook

3.2 Phishing Websites Features

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 publically, 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.

The important features that have proved to be sound and effective in predicting phishing websites are detailed below. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

S.No	Data Features
1	having_IP_Address
2	URL_Length
3	Shortining_Service
4	having_At_Symbol
5	double_slash_redirecting
6	Prefix_Suffix

7	having_Sub_Domain
8	SSLfinal_State
9	Domain_registeration_length
10	Favicon
11	port
12	HTTPS_token
13	Request_URL
14	URL_of_Anchor
15	Links_in_tags
16	SFH
17	Submitting_to_email
18	Abnormal_URL
19	Redirect
20	on_mouseover
21	RightClick
22	popUpWidnow
23	Iframe
24	age_of_domain
25	DNSRecord
26	web_traffic
27	Page_Rank
28	Google_Index
29	Links_pointing_to_page

3.3 Functional Requirement

1. A webpage is build for user interaction in which user can enter the URL for detection.

2. Alert is given to the user if the URL found to be a Phishing website

3. Model is trained automatically in the fixed interval to get updated for the new discovered phishing websites.

4. History is maintained in the database so, that the user can see for the previous results.

5. Dataset of the model is updated on the regular basis for new undiscovered websites.

3.4 Non Functional Requirement

1. Model is trained multiple times on the different algorithm to get maximum accuracy while prediction.

2. Cached result is maintained for quick prediction of previously discovered websites.

3. Model is based on ensemble learning.

4. Different dataset is used for model training.

5. Data Fitting is done in data preprocessing so that model will not become biased and not become overfit.

6. Data Scaling is done for take down data in range so that our model will not become biased.

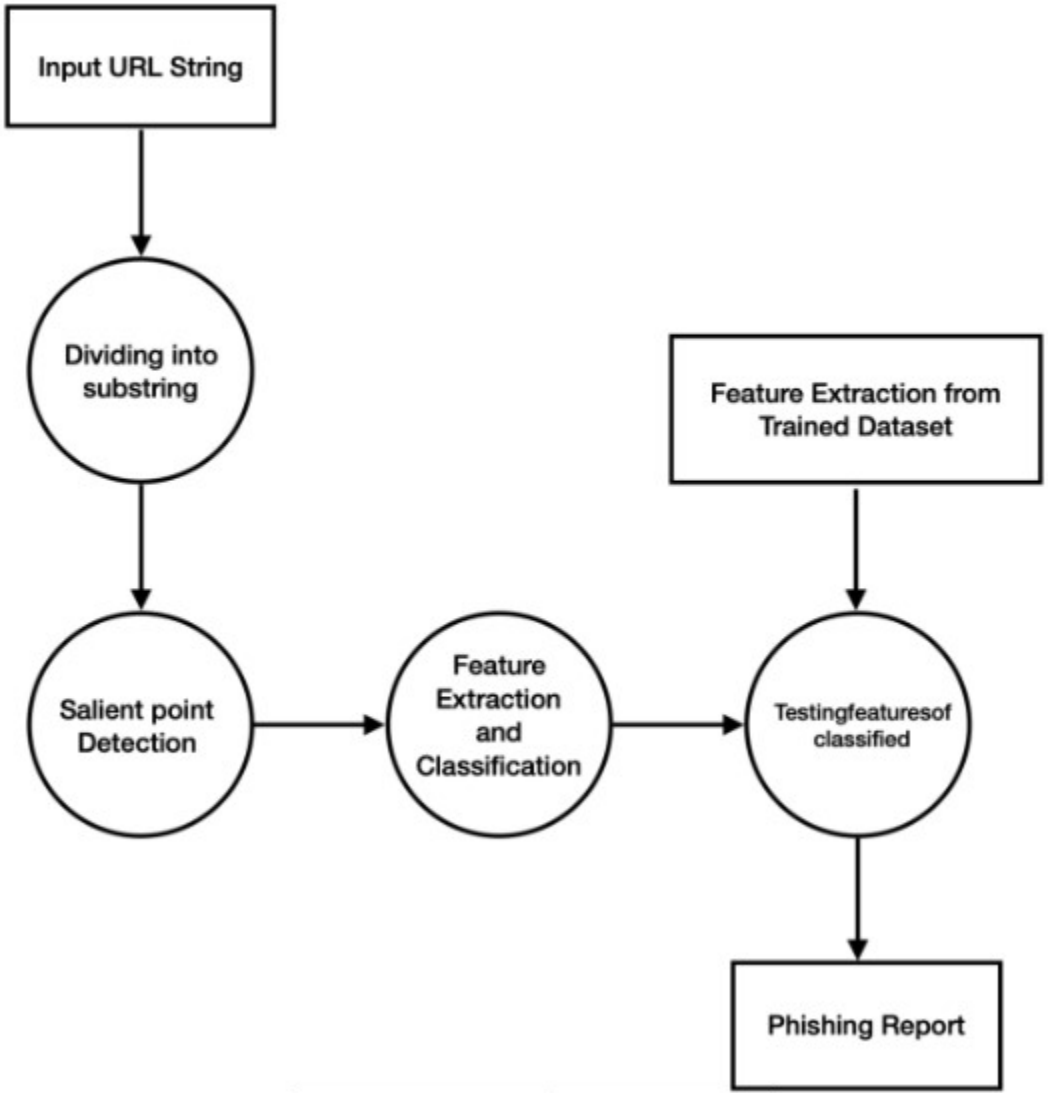
7. NAN value is removed from the dataset to increase the accuracy of the model.

8. Confusion matrix, precision, accuracy, F-1 score and recall is calculated for result analysis.

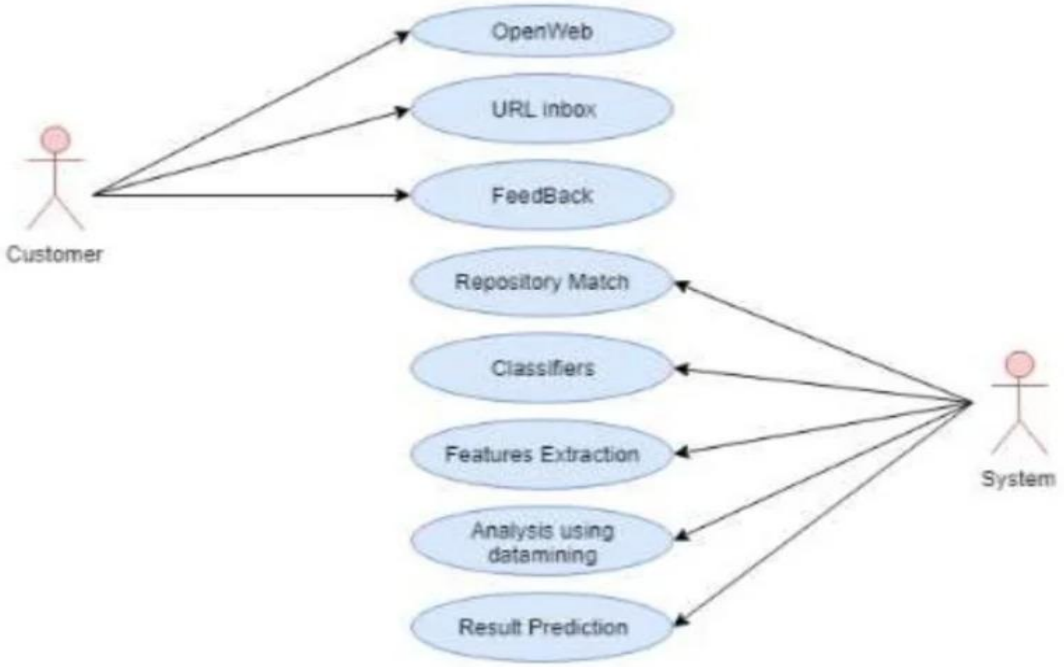
4. Detailed Design

The workflow of the application is given in a flowchart below:

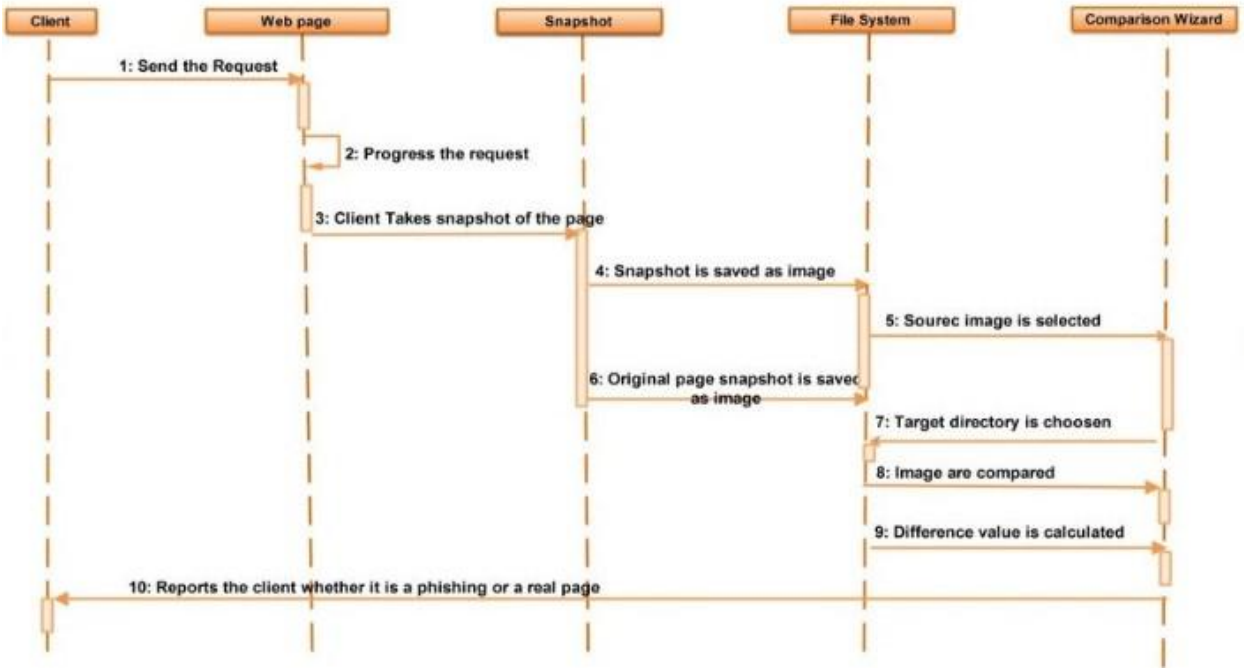
4.1 Flow Diagram



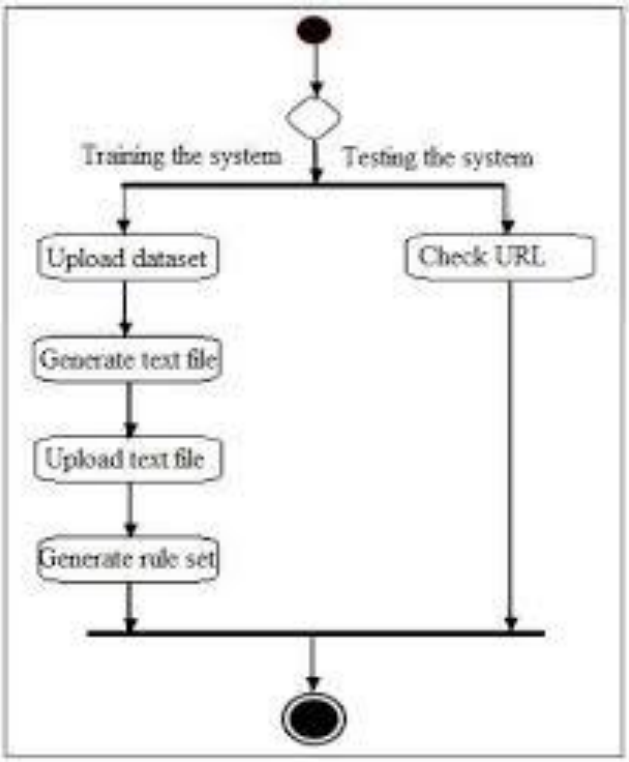
4.2 Use Case Diagram



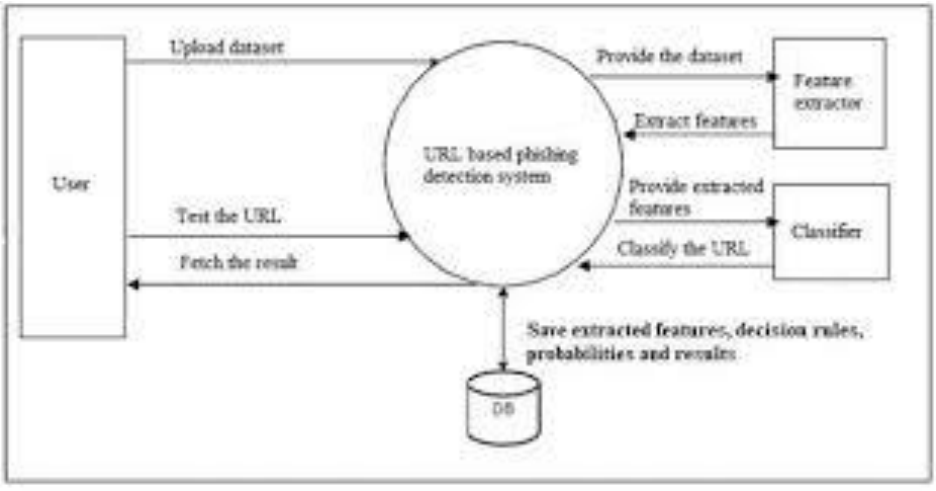
4.3 Sequence Diagram



4.4 Activity Diagram



4.5 State Diagram



Conclusion & Future Scope

In future, we wish to explore the robustness of machine learning algorithms for phishing detection in the presence of newer phishing attacks. We are also developing a real-time browser 62 add-on that will provide warnings when visiting suspicious sites. The authors believe that the phishing attacks are increasing day by day based on the literature review, though ample solutions are available. However, it is a bit challenges to educate\trained the users besides of detecting phishing attacks.

References

1. PhishLabs , 2018 PHISHING TRENDS & INTELLIGENCE REPORT. https://phishlabs.com

2. Research paper, "Real time detection of phishing websites", by Abdulghani Ali Ahmed and Nurul Amir. Published on IEEE.

3. Research paper Intelligent Phishing Website Detection using Random Forest Classifier"*, by Abdulhamit Subasi and T.J.Chaudhery published on IEEE.

4. RE LUCI Machine Learning Repository: Phishing Websites Data Set. Retrieved May 9. 2016, from https://archive.ics.uci.edu/ml/datasets/Phishing+Websites.