NALAIYA THIRAN

PROJECT REPORT

ON

WEB PHISHING DETECTION

Team ID: PNT2022TMID53652

Team Members: Sneha S

Surya V

Vaishnavi M

Yukesh Kumar S

TABLE OF CONTENTS

1.INTRODUCTION

Social engineering attack is a common security threat used to reveal private and confidential information by simply tricking the users without being detected. The main purpose of this attack is to gain sensitive information such as username, password and account numbers. According to, phishing or web spoofing technique is one example of social engineering attack. Phishing attack may appear in many types of communication forms such as messaging, SMS, VOIP and fraudster emails. Users commonly have many user accounts on various websites including social network, email and also accounts for banking. Therefore, the innocent web users are the most vulnerable targets towards this attack since the fact that most people are unaware of their valuable information, which helps to make this attack successful.

Typically phishing attack exploits the social engineering to lure the victim through sending a spoofed link by redirecting the victim to a fake web page. The spoofed link is placed on the popular web pages or sent via email to the victim. The fake webpage is created similar to the legitimate webpage. Thus, rather than directing the victim request to the real web server, it will be directed to the attacker server. The current solutions of antivirus, firewall and designated software do not fully prevent the web spoofing attack.

The implementation of Secure Socket Layer (SSL) and digital certificate (CA) also does not protect the web user against such attack. In web spoofing attack, the attacker diverts the request to fake web server. In fact, a certain type of SSL and CA can be forged while everything appears to be legitimate. According to, secure browsing connection does virtually nothing to protect the users especially from the attackers that have knowledge on how the “secure” connections actually work. This paper develops an anti-web spoofing solution based on inspecting the URLs of fake web pages. This solution developed series of steps to check characteristics of websites Uniform Resources Locators (URLs).

1.1 Project Overview

We have developed our project using a website as a platform for all the users. This is an interactive and responsive website that will be used to detect whether a website is legitimate or phishing. This website is made using different web designing languages which include HTML, CSS, Python. This project deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Decision Tree, Random Forest and support vector machine algorithms are used to detect phishing websites. Aim of the project is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

1.2 Purpose

The aim is to investigate the effectiveness of each algorithm to determine accuracy of detection and false alarms rate. So, this model will provide a clear guideline on how the research’s goals and objectives shall be achieved.

2. LITERATURE SURVEY

A literature review is a comprehensive summary of previous research on a topic. The literature review surveys scholarly articles, books, and other sources relevant to a particular area of research. The review should enumerate, describe, summarize, objectively evaluate and clarify this previous research.

2.1 Existing Problem

Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database. Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement the vulnerabilities of the stale lists.

2.2 References

[1] Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBM Internet Security Systems, 2007.

[2] https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref

[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[4] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed/January 2016

[5] http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/

[6] http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/

2.3 Problem Statement Definition

Phishing is when attackers send malicious emails designed to trick people into falling for a scam. Typically, the intent is to get users to reveal financial information, system credentials or other sensitive data. Phishing continually evolves to bypass security and human detection, so organisations must continually train staff to recognise the latest phishing strategies. It only takes one person to fall for phishing to incite a severe data breach. That's why it's one of the most critical threats to mitigate and the most difficult since it requires human defences.

Cyber criminals use phishing emails because it's easy, cheap and effective. Email addresses are easy to obtain, and emails are virtually free to send. With little effort and cost, attackers can quickly gain access to valuable data. Those who fall for phishing scams may end up with malware infections (including ransomware), identity theft and data loss. Detecting and preventing phishing offenses is a significant challenge for researchers due to the way phishers carry out the attack to bypass the existing anti-phishing techniques. Moreover, the phisher can even target some educated and experienced users by using new phishing scams. Thus, software-based phishing detection techniques are preferred for fighting against the phishing attack.

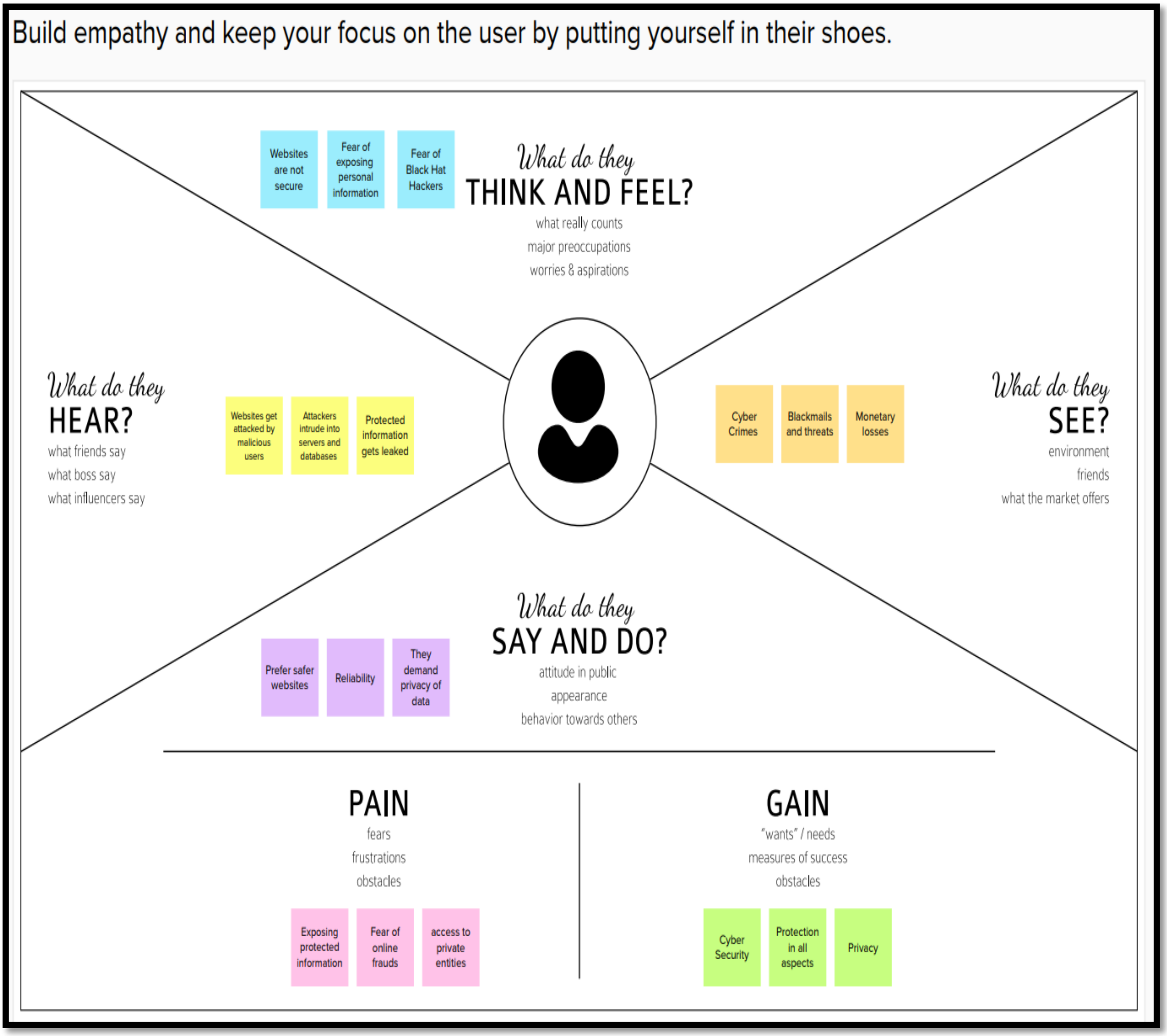
Mostly available methods for detecting phishing attacks are blacklists/whitelists, natural language processing, visual similarity, rules, machine learning techniques, etc. Techniques based on blacklists/whitelists fail to detect unlisted phishing sites as well as these methods fail when blacklisted URL is encountered with minor changes. In the machine learning based techniques, a classification model is trained using various heuristic features (i.e., URL, webpage content, website traffic, search engine, WHOIS record, and Page Rank) in order to improve detection efficiency.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

An empathy map is a collaborative tool teams can use to gain a deeper insight into their customers. It is a more in-depth version of the original empathy map, which helps identify and describe the user’s needs and pain points. And this is valuable information for improving the user experience.

Teams rely on user insights to map out what is important to their target audience, what influences them, and how they present themselves. This information is then used to create personas that help teams visualize users and empathize with them as individuals, rather than just as a vague marketing demographic or account number.



Build empathy and keep your focus on the user by putting yourself in their shoes.

What do they
THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations

Websites are not secure

Fear of exposing personal information

Fear of Black Hat Hackers

What do they
HEAR?
what friends say
what boss say
what influencers say

Websites get attacked by malicious users

Attackers intrude into servers and databases

Protected information gets leaked

What do they
SEE?
environment
friends
what the market offers

Cyber Crimes

Blackmails and threats

Monetary losses

What do they
SAY AND DO?
attitude in public
appearance
behavior towards others

Prefer safer websites

Reliability

They demand privacy of data

PAIN
fears
frustrations
obstacles

Exposing protected information

Fear of online frauds

access to private entities

GAIN
"wants" / needs
measures of success
obstacles

Cyber Security

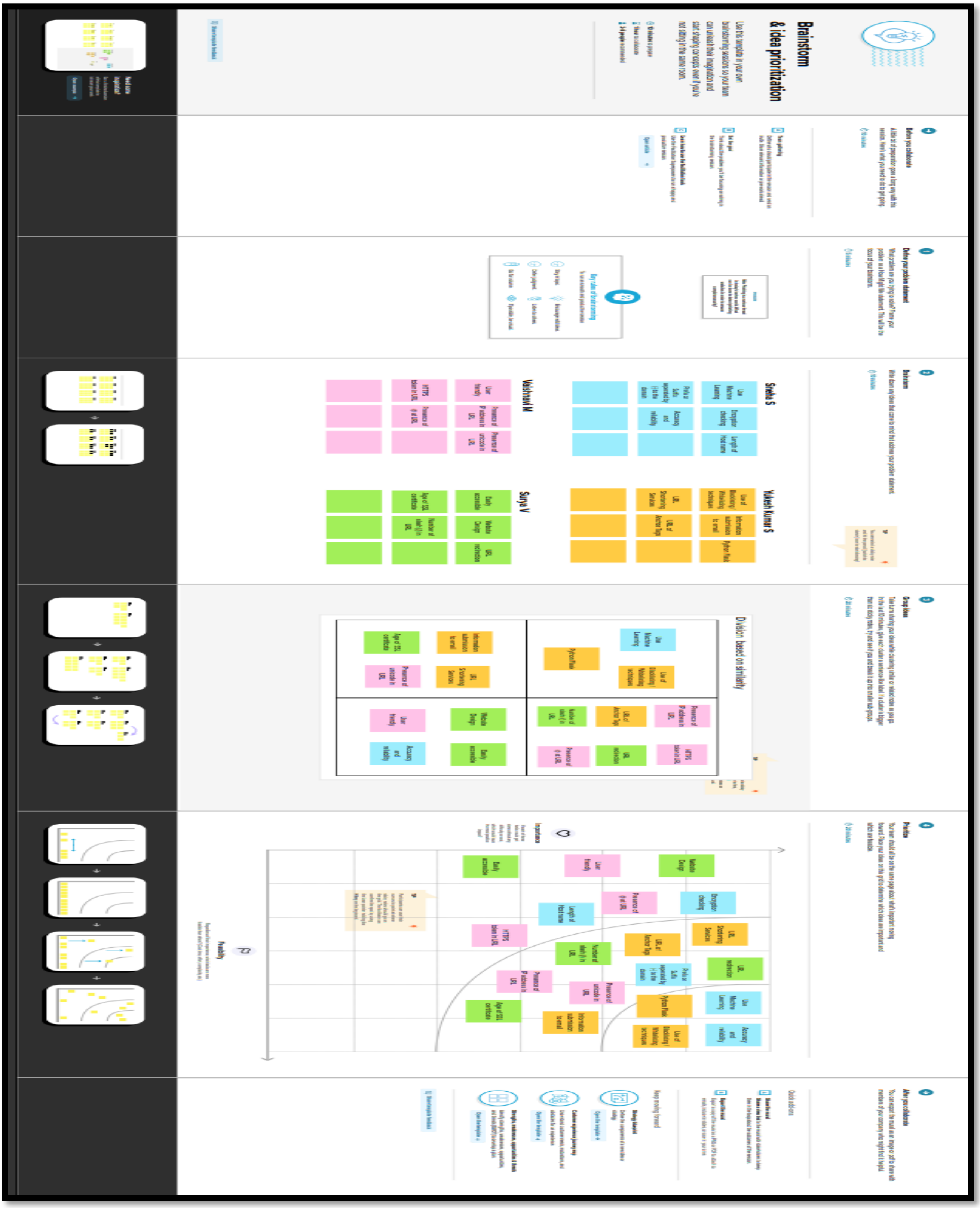
Protection in all aspects

Privacy

3.2 Ideation & Brainstorming

Ideation is often closely related to the practice of brainstorming, a specific technique that is utilized to generate new ideas. A principal difference between ideation and brainstorming is that ideation is commonly more thought of as being an individual pursuit, while brainstorming is almost always a group activity.

Brainstorming is usually conducted by getting a group of people together to come up with either general new ideas or ideas for solving a specific problem or dealing with a specific situation.



Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

10 minutes to prepare
1 hour to collaborate
3-8 people recommended

Share template feedback

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

10 minutes

A Team gathering
Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B Set the goal
Think about the problem you'll be focusing on solving in the brainstorming session.

C Learn how to use the facilitation tools
Use the Facilitation Superpowers to run a happy and productive session.

Open article

1 Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

5 minutes

PROBLEM
Web Phishing is a serious threat in today's techno world. What can be done to detect phishing websites in order to ensure complete security?

Key rules of brainstorming
To run an smooth and productive session

Stay in topic.
Encourage wild ideas.
Defer judgment.
Listen to others.
Go for volume.
If possible, be visual.

2 Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

TIP
You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

3 Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes

Division based on similarity

4 Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes

TIP
Participants can use their cursors to point at where sticky notes should go on the grid. The facilitator can confirm the spot by using the laser pointer holding the H key on the keyboard.

After you collaborate

You can export the mural as an image or pdf to share with members of your company who might find it helpful.

Quick add-ons

A Share the mural
Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.

B Export the mural
Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.

Keep moving forward

Strategy blueprint
Define the components of a new idea or strategy.
Open the template →

Customer experience journey map
Understand customer needs, motivations, and obstacles for an experience.
Open the template →

Strengths, weaknesses, opportunities & threats
Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.
Open the template →

Share template feedback

3.3 Proposed Solution

Novelty

Our model uses the power of machine learning to detect phishing sites. Python serves as a powerful tool to execute the application with Low false positives, High accuracy. Uses the latest techniques that give an efficient and great performance. It can easily differentiate the fake and safe URLs. If it's fake means, a warning message will be intimate to the users.

Feasibility of ideas

Using data visualization and machine learning algorithm, we safeguard the user's data by detecting malicious websites. This application is easy to be built we have a lot of existing software tools that aid us in creatin a web phishing detector. Faster, easier and seamless performance can be obtained.

Business model

Our model can be used by all people to secure their data from malicious websites. It's an open source tool.

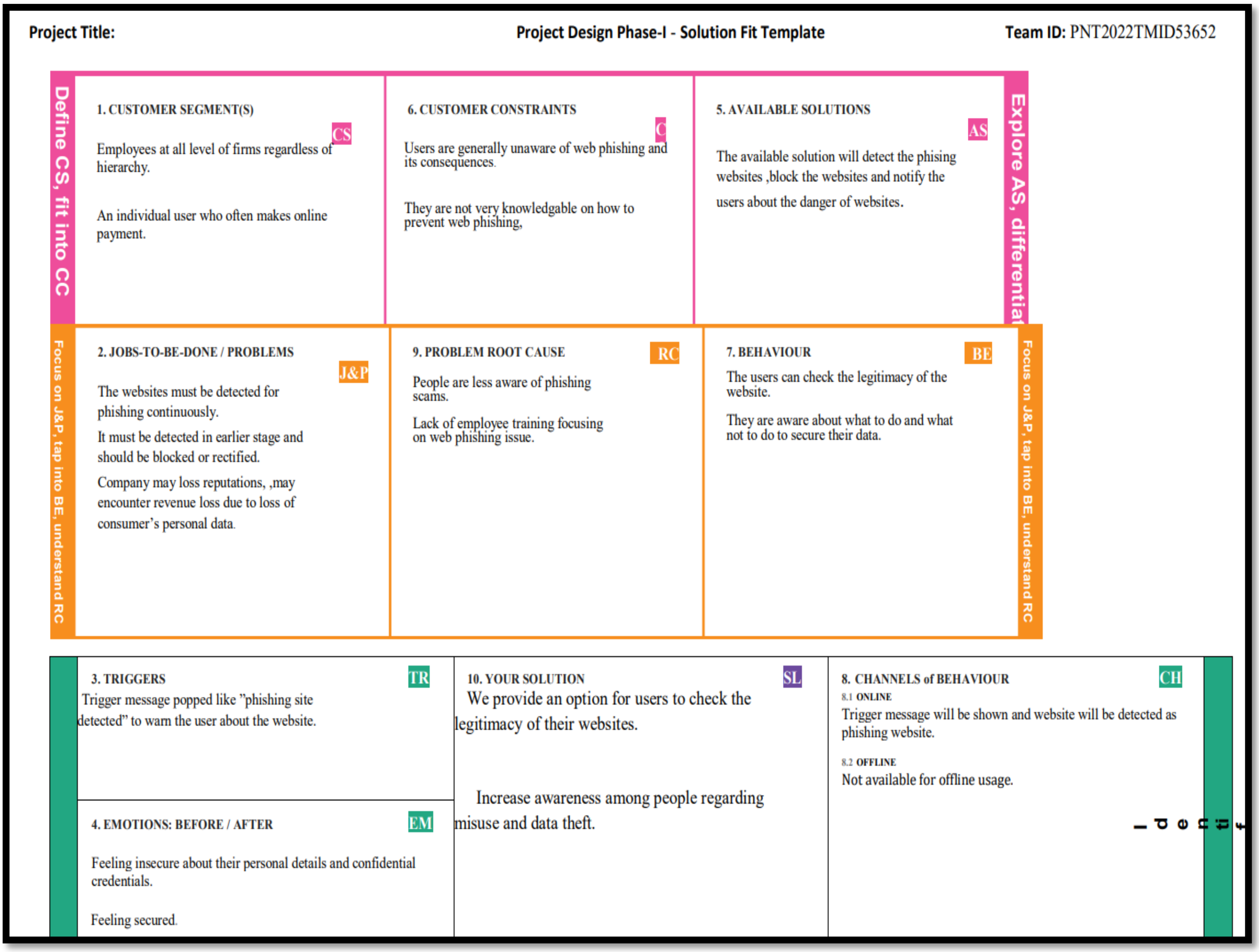
Social impact

According to recent research by Google, there was a 3505 increase in phishing websites from January to March 2020. Phishing has a list of negative effects on a business, including loss of money, loss of intellectual property, damage to reputation, and disruption of operational activities. As an impact of this model, people can able to find out fraudulent websites or fake ones. So that, they can avoid sharing sensitive data with unrecognized websites.

Scalability of solution

This project presents a proposal for scalable detection and isolation of phishing. It works on all types of websites and domains. It’s possible to make changes to software, which can accept new testing data and should also take part in training data and predict accordingly. In future prediction, modules can be improved and integrated.

3.4 Problem Solution fit



Project Title:

Project Design Phase-I - Solution Fit Template

Team ID: PNT2022TMID53652

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Employees at all level of firms regardless of hierarchy.

An individual user who often makes online payment.

6. CUSTOMER CONSTRAINTS C

Users are generally unaware of web phishing and its consequences.

They are not very knowledgable on how to prevent web phishing,

5. AVAILABLE SOLUTIONS AS

The available solution will detect the phising websites ,block the websites and notify the users about the danger of websites.

Explore AS, differentia

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

The websites must be detected for phishing continuously.

It must be detected in earlier stage and should be blocked or rectified.

Company may loss reputations, ,may encounter revenue loss due to loss of consumer’s personal data.

9. PROBLEM ROOT CAUSE RC

People are less aware of phishing scams.

Lack of employee training focusing on web phishing issue.

7. BEHAVIOUR BE

The users can check the legitimacy of the website.

They are aware about what to do and what not to do to secure their data.

Focus on J&P, tap into BE, understand RC

3. TRIGGERS TR

Trigger message popped like ”phishing site detected” to warn the user about the website.

4. EMOTIONS: BEFORE / AFTER EM

Feeling insecure about their personal details and confidential credentials.

Feeling secured.

10. YOUR SOLUTION SL

We provide an option for users to check the legitimacy of their websites.

Increase awareness among people regarding misuse and data theft.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE

Trigger message will be shown and website will be detected as phishing website.

8.2 OFFLINE

Not available for offline usage.

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration via form, email or websites.
FR-2	User Confirmation	Confirmation of registration through verification email or OTP.
FR-3	User Access	Creating their own username and passwords for access.
FR-4	User Validation	Validating email and passwords.
FR-5	User Security	Two factor Authentication.
FR-6	User Input	User gives an URL as an input to detect for web phishing and it will be detected using machine learning algorithms and will be reported to the users.

4.2 Non-Functional requirements

Following are the non-functional requirements of the proposed solution.

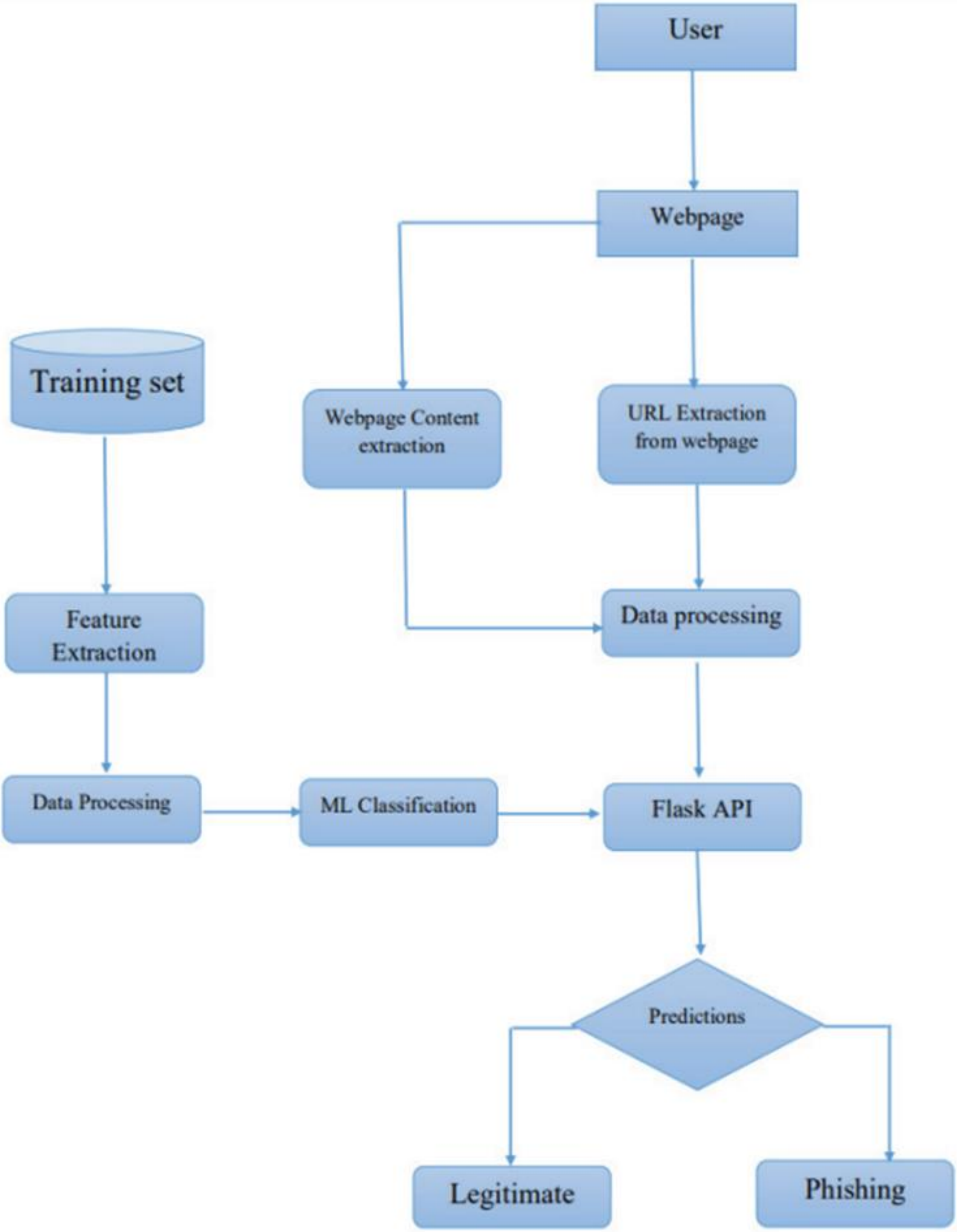
NFR No.	Non-Functional Requirement	Description
NFR-1	Usability	User Friendly. It is necessary the database backup should be provided in the case of retrieving data during virus attack.
NFR-2	Security	Appropriate User Authentication should be provided. Security questions should be setup while registering. Intimation email should be sent to the users at each time of accessing.

NFR-3	Reliability	The username and passwords should be highly reliable and secure. It should not be shared with anyone. Only those users and the admin of the web phishing detection system should access it.
NFR-4	Performance	The system should be fast and accurate.
NFR-5	Availability	The system should be available and compatible for all devices and OS.
NFR-6	Scalability	Admins should be able to make changes to the system whenever required and users can be able to use the system after making changes also.

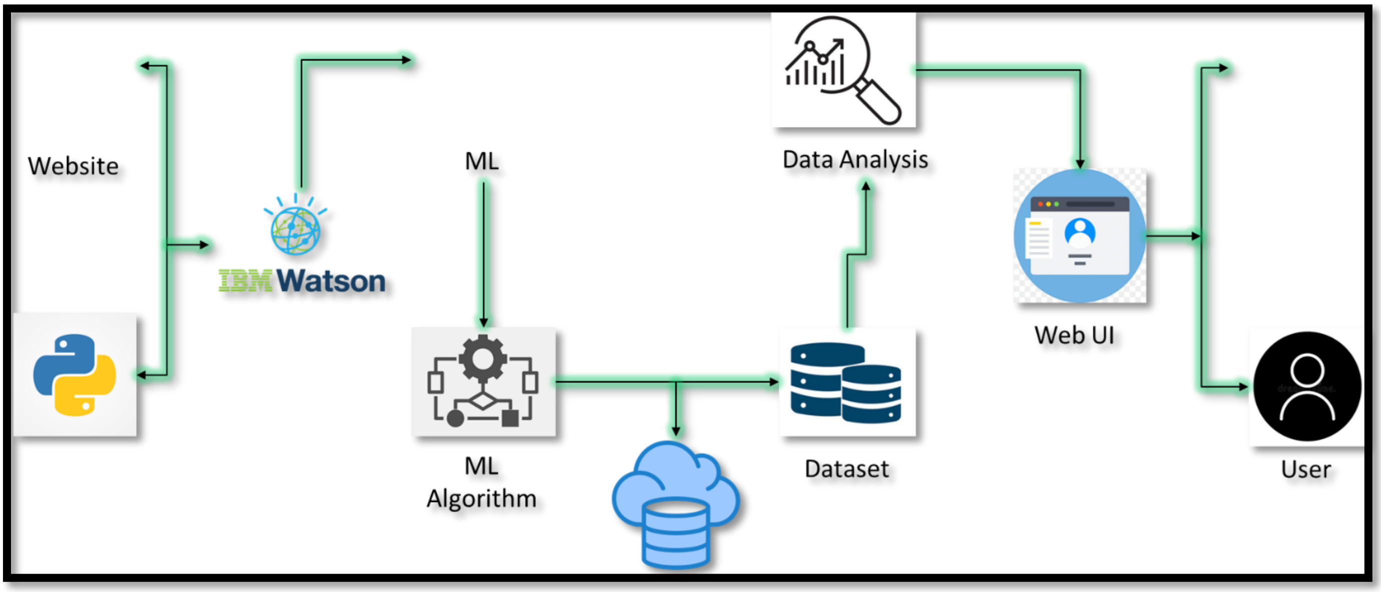
5. PROJECT DESIGN

5.1 Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.


User
Webpage
Training set
Webpage Content extraction
URL Extraction from webpage
Feature Extraction
Data processing
Data Processing
ML Classification
Flask API
Predictions
Legitimate
Phishing

5.2 Solution & Technical Architecture



S No.	Component	Description
1	Website	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.
2	Python	Code for random data
3	Machine Learning model	To identify anomalies
4	ML Algorithm	The logic behind the identification of errors
5	Infrastructure (Cloud/Server)	Storage
6	Data Analytics	Work on large data sets
7	User Interface	User feasibility
8	Web Notification	Report to user
9	User	Utilize the resources

5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can receive confirmation email & click confirm	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can access my dashboard with email login	High	Sprint-1
	Dashboard	USN-6	As a user, I can know the features of my dashboard	I can use a mobile application to access a dashboard	Medium	Sprint-1
Customer (Web user)	User Input	USN-1	As a user, I can input the particular URL in the required field and waiting for validation	I can go access the website without any problem	High	Sprint-1
Customer Care Executive	Knowledge	USN-1	As a customer care executive, I need to satisfy the customer’s query	I can access my dashboard and support the modules	High	Sprint-1
Administrator	Feature Extraction	USN-1	As an administrator, I have to extract the featured data using heuristic and visual similarity approach	I can have comparison between websites for security	High	Sprint-1
	Prediction	USN-2	Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression, KNN Model	I can have correct prediction on the particular algorithms	High	Sprint-2
	Classifier	USN-3	Here I will send all the model output to classifier in order to produce final result	I will find the correct classifier for producing the result.	Medium	Sprint-2

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

Velocity

Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let’s calculate the team’s average velocity (AV) per iteration unit (story points per day)

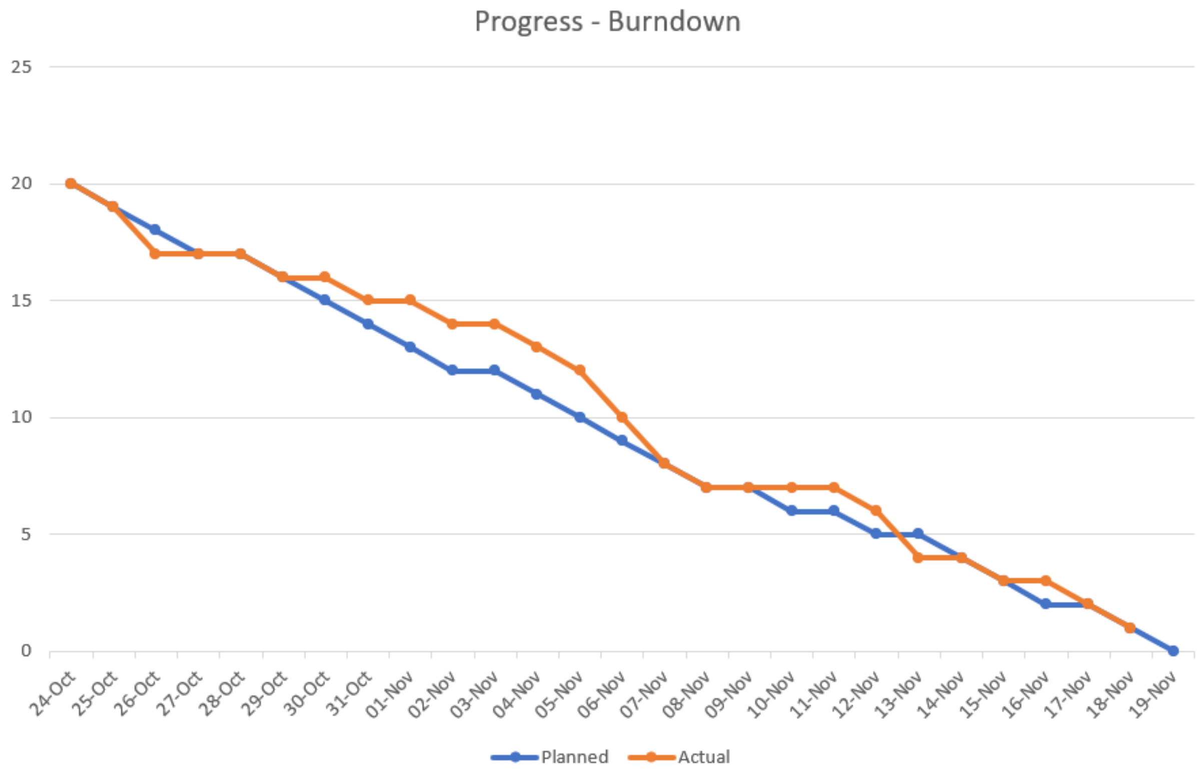
AV = (Sprint Duration / Velocity) = 20 / 10 = 2

We have a 6-day sprint duration, and the velocity of the team is 20 (points per sprint). So our team’s average velocity (AV) per iteration unit (story points per day)

AV = (Sprint Duration / Velocity) = 20 / 6 = 3.3

Burndown Chart

A burn down chart is a graphical representation of work left to do versus time. It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.



6.2 Sprint Delivery Schedule

Sprint planning is an event in scrum that kicks off the sprint. The purpose of sprint planning is to define what can be delivered in the sprint and how that work will be achieved. Sprint planning is done in collaboration with the whole scrum team.

Sprint	Functional Requirement (Epic)	User Story Number	Use Story / Task	Story Points	Priority
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High
Sprint-2		USN-3	As a user, I can register for the application through Facebook	2	Low
Sprint-1		USN-4	As a user, I can register for the application through Gmail	2	Medium
Sprint-1	Login	USN-5	As a user, I can log into the application by entering email & password	1	High
Sprint-1	Dashboard	USN-6	As a user, I can know the features of my dashboard	1	Medium
Sprint-1	User Input	USN-7	As a user, I can input the particular URL in the required field and waiting for validation	2	High
Sprint-2	Feature Extraction	USN-8	As an administrator, I have to extract the featured data using heuristic and visual similarity approach	2	High
Sprint-2	Prediction	USN-9	Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression, KNN Model	2	High
Sprint-3	Classifier	USN-10	Here I will send all the model output to classifier in order to produce final result	2	High
Sprint-4	Management	USN-11	As an admin, we can response to the user message for improvement of the website	2	Medium

6.3 Reports from JIRA

When JIRA sends either standard notifications or user invitations to a mail server, they are listed as phishing attempts rather than legitimate websites.

- Check the base URL of JIRA to see if it is set as a direct IP address with port number.

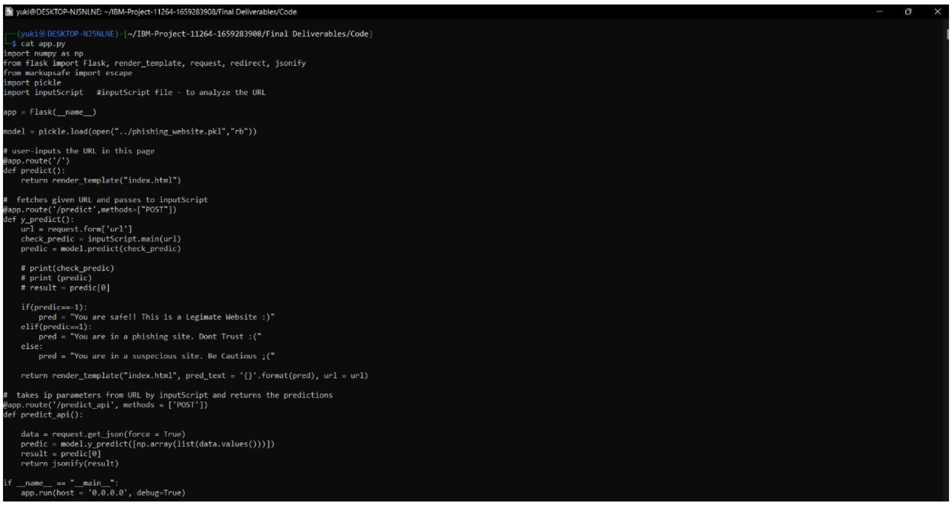
Example: **http://10.10.10.10:8080**

Some email servers (such as Microsoft Outlook) will consider messages from non-DNS URLs as phishing attempts. You can correct this behaviour by setting JIRA's base URL to a URL address such as **http://my-jira.com.**

- Sometimes when certain mail servers receive multiple emails from the same sender security measures are triggered that will then list those emails as phishing messages. For this, it is best to check with the local mail server administrator for further assistance and confirmation.

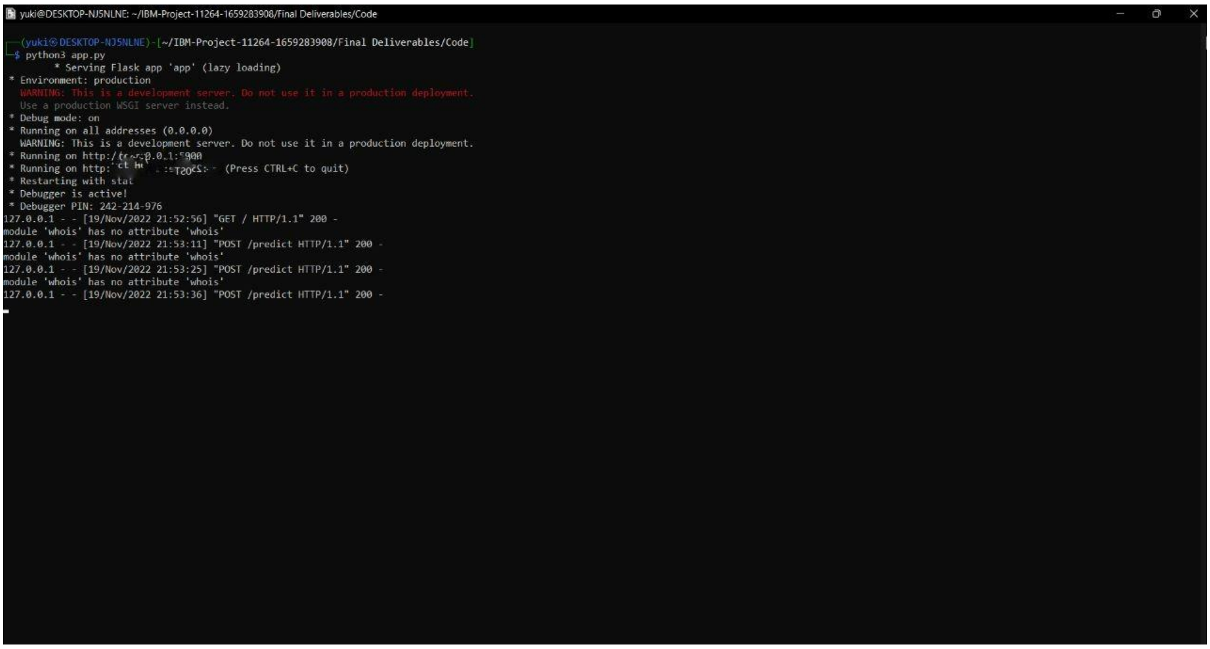
7.CODING&SOLUTIONING

CODING:

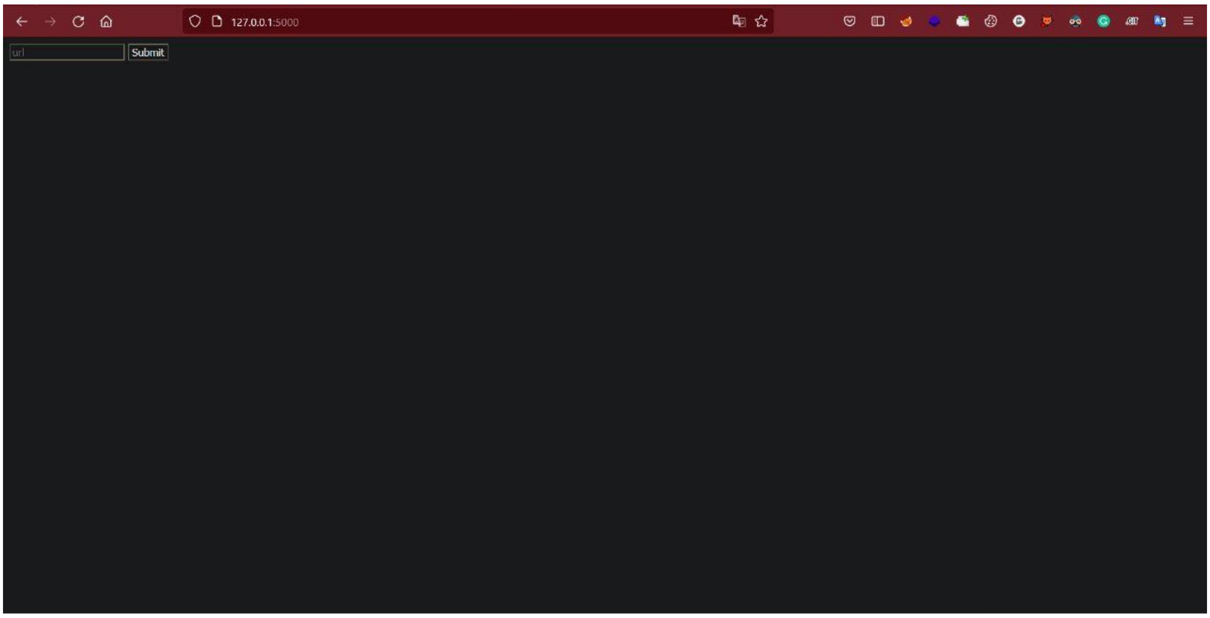


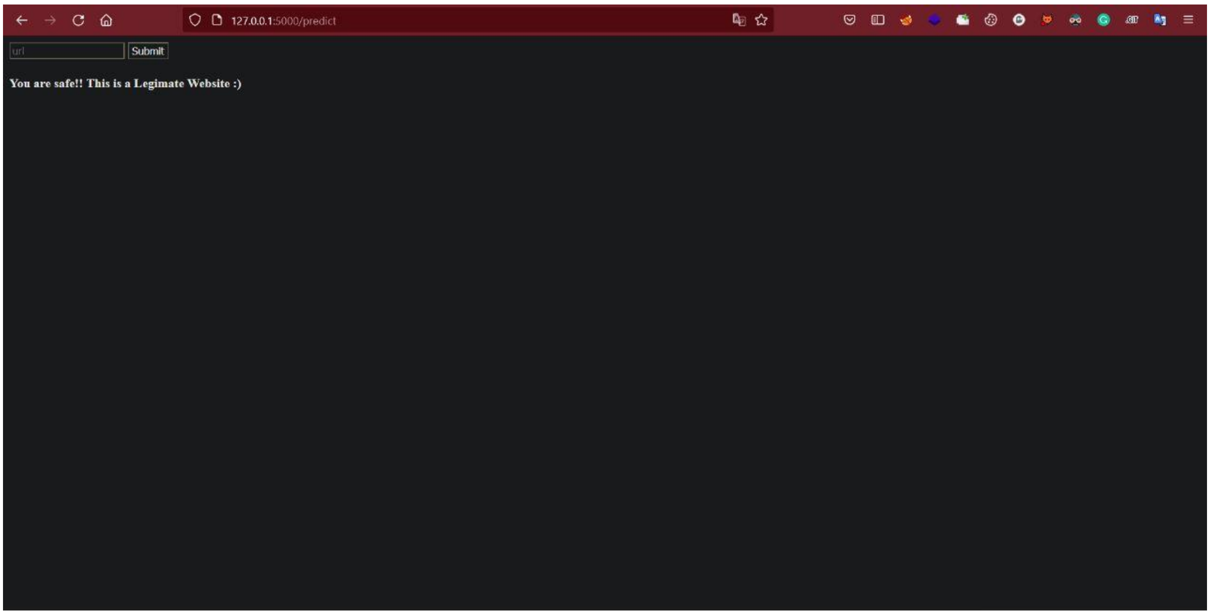
yuki@DESKTOP-NJ5NLNE: ~/IBM-Project-11264-1659283908/Final Deliverables/Code

```
┌──(yuki㉿DESKTOP-NJ5NLNE)-[~/IBM-Project-11264-1659283908/Final Deliverables/Code]
└─$ cat app.py
import numpy as np
from flask import Flask, render_template, request, redirect, jsonify
from markupsafe import escape
import pickle
import inputScript   #inputScript file - to analyze the URL

app = Flask(__name__)

model = pickle.load(open("../phishing_website.pkl","rb"))

# user-inputs the URL in this page
@app.route('/')
def predict():
    return render_template("index.html")

#  fetches given URL and passes to inputScript
@app.route('/predict',methods=["POST"])
def y_predict():
    url = request.form['url']
    check_predic = inputScript.main(url)
    predic = model.predict(check_predic)

    # print(check_predic)
    # print (predic)
    # result = predic[0]

    if(predic==-1):
        pred = "You are safe!! This is a Legimate Website :)"
    elif(predic==1):
        pred = "You are in a phishing site. Dont Trust :("
    else:
        pred = "You are in a suspecious site. Be Cautious ;("

    return render_template("index.html", pred_text = '{}'.format(pred), url = url)

#  takes ip parameters from URL by inputScript and returns the predictions
@app.route('/predict_api', methods = ['POST'])
def predict_api():

    data = request.get_json(force = True)
    predic = model.y_predict([np.array(list(data.values()))])
    result = predic[0]
    return jsonify(result)

if __name__ == "__main__":
    app.run(host = '0.0.0.0', debug=True)
```



```
yuki@DESKTOP-NJ5NLNE: ~/IBM-Project-11264-1659283908/Final Deliverables/Code

┌──(yuki㉿DESKTOP-NJ5NLNE)-[~/IBM-Project-11264-1659283908/Final Deliverables/Code]
└─$ python3 app.py
       * Serving Flask app 'app' (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Running on all addresses (0.0.0.0)
   WARNING: This is a development server. Do not use it in a production deployment.
 * Running on http:/[illegible]0.0.1:5000
 * Running on http:[illegible] (Press CTRL+C to quit)
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 242-214-976
127.0.0.1 - - [19/Nov/2022 21:52:56] "GET / HTTP/1.1" 200 -
module 'whois' has no attribute 'whois'
127.0.0.1 - - [19/Nov/2022 21:53:11] "POST /predict HTTP/1.1" 200 -
module 'whois' has no attribute 'whois'
127.0.0.1 - - [19/Nov/2022 21:53:25] "POST /predict HTTP/1.1" 200 -
module 'whois' has no attribute 'whois'
127.0.0.1 - - [19/Nov/2022 21:53:36] "POST /predict HTTP/1.1" 200 -
```





You are safe!! This is a Legimate Website :)

FEATURES:

- Website is created in which user can post their domain and can check for phishing URL'S.
- User Friendly. Everyone can check for phishing websites.
- Simple and efficient Machine Learning Algorithms are used to design the system.

8.TESTING

Testcases:

TESTSTCASE 1: If phishing is detected it shows the warning message as "PHISHING WEBSITE".

TESTCASE 2: If not it shows as "PHISHING FREE WEBSITE".

9.RESULTS

PERFORMANCE METRICS:

Highly Reliable

User Friendly

Fast and accurate response

10.ADVANTAGES & DISADVANTAGES

ADVANTAGES:

Eliminate the cyberthreat risk level.

Protect valuable corporate and personal data.

DISADVANTAGES:

Need many training data.

Need expert to review and monitor the network.

11.CONCLUSION

Phishing is a growing crime and one that we must be aware of. Phishing attacks are major threat to e-commerce and banking applications. It affects both consumers and organizations. Our proposed a system find those phishing websites and reports it to the user such that they can be aware of phishing attacks.

12.FUTURE SCOPE

Further enhancements which has to be made are generating a report about phishing websites and send it to the user, providing subscriptions to the user by which they can scan their websites limitlessly.