WEB PHISHING DETECTION

NALAIYA THIRAN - PROJECT REPORT

PROJECT ID: PNT2022TMID00904

Submitted by

KARTHICK ARAVIND B [211419104318]

AMARNATH R [211419104010]

BALAMURUGAN S [211419104036]

PENIAL BRANHAM T [211419104321]

In partial fulfillment for the award of the degree

Of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE, CHENNAI-600123.

(AN AUTONOMOUS INSTITUTION , AFFILIATED TO ANNA UNIVERSITY)

NOVEMBER 2022

PANIMALAR ENGINEERING COLLEGE, CHENNAI-600123.

(AN AUTONOMOUS INSTITUTION , AFFILIATED TO ANNA UNIVERSITY)

BONAFIDE CERTIFICATE

Certified that this project report

"**WEB PHISHING DETECTION– PNT2022TMID00904"**

is the bonafide work of

KARTHICK ARAVIND B	[211419104318]
AMARNATH R	[211419104010]
BALAMURUGAN S	[211419104036]
PENIAL BRANHAM T	[211419104321]

who carried out the NALAIYA THIRAN project work under the supervision.

SANDESH P
INDUSTRY MENTOR
IBM

PUGHAZENDI N
FACULTY MENTOR
Department of CSE
Panimalar Engineering College

INDEX

1.INTRODUCTION

1.1 Project Overview

Phishing is one of the most severe cyber-attacks where researchers are interested to find a solution. In phishing, attackers lure end-users and steal their personal information. To minimize the damage caused by phishing must be detected as early as possible. There are various phishing attacks like spear phishing, whaling, vishing, smishing, pharming and so on. There are various phishing detection techniques based on whitelist, black-list, content-based, URL-based, visualsimilarity and machine-learning. In this paper, we discuss various kinds of phishing attacks, attack vectors and detection techniques for detecting the phishing sites. Performance comparison of 18 different models along with nine different sources of datasets are given. Challenges in phishing detection techniques are also given.

In recent days cyber-attacks are increasing at an unprecedented rate. Phishing is one among those cyberattacks. In phishing, attackers lure the end-users by making them click the hyper-links which make them lose their personally identifiable information, banking and credit card details, and passwords. In this attack the attackers disguise themselves as trusted entities such as service providers, employees of the organization or technical-support team from the organization so that end-users never doubt them. It is mainly done through emails asking to update the system, or saying that account has been suspended, or asking to claim the prize and so on [59]. The main goal of phishing is to make end-users share their sensitive information. Now-a-days information regarding anything is available online and that information is stored in websites. Websites help the end-users by providing them information about their respective products,services or helping the end-users if they face any problem by chatbots, message forums and so on. Websites also store the personal information of the end-users. As websites help the end-users in gaining information they can be used as bait for trapping the end-users to obtain confidential information from them.

1.2 Purpose

Nowadays phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. main aim of the attacker is to steal banks account credentials. phishing attacks are becoming successful because lack of user awareness. since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. phishing may be a style of broad extortion that happens once a pernicious web site act sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or mastercard numbers. all the same, the means that there square measure some of contrary to phishing programming and techniques for recognizing potential phishing tries in messages and characteristic phishing substance on locales, phishes think about new and crossbreed procedures to bypass the open programming and frameworks. phishing may be a fraud framework that uses a mixture of social designing what is additional, advancement to sensitive and personal data, as an example, passwords associate degree open-end credit unpretentious elements by presumptuous the highlights of a reliable individual or business in electronic correspondence. phishing makes use of parody messages that square measure created to seem substantial and instructed to start out from true blue sources like money connected institutions, online business goals, etc, to draw in customers to go to phony destinations through joins gave within the phishing websites.

2.LITERATURE SURVEY

2.1 Existing problem

Machine learning classifiers with wrapper features were proposed in this study. Their results were compared with the benchmark models. Machine learning with wrapper-based features outperformed the other feature selection methods. Some limitations were noticed after the evaluation of the research that was conducted in. One of these limitations is that it can't detect the embedded objects, including iframes, Flash, and HTML files to provide detection for multiple heuristics-based approaches.

Nguyen et al. presented a novel methodology for detection of phishing website based on a ML classifier as well as a wrapper features selection technique. The authors had achieved the detection by using selected supervised ML techniques. The key feature was selected by using the ML-based wrapper features technique that demonstrated a high performance for detection of phishing websites. The experimental results from this study presented better performance of the ML techniques because wrapper features selection was embedded with the proposed approach. Moreover, the ML technique and the wrapper-based features selection offered researchers an opportunity to extend their research to improve phishing websites' classification and detection. As compared to a single ML technique, the combined method worked better to achieve the targeted goals of detecting phishing websites.

Applications of ML techniques to identify phishing attacks were reported in the form of positive rate and negative rate. In this research, the authors had identified the most suitable ML algorithm for anti-phishing attacks. They had proposed a phishing classification method that captures attributes that are useful to overcome the shortcomings of phishing detection techniques. In this research, the authors had applied the use of numeric representation. Metadata of URLs were used for the determination of a website that either legitimate one or not. The authors had used ML algorithms: Random Forest, KNN, D-Tree, Linear-SVC classifier, SVM classifier, and wrapper-based (W-B) features selection. Random Forest and SVM models outperformed the rest of the models.

2.2 References

1. Phishing Activity Trends Report: 4rd Quarter 2020. Anti-Phishing Work. Group. Retrieved April 2021, 30, 2020.
2. FBI. 2019 Internet Crime Report Released-FBI. Available online: https://www.fbi.gov/news/stories/2019-internet-crime-report-released-021120. (accessed on 11 February 2020).
3. Mohammad, R.M.; Thabtah, F.; McCluskey, L. Tutorial and critical analysis of phishing websites methods. Comput. Sci. Rev. 2015, 17, 1–24. [Google Scholar] [CrossRef][Green Version]
4. Almomani, A.; Wan, T.C.; Altaher, A.; Manasrah, A.; ALmomani, E.; Anbar, M.; ALomari, E.; Ramadass, S. Evolving fuzzy neural network for phishing emails detection. J. Comput. Sci. 2012, 8, 1099. [Google Scholar]
5. Prakash, P.; Kumar, M.; Kompella, R.R.; Gupta, M. Phishnet: Predictive blacklisting to detect phishing attacks. In Proceedings of the 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, 14–19 March 2010; pp. 1–5. [Google Scholar]
6. Zhang, J.; Porras, P.A.; Ullrich, J. Highly Predictive Blacklisting. In Proceedings of the USENIX Security Symposium, San Jose, CA, USA, 28 July–1 August 2008; pp. 107–122. [Google Scholar]
7. Cao, Y.; Han, W.; Le, Y. Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM Workshop on Digital Identity Management, Alexandria, VA, USA, 31 October 2008; pp. 51–60. [Google Scholar]
8. Srinivasa Rao, R.; Pais, A.R. Detecting phishing websites using automation of human behavior. In Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security, Abu Dhabi, United Arab Emirates, 2–4 April 2017; pp. 33–42. [Google Scholar]
9. Rao, R.S.; Ali, S.T. Phishshield: A desktop application to detect phishing webpages through heuristic approach. Procedia Comput. Sci. 2015, 54, 147–156. [Google Scholar] [CrossRef][Green Version]
10. Joshi, Y.; Saklikar, S.; Das, D.; Saha, S. PhishGuard: A browser plug-in for protection from phishing. In Proceedings of the 2008 2nd International Conference on Internet Multimedia Services Architecture and Applications, Las Vegas, NV, USA, 14–17 July 2008; pp. 1–6. [Google Scholar]

2.3 Problem Statement Definition

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.

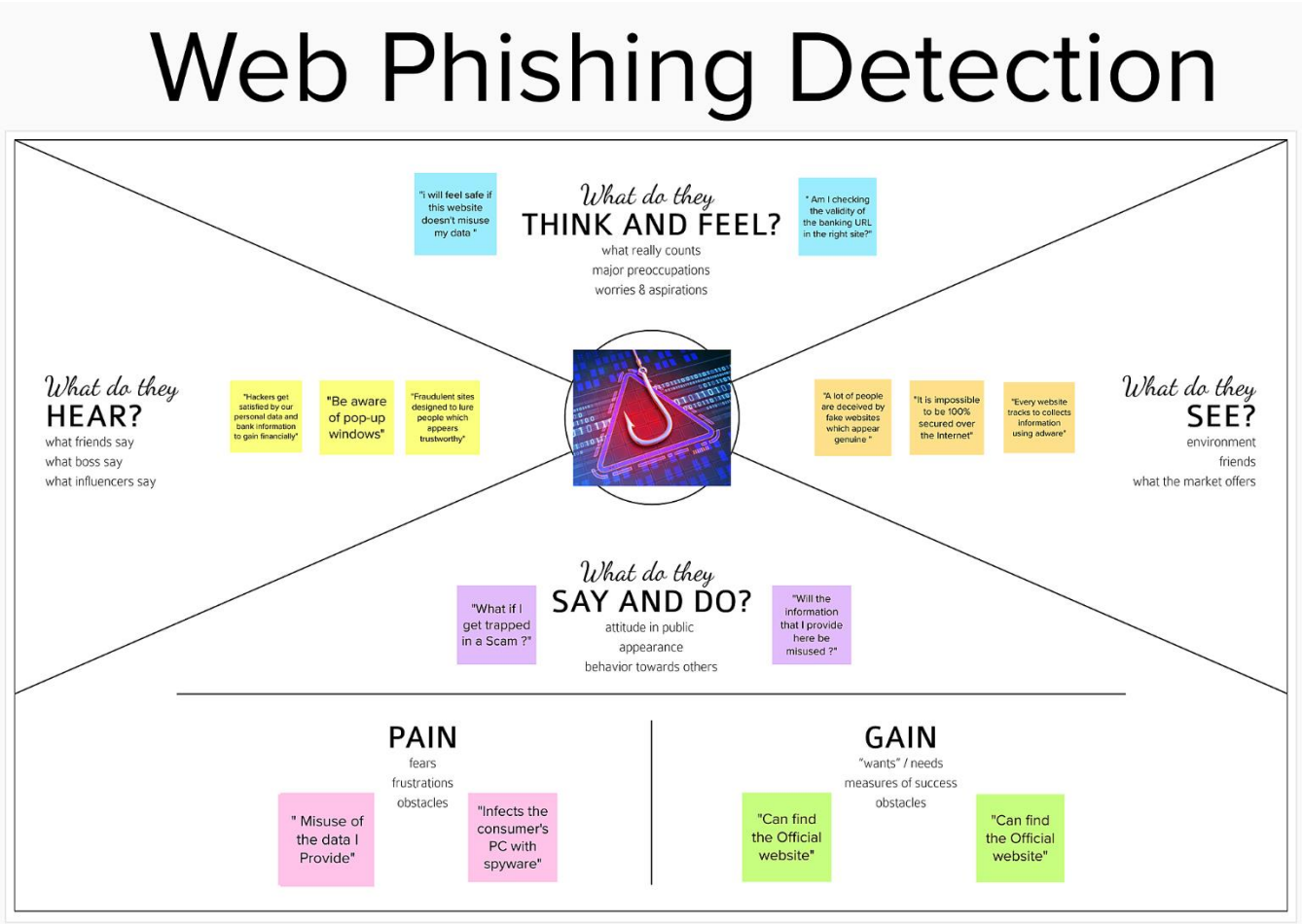
Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.

- It will lead to information disclosure and property damage.
- Large organizations may get trapped in different kinds of scams.

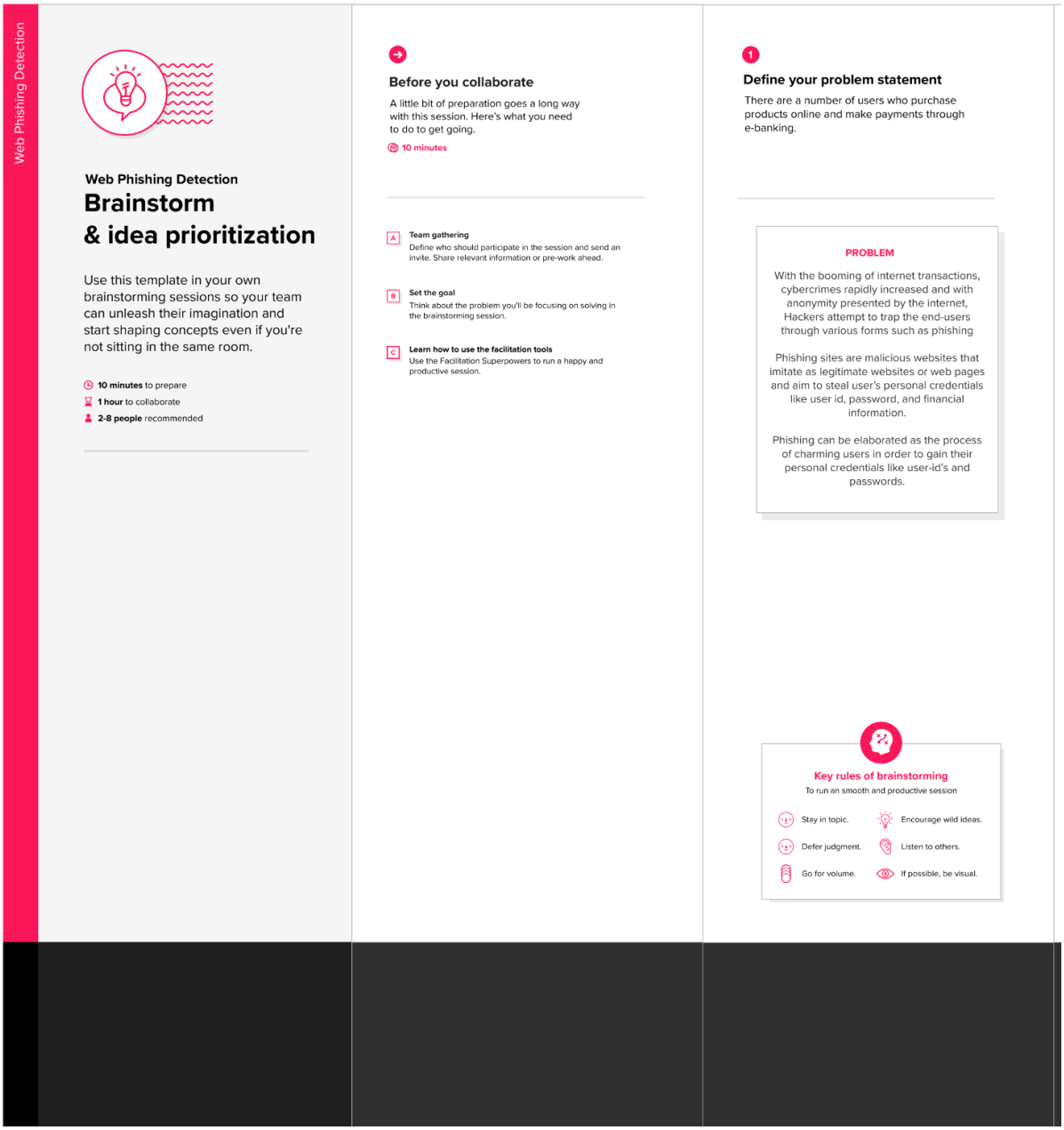
This project mainly focuses on applying a machine-learning algorithm to detect Phishing websites.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming



2

Brainstorm

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

Karthick Aravind B

- Perform Regular Scan
- Cumulative Accuracy Calculation
- Check if form resources are referencing to a malicious site.
- Prediction Model (Ensemble Learning)
- Evaluating with Test Dataset

Amarnath R

- Find special characters in the domain name to identify phishing sites.
- Phishers use long URLs to hide sensitive part of content.
- IP address in URL isn't safe
- Check if there are O's replaced with 0's or change of alphabetic cases or the presence of '@'.
- Redirection "//" should be present right after the protocol(http or https).

Peniel Branham T

- Training of Dataset using ML techniques (Random Forest, Support vector machine)
- ML Technique (Supervised Learning)
- Block the Phishing activity
- Implementation of Phishing detection
- Cloud based model is developed

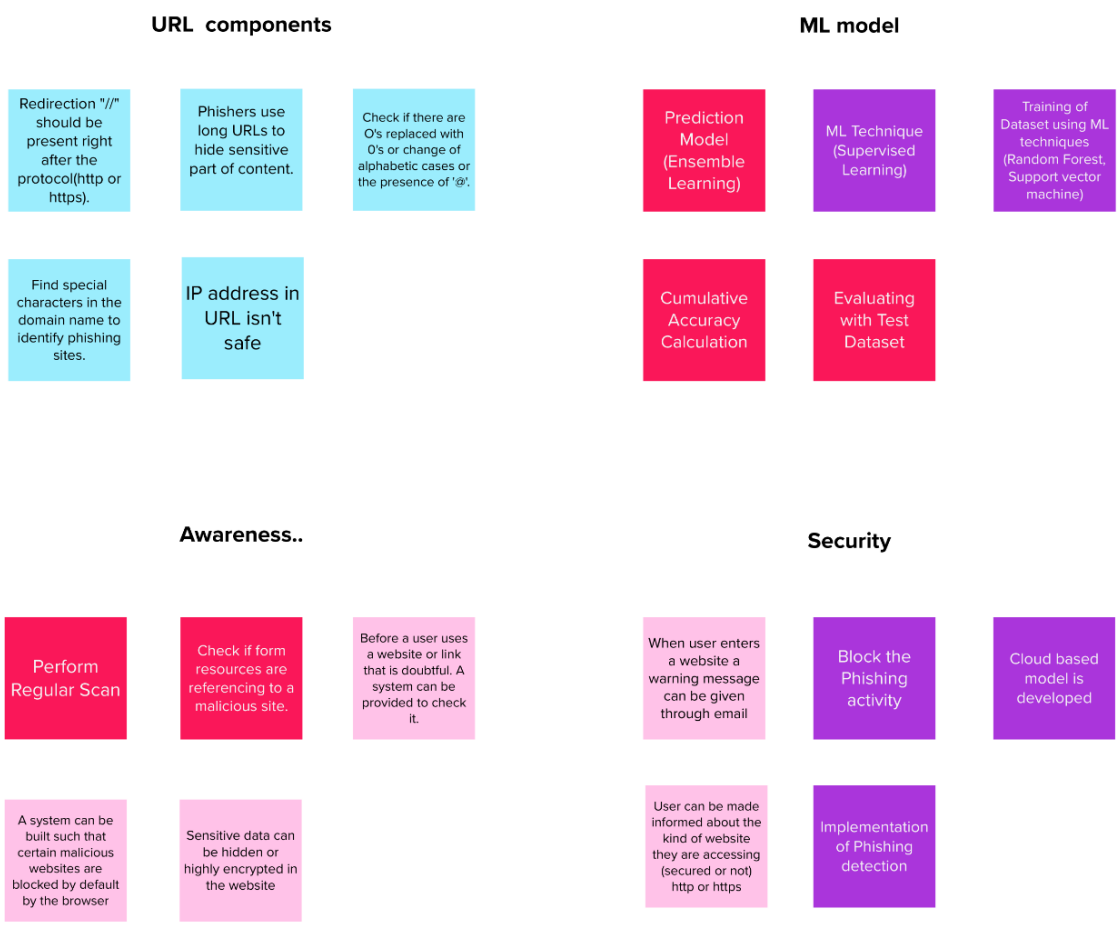
Balamurugan S

- Before a user uses a website or link that is doubtful. A system can be provided to check it.
- A system can be built such that certain malicious websites are blocked by default by the browser
- When user enters a website a warning message can be given through email
- Sensitive data can be hidden or highly encrypted in the website
- User can be made informed about the kind of website they are accessing (secured or not) http or https

3

Group ideas

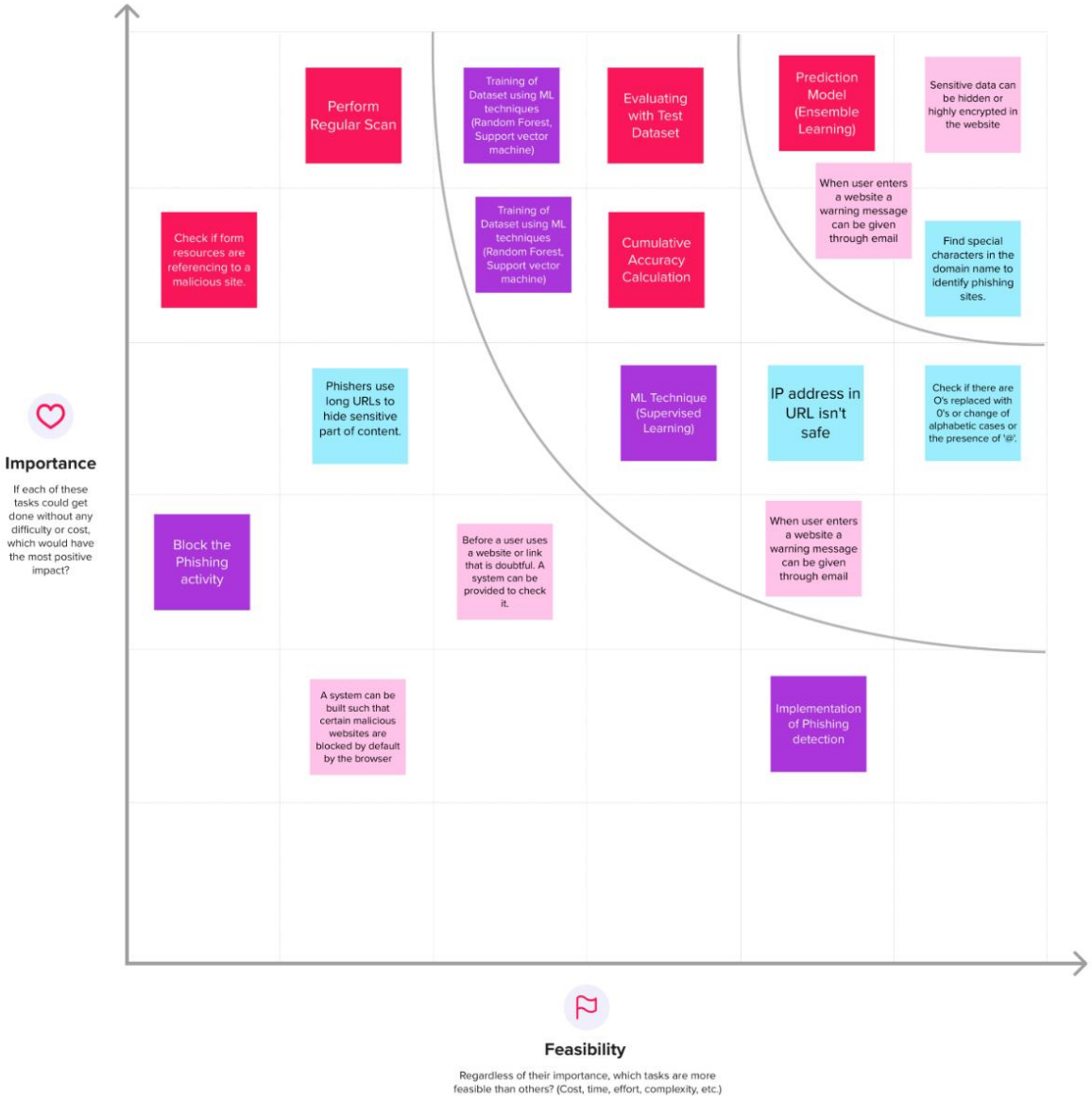
Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.



4

Prioritize

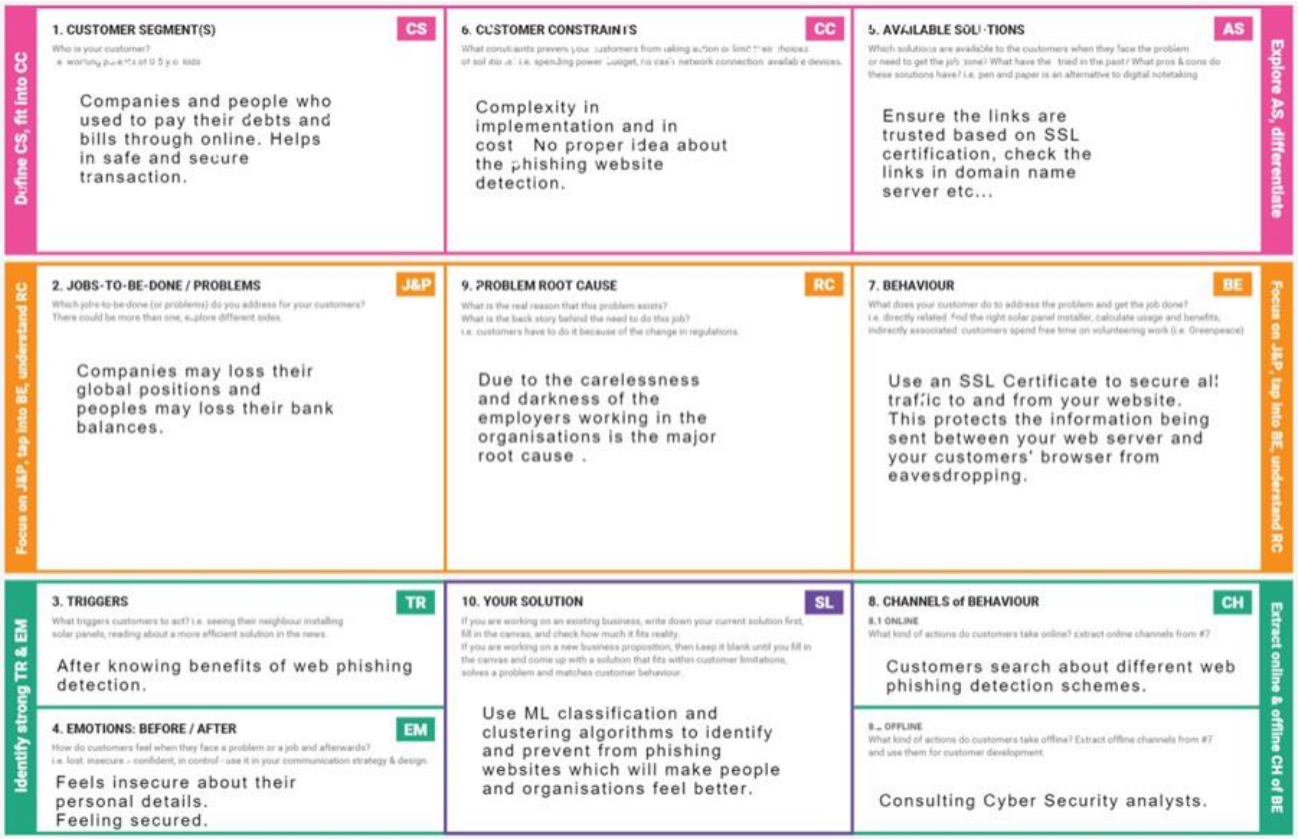
Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.



3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Phishing sites are harmful websites that mimic trustworthy websites or web pages in an effort to steal users' personal information, including their user's name, password, and credit card number. Because phishing is mostly a semantics-based assault that focuses on human vulnerabilities rather than network or software flaws, identifying these phishing websites can be difficult.
2.	Idea / Solution description	A deep learning-based framework by implementing it as a browser plug-in capable of determining whether there is a phishing risk in real-time when the user visits a web page and gives a warning message. The real-time prediction includes whitelist filtering, blacklist interception, and machine learning (ML) prediction.
3.	Novelty / Uniqueness	Feel protected by using the website as the business-related credentials will be safe. Parents can be relaxed when kids explore educational website as the fraudulent website will be detected by our website
4.	Social Impact / Customer Satisfaction	The customer will come to know whether their details are safe/ not and the customer will be restricted from entering into the phishing websites.
5.	Business Model (Revenue Model)	Visitors engage with their ads, by generating impressions, engagements or clicks.
6.	Scalability of the Solution	Cost-effective and time-saving for global users residing at global locations.

3.4 Problem Solution fit



4. REQUIREMENT ANALYSIS

4.1 Functional requirement

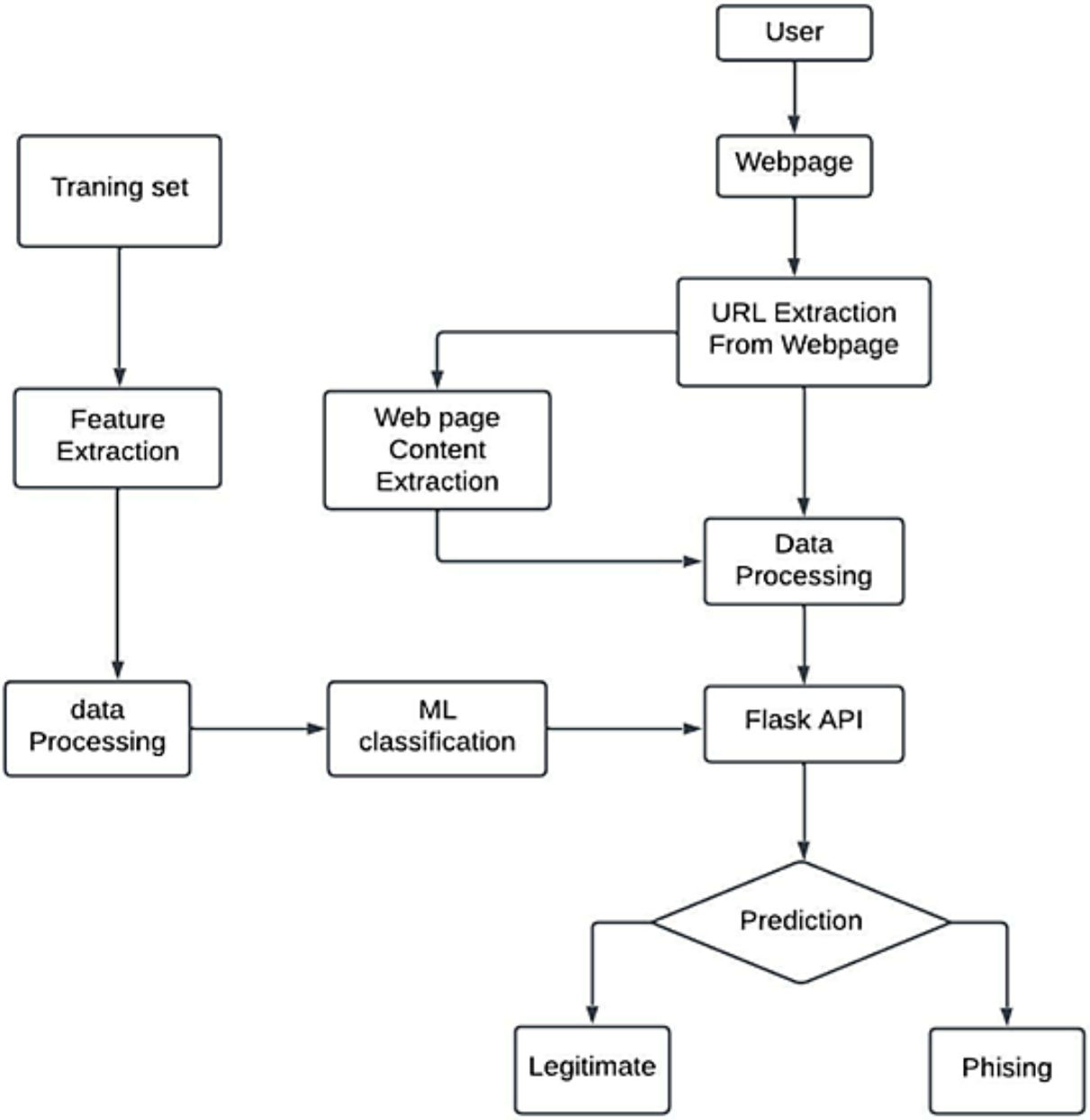
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Register by entering details such as name, email, Password.
FR-2	User Login	Login using the registered email id and password.
FR-3	Website Comparison	Blacklist filtering and Whitelist filtering techniques are used to compare the website URL.
FR-4	Feature Selection	Based on the length of an URL, number of dots in URL and check for the correct spelling and grammar.
FR-5	Feature Vectorization	Training and Testing dataset should be developed.
FR-6	Classifier	Model sends all output 10 classifier and produces final result.
FR-7	Results	Model then displays whether website is a legal site or a phishing site.

4.2 Non-Functional requirements

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	User can access to several website easily using web phishing detection without losing any data.
NFR-2	**Security**	Alert message must be sent to the users to enable secure browsing.
NFR-3	**Reliability**	The web phishing websites must detect accurately and the result must be reliable.
NFR-4	**Performance**	The performance should be faster and user friendly for the effective performance.
NFR-5	**Availability**	The system will be accessible to the user at any point in time through a web browser.
NFR-6	**Scalability**	It must be able to handle an increase in users and loads without disrupting the end users.

5. PROJECT DESIGN

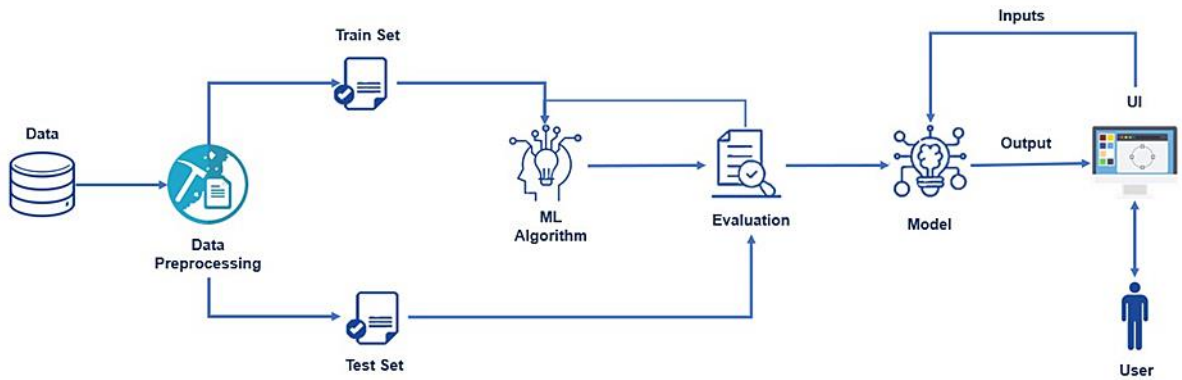
5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

1. Find the best tech solution to solve existing business problems.
2. Describe the structure, characteristics, behaviour, and other aspects of the software to project stakeholders.
3. Define features, development phases, and solution requirements.
4. Provide specifications according to which the solution is defined, managed, and delivered.



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Registration	USN-1	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
	Login	USN-2	As a user, I can log into the application by entering email & password.	I can access the website	High	Sprint-1
	Website	USN-3	As a user, I enter a website to check whether the URL is safe to enter or not.	Website should be user Friendly	High	Sprint-1
	Notification	USN-4	If the Link is Malicious, Notification has to be sent to me.	I can receive a Notification	Medium	Sprint-2
	Dashboard	USN-5	As a user, I can see the Result	I can view that it is a Safe site or not	High	Sprint-2
Customer Care Executive	Help	USN-6	As a user, I can share my Queries in the Help Textbox	I can send my Queries through it	Medium	Sprint-3
Administrator	Contact	USN-7	As a administrator, I can Answer the User Queries	I sent the Solution through User provided Email	Low	Sprint-3
		USN-8	As a Administrator, I can Improve the Accuracy	I can update the Website	High	Sprint-4

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	a user, I can register for the application by Entering my email, password, and confirming My password.	5	High	Karthick Aravind B, Peniel Branham T
Sprint-1	Login	USN-2	As a user, I can log into the application by entering email & password.	10	High	Amarnath R
Sprint-1	Website	USN-3	As a user, I enter a website to check whether the URL is safe to enter or not.	15	High	Balamurugan S, Amarnath R
Sprint-2	Notification	USN-4	If the Link is Malicious, Notification has to be sent to me.	5	Medium	Peniel Branham T, Karthick Aravind B
Sprint-2	Dashboard	USN-5	As a user, I can see the Result	15	High	Amarnath R
Sprint-3	Help	USN-6	As a user, I can share my Queries in the Help Textbox	10	Medium	Karthick Aravind B
Sprint-3	Contact	USN-7	As a administrator, I can Answer the User Queries	5	Low	Peniel Branham T
Sprint-4		USN-8	As a Administrator, I can Improve the Accuracy	10	High	Balamurugan S, Karthick Aravind B

Velocity:

Imagine we have a 6-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)

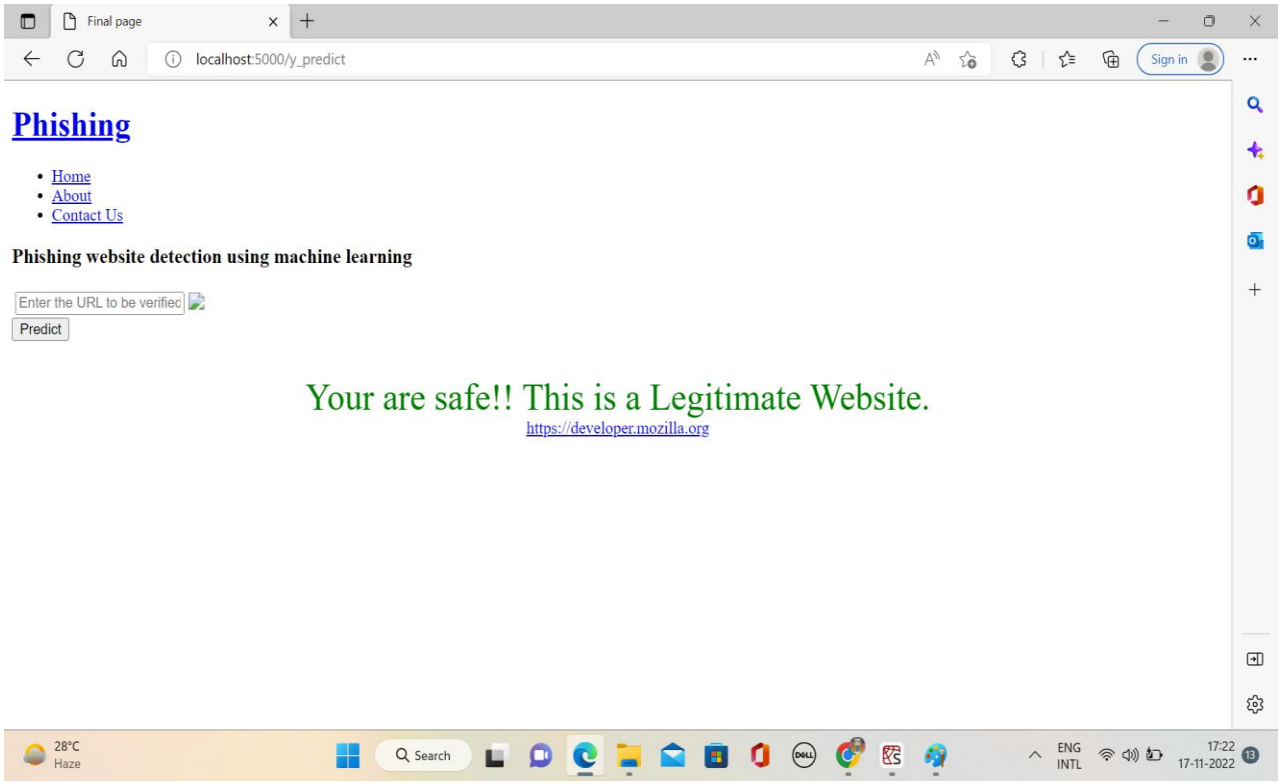
$$AV = (\text{Sprint Duration} / \text{Velocity}) = 20/6 = 3.33$$

6.2 Sprint Delivery Schedule

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

7. CODING & SOLUTIONING

7.1 Feature 1


Final page
localhost:5000/y_predict
Sign in
Phishing
Home
About
Contact Us
Phishing website detection using machine learning
Enter the URL to be verified
Predict
Your are safe!! This is a Legitimate Website.
https://developer.mozilla.org
28°C
Haze
Search
ENG
INTL
17:22
17-11-2022

8. TESTING

8.1 Test Cases

TEST CASE ID	TESTCASE/ACTION TO BE PERFORMED	EXPECTED RESULT	ACTUAL RESULT	PASS/ FAIL
1	Clicking the “Scan Now” Button	Display Scan Page	Display Scan Page	Pass
2	Entering thr URL to verified	URL	URL	Pass
3	Clicking the “Predict” Button	Your are safe!! This is a Legitimate Website	Your are safe!! This is a Legitimate Website	Pass
4	Clicking the “Predict” Button	Your are Not safe!! This is a Not Legitimate Website	Your are Not safe!! This is a Not Legitimate Website	Pass
5	Clicking the “Home” Button	Returns to Home Page	Returns to Home Page	Pass

8.2 User Acceptance Testing

ACCEPTANCE	AGREE	DISAGREE	STRONGLY AGREE
1. This Website helps me to keep my Data Secure	1		2
2. This Website is used to Detect Web Phishing Websites			3

9. RESULTS

9.1 Performance Metrics

Confusion Matrix :
[[960 54]
 [18 1179]]
Accuracy Score is 0.9674355495251018
Classification Report :

	precision	recall	f1-score	support
-1	0.98	0.95	0.96	1014
1	0.96	0.98	0.97	1197
accuracy			0.97	2211
macro avg	0.97	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

AUC-ROC: 0.9658539840726075
LOGLOSS Value is 1.1247558022138404

10. ADVANTAGES & DISADVANTAGES

Detection Technique	Advantages	Disadvantages
Blacklists	-Requiring low resources on host machine -Effective when minimal FP rates are required.	-Mitigation of zero-hour phishing attacks. -Can result in excessive queries with heavily loaded servers.
Heuristics and visual similarity	-Mitigate zerohour attacks.	-Higher FP rate than blacklists. -High computatio- nal cost.
Machine Learning	-Mitigate zerohour attacks. -Constuct own classification models.	-Time consuming. -Costly. -Huge number of rules.

11. CONCLUSION

CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. In Future System can upgrade to automatic Detect the web page and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. Phish Checker application also can be upgraded into the web phone application in detecting phishing on the mobile platform. There are many features that can be improved in the work, for various other issues. The heuristics can be further developed to detect phishing attacks in the presence of embedded objects like flash. Identity extraction is an important operation and it was improved with the Optical Character Recognition (OCR) system to extract the text and images. More effective inferring rules for identifying a given suspicious web page, and strategies for discovering if it is a phishing target, should be designed in order to further improve the overall performance of this system. Moreover, it is an open challenge to develop a robust malware detection method, retaining accuracy for future phishing emails. In addition, the dynamic and static features complement each other, and therefore both are considered important in achieving high accuracy.

12. FUTURE SCOPE

FUTURE SCOPE

In future if we get structured dataset of phishing we can perform phishing detection much more faster than any other technique. In future we can use a combination of any other two or more classifier to get maximum accuracy. We also plan to explore various phishing techniques that uses Lexical features, Network based features, Content based features, Webpage based features and HTML and JavaScript features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside, but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning.

13. APPENDIX

Source Code

```
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle
import inputscript
from gevent.pywsgi import WSGIServer

app = Flask(__name__)
model = pickle.load(open('phishing_Website.pkl', 'rb'))

@app.route('/')
def predict1():
    return render_template('index.html')

@app.route('/predict')
def predict():
    return render_template('final.html')

@app.route('/y_predict',methods=['POST'])
def y_predict():
    '''
    For rendering results on HTML GUI
    '''
    url = request.form['URL']
    checkprediction = inputscript.main(url)
    prediction = model.predict(checkprediction)
    print(prediction)
    output=prediction[0]
    if(output==1):
        pred="Your are safe!!  This is a Legitimate Website."

    else:
        pred="You are on the wrong site. Be cautious!"
    return render_template('final.html', prediction_text='{}'.format(pred),url=url)

@app.route('/predict_api',methods=['POST'])
def predict_api():
    '''
    For direct API calls trought request
    '''
    data = request.get_json(force=True)
    prediction = model.y_predict([np.array(list(data.values()))])


     output = prediction[0]
    return jsonify(output)
if __name__=='__main__':
    app.run(debug=False)


<!DOCTYPE html>
<html lang="en">
```

```
<head>
   <!-- meta tags-->
   <meta charset="utf-8">
   <meta name="viewport" content="width=device-width, initial-scale=1.0">

   <!-- Css Attachment-->
   <link rel="stylesheet" type="text/css" href="{{ url_for('static', filename='css/style.css') }}">
   <link rel="stylesheet" href="style.css">

   <title>Final page</title>
</head>

<style>
  .login{
  top: 20%;
 }
  </style>
  </head>

<body>

   <!--Header starts -->
   <header class="header" id="navbar">
      <h1 class="logo"><a href="#">Phishing</a></h1>
      <ul class="main-nav">
         <li><a href="#">Home</a></li>
         <li><a href="#">About</a></li>
         <li><a href="#">Contact Us</a></li>
      </ul>
   </header>
   <!--Header ends -->

   <!--Body starts -->
  <h3 class="headTitle">Phishing website detection using machine learning</h3>
  <form action="{{ url_for('y_predict')}}"method="post">
  <div class="boxContainer">
   <table class="elementContainer">
     <tr>
      <td>
      <input type="text" name="URL" placeholder="Enter the URL to be verified" class="search"
required="required"/>
        </td>
      <td>
        <img src="{{ url_for('static', filename='images/search.png') }}"></a>
      </td>
      </tr>
    </table>

  </div>

  <button type="submit" class="btn">Predict</button>
```

```
</form>
<div style="text-align: center ;">
<div id='result', style="color: green;padding-top: 2rem;font-size: 2.2rem;" font-size:30px;>{{
prediction_text }}</div>
  <a href=" {{ url }} "> {{ url }} </a>
</div>
  <!--Body ends -->
</html>

<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="utf-8">
  <meta content="width=device-width, initial-scale=1.0" name="viewport">

  <title>IBM Project</title>
  <meta content="" name="description">
  <meta content="" name="keywords">



  <!-- Google Fonts -->
  <link
href="https://fonts.googleapis.com/css?family=Open+Sans:300,300i,400,400i,600,600i,700,
700i|Jost:300,300i,400,400i,500,500i,600,600i,700,700i|Poppins:300,300i,400,400i,500,500i,6
00,600i,700,700i" rel="stylesheet">

  <!-- Vendor CSS Files -->
   <link href="assets/vendor/bootstrap/css/bootstrap.min.css" rel="stylesheet">
  <link href="assets/vendor/icofont/icofont.min.css" rel="stylesheet">
  <link href="assets/vendor/boxicons/css/boxicons.min.css" rel="stylesheet">
  <link href="assets/vendor/remixicon/remixicon.css" rel="stylesheet">
  <link href="assets/vendor/venobox/venobox.css" rel="stylesheet">
  <link href="assets/vendor/owl.carousel/assets/owl.carousel.min.cs"S" rel="stylesheet">
  <link href="assets/vendor/aos/aos.css" rel="stylesheet">


          <!--link href="{{ url_for('static', filename='/vendor/bootstrap/css/bootstrap.min.css')
}}" rel="stylesheet">
          <link href="{{ url_for('static', filename='/vendor/icofont/icofont.min.css') }}"
rel="stylesheet">
          <link href="{{ url_for('static', filename='/vendor/boxicons/css/boxicons.min.css') }}"
rel="stylesheet">
          <link href="{{ url_for('static', filename='/vendor/remixicon/remixicon.css') }}"
rel="stylesheet">
          <link href="{{ url_for('static', filename='/vendor/venobox/venobox.css') }}"
rel="stylesheet">
          <link href="{{ url_for('static',
filename='/vendor/owl.carousel/assets/owl.carousel.min.css') }}" rel="stylesheet">
          <link href="{{ url_for('static', filename='/vendor/aos/aos.css') }}" rel="stylesheet"-->

  <!-- Template Main CSS File -->
```

```
  <link href="assets/css/style.css" rel="stylesheet">
  <!--link href="{{ url_for('static', filename='css/style.css') }}" rel="stylesheet"-->


</head>

<body>

  <!-- ======= Header ======= -->
  <header id="header" class="fixed-top ">
    <div class="container d-flex align-items-center">

      <h1 class="logo mr-auto"><a href="index.html">Web Phishing Detection</a></h1>
      <!-- Uncomment below if you prefer to use an image logo -->
      <!-- <a href="index.html" class="logo mr-auto"><img src="assets/img/logo.png" alt=""
class="img-fluid"></a>-->

    </div>
  </header><!-- End Header -->

  <!-- ======= Hero Section ======= -->
  <section id="hero" class="d-flex align-items-center">

    <div class="container">
      <div class="row">
        <div class="col-lg-6 d-flex flex-column justify-content-center pt-4 pt-lg-0 order-2 order-lg-
1" data-aos="fade-up" data-aos-delay="200">
          <h1>Let's Find The Phishing Sites         </h1>
          <h2>Don't Let Someone To Steal Your Data</h2>
          <div class="d-lg-flex">
            <a href="http://localhost:5000/predict" class="btn-get-started scrollto">Scan
Now</a>
          </div>
        </div>
        <div class="col-lg-6 order-1 order-lg-2 hero-img" data-aos="zoom-in" data-aos-
delay="200">
          <img src="https://media.tenor.com/SRy6HR5ibvsAAAAC/phishing-phisher.gif"
class="img-fluid animated" alt="">
        </div>
      </div>
    </div>

  </section><!-- End Hero -->

  <main id="main">


    <!-- ======= About Us Section ======= -->
    <section id="about" class="about">
      <div class="container" data-aos="fade-up">

        <div class="section-title">
```

```
        <h2>About The Project</h2>
       </div>

       <div class="row content">
        <div class="col-lg-6">
         <p>
          There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.
         </p>

        </div>
        <div class="col-lg-6 pt-4 pt-lg-0">
         <p>
          In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms.  We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and security and encryption criteria in the final phishing detection rate. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.
         </p>

        </div>
       </div>

     </div>
   </section>
 <footer id="footer">



   <div class="footer-top">
    <div class="container">
     <div class="row">


     </div>
    </div>
   </div>

   <div class="container footer-bottom clearfix">

 </footer>

 <a href="#" class="back-to-top"><i class="ri-arrow-up-line"></i></a>
 <div id="preloader"></div>
```

```
  <script src="assets/vendor/jquery/jquery.min.js"></script>
  <script src="assets/vendor/bootstrap/js/bootstrap.bundle.min.js"></script>
  <script src="assets/vendor/jquery.easing/jquery.easing.min.js"></script>
  <script src="assets/vendor/php-email-form/validate.js"></script>
  <script src="assets/vendor/waypoints/jquery.waypoints.min.js"></script>
  <script src="assets/vendor/isotope-layout/isotope.pkgd.min.js"></script>
  <script src="assets/vendor/venobox/venobox.min.js"></script>
  <script src="assets/vendor/owl.carousel/owl.carousel.min.js"></script>
  <script src="assets/vendor/aos/aos.js"></script>

  <script src="assets/js/main.js"></script>


</body>

</html>

import regex
from tldextract import extract
import ssl
import socket
from bs4 import BeautifulSoup
import urllib.request
import whois
import datetime
import requests
import favicon
import re
import google
import xmltodict
from googlesearch import search
def having_IPhaving_IP_Address(url):
    match=regex.search(
  '(([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-
4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\/)|'  #IPv4
                '((0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-
F]{1,2})\\/)'  #IPv4 in hexadecimal
                '(?:[a-fA-F0-9]{1,4}:){7}[a-fA-F0-9]{1,4}',url)     #Ipv6
    if match:
      return -1
    else:
      return 1

def URLURL_Length (url):
   length=len(url)
   if(length<=75):
       if(length<54):
          return 1
       else:
          return 0
   else:
      return -1
```

```
def Shortining_Service (url):

match=regex.search('bit\.ly|goo\.gl|shorte\.st|go2l\.ink|x\.co|ow\.ly|t\.co|tinyurl|tr\.im|is\.gd|
cli\.gs|'

'yfrog\.com|migre\.me|ff\.im|tiny\.cc|url4\.eu|twit\.ac|su\.pr|twurl\.nl|snipurl\.com|'

'short\.to|BudURL\.com|ping\.fm|post\.ly|Just\.as|bkite\.com|snipr\.com|fic\.kr|loopt\.us|'

'doiop\.com|short\.ie|kl\.am|wp\.me|rubyurl\.com|om\.ly|to\.ly|bit\.do|t\.co|lnkd\.in|'
                'db\.tt|qr\.ae|adf\.ly|goo\.gl|bitly\.com|cur\.lv|tinyurl\.com|ow\.ly|bit\.ly|ity\.im|'

'q\.gs|is\.gd|po\.st|bc\.vc|twitthis\.com|u\.to|j\.mp|buzurl\.com|cutt\.us|u\.bb|yourls\.org|'

'x\.co|prettylinkpro\.com|scrnch\.me|filoops\.info|vzturl\.com|qr\.net|1url\.com|tweez\.me|v
\.gd|tr\.im|link\.zip\.net',url)
    if match:
        return -1
    else:
        return 1

def having_At_Symbol(url):
    symbol=regex.findall(r'@',url)
    if(len(symbol)==0):
        return 1
    else:
        return -1

def double_slash_redirecting(url):
    for i in range(8,len(url)):
        if(url[i]=='/'):

            if(url[i-1]=='/'):
                return -1
    return 1

def Prefix_Suffix(url):
    subDomain, domain, suffix = extract(url)
    if(domain.count('-')):
        return -1
    else:
        return  1

def having_Sub_Domain(url):
    subDomain, domain, suffix = extract(url)
    if(subDomain.count('.')<=2):
        if(subDomain.count('.')<=1):
            return 1
        else:
            return 0
    else:
```

```
        return -1

def SSLfinal_State(url):
    try:
        response = requests.get(url)
        return 1
    except Exception as e:
        return -1

def Domain_registeration_length(url):
    try:
        domain = whois.whois(url)
        exp=domain.expiration_date[0]
        up=domain.updated_date[0]
        domainlen=(exp-up).days
        if(domainlen<=365):
            return -1
        else:
            return 1
    except:
        return -1

def Favicon(url):
    subDomain, domain, suffix = extract(url)
    b=domain
    try:
        icons = favicon.get(url)
        icon = icons[0]
        subDomain, domain, suffix =extract(icon.url)
        a=domain
        if(a==b):
            return 1
        else:
            return -1
    except:
        return -1

def port(url):
    try:
        a_socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
        location=(url[7:],80)
        result_of_check = a_socket.connect_ex(location)
        if result_of_check == 0:
            return 1
        else:
            return -1
        a_socket.close
    except:
        return -1

def HTTPS_token(url):
    match=re.search('https://|http://',url)
```

```
    if (match.start(0)==0):
        url=url[match.end(0):]
    match=re.search('http|https',url)
    if match:
        return -1
    else:
        return 1


def Request_URL(url):
    try:
        subDomain, domain, suffix = extract(url)
        websiteDomain = domain

        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        imgs = soup.findAll('img', src=True)
        total = len(imgs)

        linked_to_same = 0
        avg =0
        for image in imgs:
            subDomain, domain, suffix = extract(image['src'])
            imageDomain = domain
            if(websiteDomain==imageDomain or imageDomain==''):
                linked_to_same = linked_to_same + 1
        vids = soup.findAll('video', src=True)
        total = total + len(vids)

        for video in vids:
            subDomain, domain, suffix = extract(video['src'])
            vidDomain = domain
            if(websiteDomain==vidDomain or vidDomain==''):
                linked_to_same = linked_to_same + 1
        linked_outside = total-linked_to_same
        if(total!=0):
            avg = linked_outside/total

        if(avg<0.22):
            return 1
        else:
            return -1
    except:
        return -1

def URL_of_Anchor(url):
    try:
        subDomain, domain, suffix = extract(url)
        websiteDomain = domain

        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
```

```
        anchors = soup.findAll('a', href=True)
        total = len(anchors)
        linked_to_same = 0
        avg = 0
        for anchor in anchors:
            subDomain, domain, suffix = extract(anchor['href'])
            anchorDomain = domain
            if(websiteDomain==anchorDomain or anchorDomain==''):
                linked_to_same = linked_to_same + 1
        linked_outside = total-linked_to_same
        if(total!=0):
            avg = linked_outside/total

        if(avg<0.31):
            return 1
        elif(0.31<=avg<=0.67):
            return 0
        else:
            return -1
    except:
        return 0


def Links_in_tags(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')

        no_of_meta =0
        no_of_link =0
        no_of_script =0
        anchors=0
        avg =0
        for meta in soup.find_all('meta'):
            no_of_meta = no_of_meta+1
        for link in soup.find_all('link'):
            no_of_link = no_of_link +1
        for script in soup.find_all('script'):
            no_of_script = no_of_script+1
        for anchor in soup.find_all('a'):
            anchors = anchors+1
        total = no_of_meta + no_of_link + no_of_script+anchors
        tags = no_of_meta + no_of_link + no_of_script
        if(total!=0):
            avg = tags/total

        if(avg<0.25):
            return -1
        elif(0.25<=avg<=0.81):
            return 0
        else:
            return 1
```

```
    except:
        return 0


def SFH(url):
    return -1

def Submitting_to_email(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        if(soup.find('mailto:','mail():')):
            return -1
        else:
            return 1
    except:
        return -1

def Abnormal_URL(url):
    subDomain, domain, suffix = extract(url)
    try:
        domain = whois.whois(url)
        hostname=domain.domain_name[0].lower()
        match=re.search(hostname,url)
        if match:
            return 1
        else:
            return -1
    except:
        return -1

def Redirect(url):
    try:
        request = requests.get(url)
        a=request.history
        if(len(a)<=1):
            return 1
        else:
            return 0

    except:
        return 0

def on_mouseover(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')

        no_of_script =0
        for meta in soup.find_all(onmouseover=True):
            no_of_script = no_of_script+1
        if(no_of_script==0):
```

```
            return 1
        else:
            return -1
    except:
        return -1

def RightClick(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        if(soup.find_all('script',mousedown=True)):
            return -1
        else:
            return 1
    except:
        return -1

def popUpWidnow(url):
    return 1

def Iframe(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        nmeta=0
        for meta in soup.findAll('iframe',src=True):
            nmeta= nmeta+1
        if(nmeta!=0):
              return -1
        else:
            return 1
    except:
        return -1

def age_of_domain(url):
    try:
        w = whois.whois(url).creation_date[0].year
        if(w<=2018):
            return 1
        else:
            return -1
    except Exception as e:
        return -1

def DNSRecord(url):

    subDomain, domain, suffix = extract(url)
    try:
        dns = 0
        domain_name = whois.whois(url)
    except:
        dns = 1
```

```
    if(dns == 1):
        return -1
    else:
        return 1

def web_traffic(url):+ url)
        dict_data = xmltodict.parse(response.content)
        rank=dict_data['ALEXA']['SD'][1]['REACH']['@RANK']


    except TypeError:
        return -1
    rank= int(rank)
    if (rank<100000):
        return 1
    else:
        return 0

def Page_Rank(url):
    return 1

def Google_Index(url):
    try:
        subDomain, domain, suffix = extract(url)
        a=domain + '.' + suffix
        query = url
        for j in search(query, tld="co.in", num=5, stop=5, pause=2):
            subDomain, domain, suffix = extract(j)
            b=domain + '.' + suffix
        if(a==b):
            return 1
        else:
            return -1
    except:
        return -1

def Links_pointing_to_page (url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        count = 0
     try:
        response = requests.get("http://data.alexa.com/data?cli=10&dat=s&url="

        for link in soup.find_all('a'):
            count += 1
        if(count>=2):
            return 1
        else:
            return 0
    except:
```

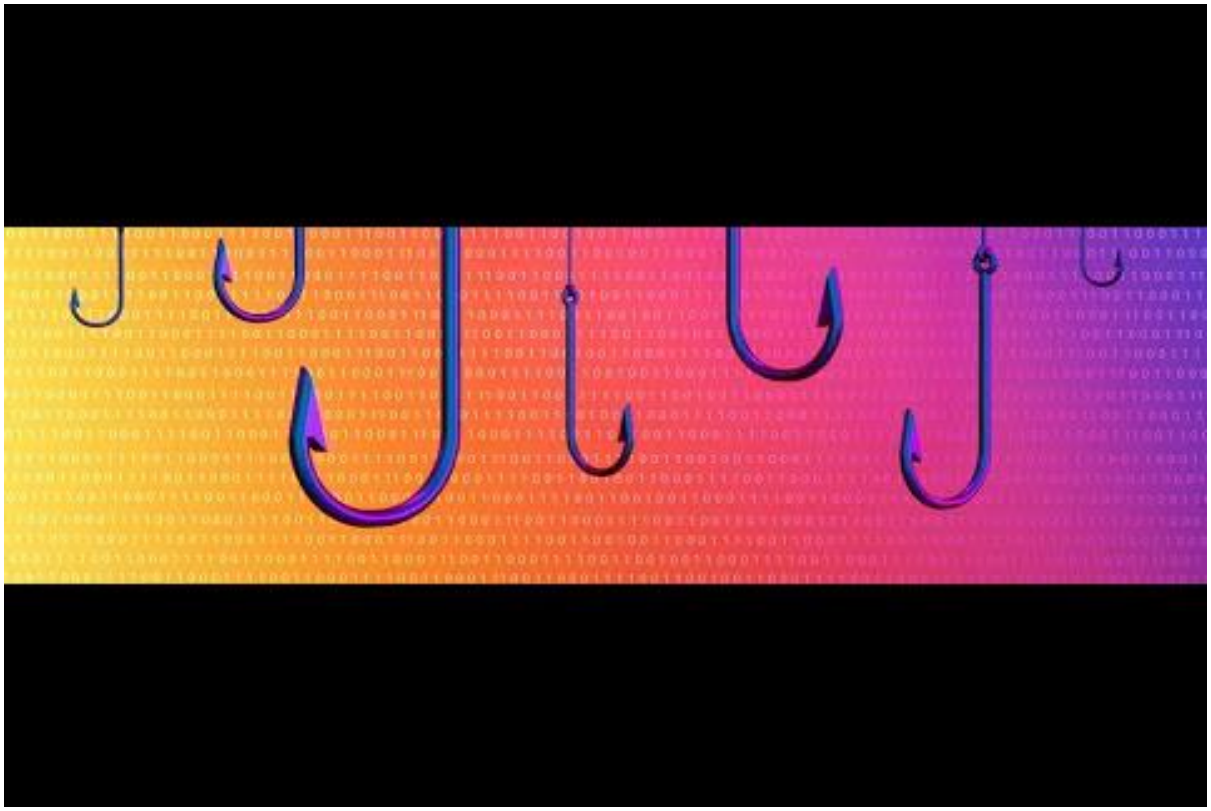
```
        return -1

def Statistical_report (url):
    hostname = url
    h = [(x.start(0), x.end(0)) for x in
regex.finditer('https://|http://|www.|https://www.|http://www.', hostname)]
    z = int(len(h))
    if z != 0:
        y = h[0][1]
        hostname = hostname[y:]
        h = [(x.start(0), x.end(0)) for x in regex.finditer('/', hostname)]
        z = int(len(h))
        if z != 0:
            hostname = hostname[:h[0][0]]

url_match=regex.search('at\.ua|usa\.cc|baltazarpresentes\.com\.br|pe\.hu|esy\.es|hol\.es|s
weddy\.com|myjino\.ru|96\.lt|ow\.ly',url)
    try:
        ip_address = socket.gethostbyname(hostname)

ip_match=regex.search('146\.112\.61\.108|213\.174\.157\.151|121\.50\.168\.88|192\.185\.
217\.116|78\.46\.211\.158|181\.174\.165\.13|46\.242\.145\.103|121\.50\.168\.40|83\.125\
.22\.219|46\.242\.145\.98|107\.151\.148\.44|107\.151\.148\.107|64\.70\.19\.203|199\.184
\.144\.27|107\.151\.148\.108|107\.151\.148\.109|119\.28\.52\.61|54\.83\.43\.69|52\.69\.1
66\.231|216\.58\.192\.225|118\.184\.25\.86|67\.208\.74\.71|23\.253\.126\.58|104\.239\.1
57\.210|175\.126\.123\.219|141\.8\.224\.221|10\.10\.10\.10|43\.229\.108\.32|103\.232\.2
15\.140|69\.172\.201\.153|216\.218\.185\.162|54\.225\.104\.146|103\.243\.24\.98|199\.59
\.243\.120|31\.170\.160\.61|213\.19\.128\.77|62\.113\.226\.131|208\.100\.26\.234|195\.1
6\.127\.102|195\.16\.127\.157|34\.196\.13\.28|103\.224\.212\.222|172\.217\.4\.225|54\.7
2\.9\.51|192\.64\.147\.141|198\.200\.56\.183|23\.253\.164\.103|52\.48\.191\.26|52\.214\.
197\.72|87\.98\.255\.18|209\.99\.17\.27|216\.38\.62\.18|104\.130\.124\.96|47\.89\.58\.14
1|78\.46\.211\.158|54\.86\.225\.156|54\.82\.156\.19|37\.157\.192\.102|204\.11\.56\.48|11
0\.34\.231\.42',ip_address)
    except:
        return -1

    if url_match:
        return -1
    else:
        return 1

def main(url):



    check = [[having_IPhaving_IP_Address
(url),URLURL_Length(url),Shortining_Service(url),having_At_Symbol(url),

double_slash_redirecting(url),Prefix_Suffix(url),having_Sub_Domain(url),SSLfinal_State(url),

Domain_registeration_length(url),Favicon(url),port(url),HTTPS_token(url),Request_URL(url),
```

```
URL_of_Anchor(url),Links_in_tags(url),SFH(url),Submitting_to_email(url),Abnormal_URL(url),
        Redirect(url),on_mouseover(url),RightClick(url),popUpWidnow(url),Iframe(url),

age_of_domain(url),DNSRecord(url),web_traffic(url),Page_Rank(url),Google_Index(url),
        Links_pointing_to_page(url),Statistical_report(url)]]


    print(check)
    return check
```

GitHub Link

https://github.com/IBM-EPBL/IBM-Project-3867-1658667541

Project Demo Link



[OR]

https://drive.google.com/file/d/11QliBPxZRMTxPUATO6SUWQNRB9DArXCD/view?usp=sharing