GOVERNMENT COLLEGE OF ENGINEERING(Formerly IRTT) ERODE-638 316

BONAFIDE CERTIFICATE

Certified that this project titled **“WEB PHISHING DETECTION USING DATA SCIENCE”** is the bonafide work of **“SAGAR M(731119104040), SATHISH G(731119104042), SETHU K(731119104043), YUVAN JEYAN G(731119104054)”** who carried out the project work under my supervision.

SIGNATURE OF HOD	**SIGNATURE OF SPOC**
Dr.A SARADHA,M.E.,Ph.D.,	Dr.G.GOWRISON, M.E.,Ph.D.,
HEAD OF THE DEPARTMENT	ASSISTANT PROFESSOR(SR)
DEPARTMENT OF CSE,	DEPARTMENT OF ECE,
GOVERNMENT COLLEGE OF	GOVERNMENT COLLEGE OF
ENGINEERING, ERODE – 638316	ENGINEERING,ERODE - 638316
SIGNATURE OF FACULTY MENTOR	**SIGNATURE OF FACULTY**
EVALUATOR	
Dr.A SARADHA,M.E.,Ph.D.,	Dr.A KAVIDHA,M.E.,Ph.D.,
ASSISTANT PROFESSOR(SR)	HEAD OF THE DEPARTMENT
DEPARTMENT OF CSE,	DEPARTMENT OF CSE,
GOVERNMENT COLLEGE OF	GOVERNMENT COLLEGE OF
ENGINEERING, ERODE – 638316	ENGINEERING, ERODE - 638316

Project Report Format

1.INTRODUCTION

1.1PROJECT OVERVIEW

Today's growing phishing websites pose significant threats due to their extremely undetectable risk. They anticipate internet users to mistake them as genuine ones in order to reveal user information and privacy, such as login ids, pass-words, credit card numbers, etc. without notice. This paper proposes a new approach to solve the anti-phishing problem. The new features of this approach can be represented by URL character sequence without phishing prior knowledge, various hyperlink information, and textual content of the webpage, which are combined and fed to train the XGBoost classifier. One of the major contributions of this paper is the selection of different new features, which are capable enough to detect 0-h attacks, and these features do not depend on any third-party services. In particular, we extract character level Term Frequency-Inverse Document Frequency (TF-IDF) features from noisy parts of HTML and plaintext of the given webpage. Moreover, our proposed hyperlink features determine the relationship between the content and the URL of a webpage. Due to the absence of publicly available large phishing data sets, we needed to create our own data set with 60,252 webpages to validate the proposed solution. This data contains 32,972 benign webpages and 27,280 phishing webpages. For evaluations, the performance of each category of the proposed feature set is evaluated, and various classification algorithms are employed. From the empirical results, it was observed that the proposed individual features are valuable for phishing detection. However, the integration of all the features improves the detection of phishing sites with significant accuracy. The proposed approach achieved an accuracy of 96.76% with only 1.39% false-positive rate on our dataset, and an accuracy of 98.48% with 2.09% false-positive rate on benchmark dataset, which outperforms the existing baseline approaches.

1.2 PURPOSE

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Phishing is a form of fraudulent attack where the attacker tries to gain sensitive information by posing as a reputable source. In a typical phishing attack, a victim opens a compromised link that poses as a credible website. The victim is then asked to enter their credentials, but since it is a “fake” website, the sensitive information is routed to the hacker and the victim gets hacked.

Phishing is popular since it is a low effort, high reward attack. Most modern web browsers, antivirus software and email clients are pretty good at detecting phishing websites at the source, helping to prevent attacks. The situation worsens when a lazy algorithm is trained and tested with a large dataset. Therefore, the performance of the research methodology used in this project may not perform so well if the wrong classifier is trained and tested with dataset size more than the classifier’s capacity.

LITERATURE SURVEY

2.1 EXISITING PROBLEM

- Users are easily deceived by fake websites, if they are first time consumers of the official product or service.

- User may loss very information to these sites and may face unrecoverable loss of economy or resources.

- The major issue is the lack of awareness due to which the user may loss a great range of personal information andsensitive paves way for the attack. This may also create blame on the user for unauthorised purchase, identity theft and other cybercriminal activities.

- This issue occurs when the user is in a hurry that makes them unaware of the details that indicate the fabricated site.

- Also occurs when the user is a beginner consumer of the service or website of concern.

- 96% of phishing attacks arrive by email. Another 3% are carried out through malicious websites and just 1% via phone.

2.2 REFERENCES

1. Dhanalakshmi, R & Prabhu, C & Chellapan, C ’ Detection of Phishing Websites and Secure Transactions Detection of Phishing Websites and Secure Transactions’. International Journal Communication & Network Security (IJCNS).

Jian Mao, Wenqian Tian, Pei Li, Tao Wei and Zhenkai Liang ‘Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity’. IEEE Access (Volume: 5) 23 August 2017

2. Ramana Rao Kompella, and Minaxi Gupta. ‘A machine learning based approach for phishing detection using hyperlinks information’ vol.12, no.2, pp.1–27, 2007.

3. Pawan Prakash, Manish Kumar ‘Phish net: predictive blacklisting to detect phishing attacks. SANS Institute, 2007. Accessed Jan 2018.

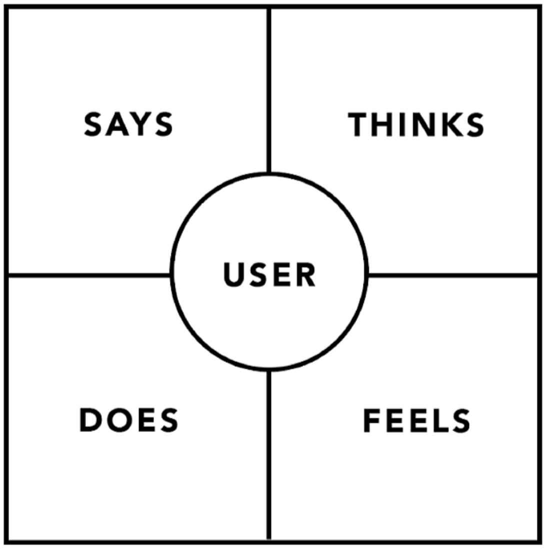
4. Dr. Gunikhan Sonowal: ‘Phishing Scams Cost American Businesses Half A Billion Dollars a Year’. Forbes, 5 May 2017. Accessed Jan 2018.

2.3 PROBLEM STATEMENT DEFINITION

The URL of phishing websites may be very similar to real websites to the human eye, but they are different in IP. The content-based detection usually refers to the detection of phishing sites through the pages of elements, such as form information, field names, and resource reference.

IDEATION &PROPOSED SOLUTION

3.1 **EMPATHY MAP CANVAS**



An **empathy map** is a collaborative visualization used to articulate what we know about a particular type of user. It externalizes knowledge about users in order to

- create a shared understanding of user needs, and
- aid in decision making.

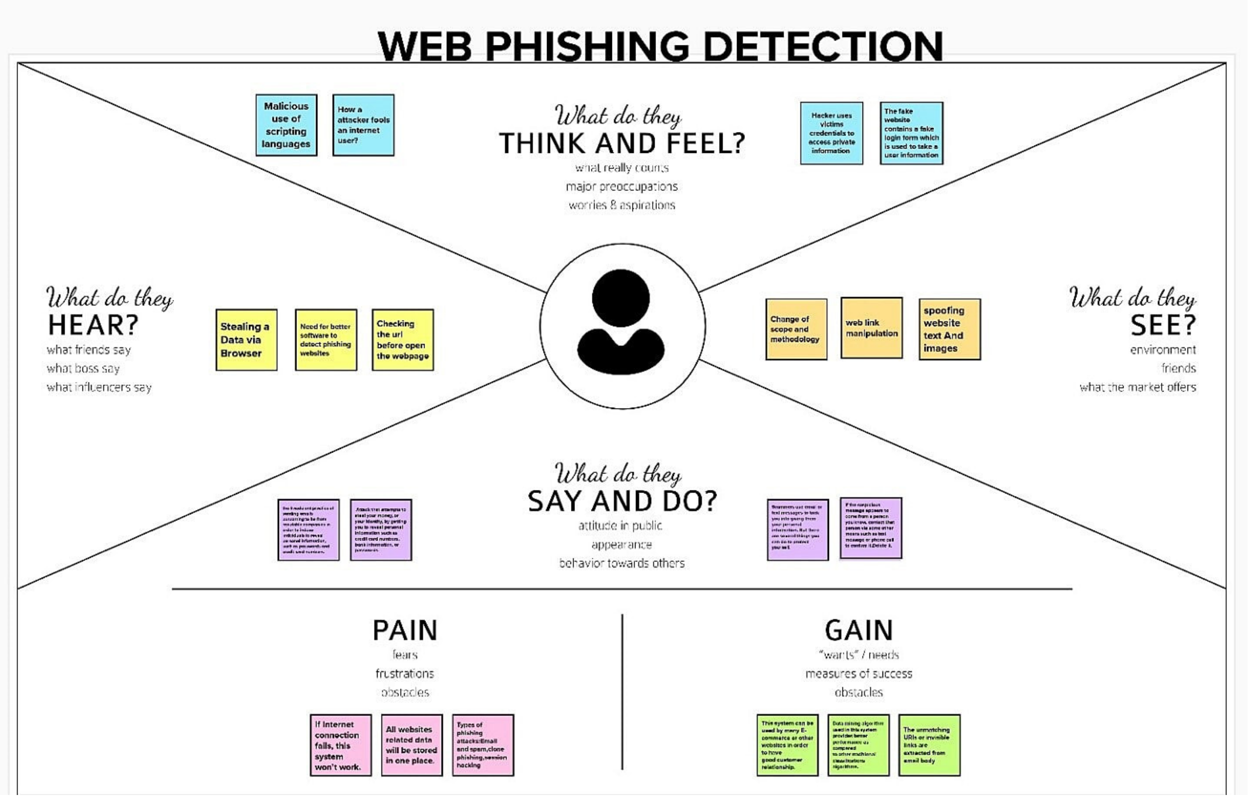
Traditional empathy maps are split into 4 quadrants (*Says*, *Thinks*, *Does*, and *Feels*), with the user or persona in the middle. Empathy maps provide a glance into who a user is as a whole and are **not** chronological or sequential

- ***Says*** quadrant contains what the user says out loud in an interview or some other usability study.
- ***Thinks*** quadrant captures what the user is thinking throughout the experience.
- ***Does*** quadrant encloses the actions the user takes.
- ***Feels*** quadrant is the user's emotional state, often represented as an adjective plus a short sentence for context.

As their name suggests, empathy maps simply help us build empathy with our end users. When based on real data and when combined with other mapping method, they can:

- Remove bias from our designs and align the team on a single, shared understanding of the user
- Discover weaknesses in our research

- ➤ Uncover user needs that the user themselves may not even be aware of
- ➤ Understand what drives users' behaviours
- ➤ Guide us towards meaningful innovation


WEB PHISHING DETECTION
What do they THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
Malicious use of scripting languages
How a attacker fools an internet user?
Hacker uses victims credentials to access private information
The fake website contains a fake login form which is used to take a user information
What do they HEAR?
what friends say
what boss say
what influencers say
Stealing a Data via Browser
Need for better software to detect phishing websites
Checking the url before open the webpage
What do they SEE?
environment
friends
what the market offers
Change of scope and methodology
web link manipulation
spoofing website text And images
What do they SAY AND DO?
attitude in public
appearance
behavior towards others
the fraudulent practice of sending emails purporting to be from reputable companies in order to induce individuals to reveal personal information, such as passwords and credit card numbers.
Attack that attempts to steal your money, or your identity, by getting you to reveal personal information such as credit card numbers, bank information, or passwords.
Scammers use email or text messages to trick you into giving them your personal information. But there are several things you can do to protect yourself.
If the suspicious message appears to come from a person you know, contact that person via some other means such as text message or phone call to confirm it.Delete it.
PAIN
fears
frustrations
obstacles
If internet connection fails, this system won't work.
All websites related data will be stored in one place.
Types of phishing attacks:Email and spam,clone phishing,session hacking
GAIN
"wants" / needs
measures of success
obstacles
This system can be used by many E-commerce or other websites in order to have good customer relationship.
Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
The unmatching URLs or invisible links are extracted from email body

3.2 IDEATION & BRAINSTORMING

Brainstorming is a great way to generate a lot of ideas that you would not be able to generate by just sitting down with a pen and paper. The intention of brainstorming is to leverage the collective thinking of the group, by engaging with each other, listening, and building on other ideas. Conducting a brainstorm also creates a distinct segment of time when you intentionally turn up the generative part of your brain and turn down the evaluative part. You can use brainstorming throughout any design or work process, of course, to generate ideas for design solutions, but also any time you are trying to generate ideas, such as planning where to do empathy work, or thinking about product and services related to our project.

Brainstorming has remained the cornerstone of the creative industry for decades and has evolved over the years as experience and learning developed from its variety of applications. Brainstorming essentially relies on a group of people coming together with their prior knowledge and research in order to gather ideas for solving the stated problem. It evokes images of exploration, experimental thinking, and wild ideas.

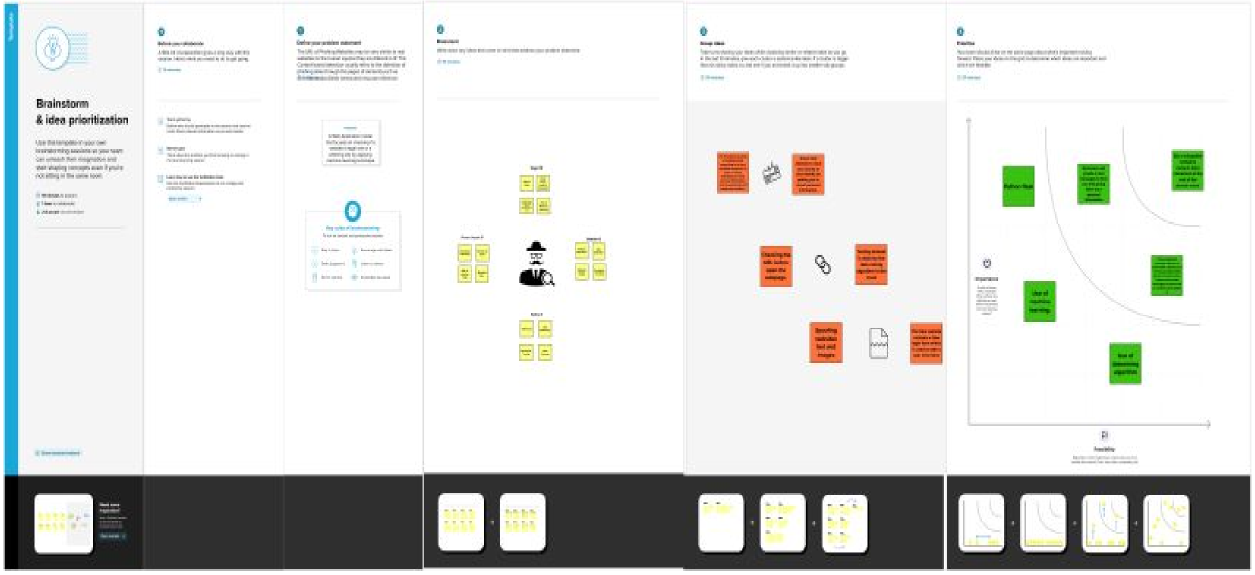
However, all too often it takes the form of controlled sessions where dominant figures assert themselves over others and creativity ends up getting stifled. Or, in other cases, the facilitator does not succeed in helping the team steer towards the goal by keeping the user, the user's need and the team's insights about them in a problem statement – also called Point-of-View at the front of the team's minds.

Ideation, is an art form, which is dependent on appointing an experienced facilitator and having an experienced team. However, we're all here to learn, and here is how you can start learning to become a successful facilitator of brainstorming sessions. Brainstorming is about setting a safe, creative space for people to feel like they can say anything and be wild—and know that they will not be judged for doing so—so that new ideas can be born.

- Set a time limit
- Start with a problem statement, Point of View, How Might We questions, a plan or a goal – and stay focused on the topic
- Defer judgement or criticism, including non-verbal
- Encourage weird, wacky, and wild ideas
- Aim for quantity
- Build on each other's ideas
- Be visual
- One conversation at a time

BRAINSTORM AND IDEATION FOR

WEB PHISHING DETECTION



Brainstorm & idea prioritization

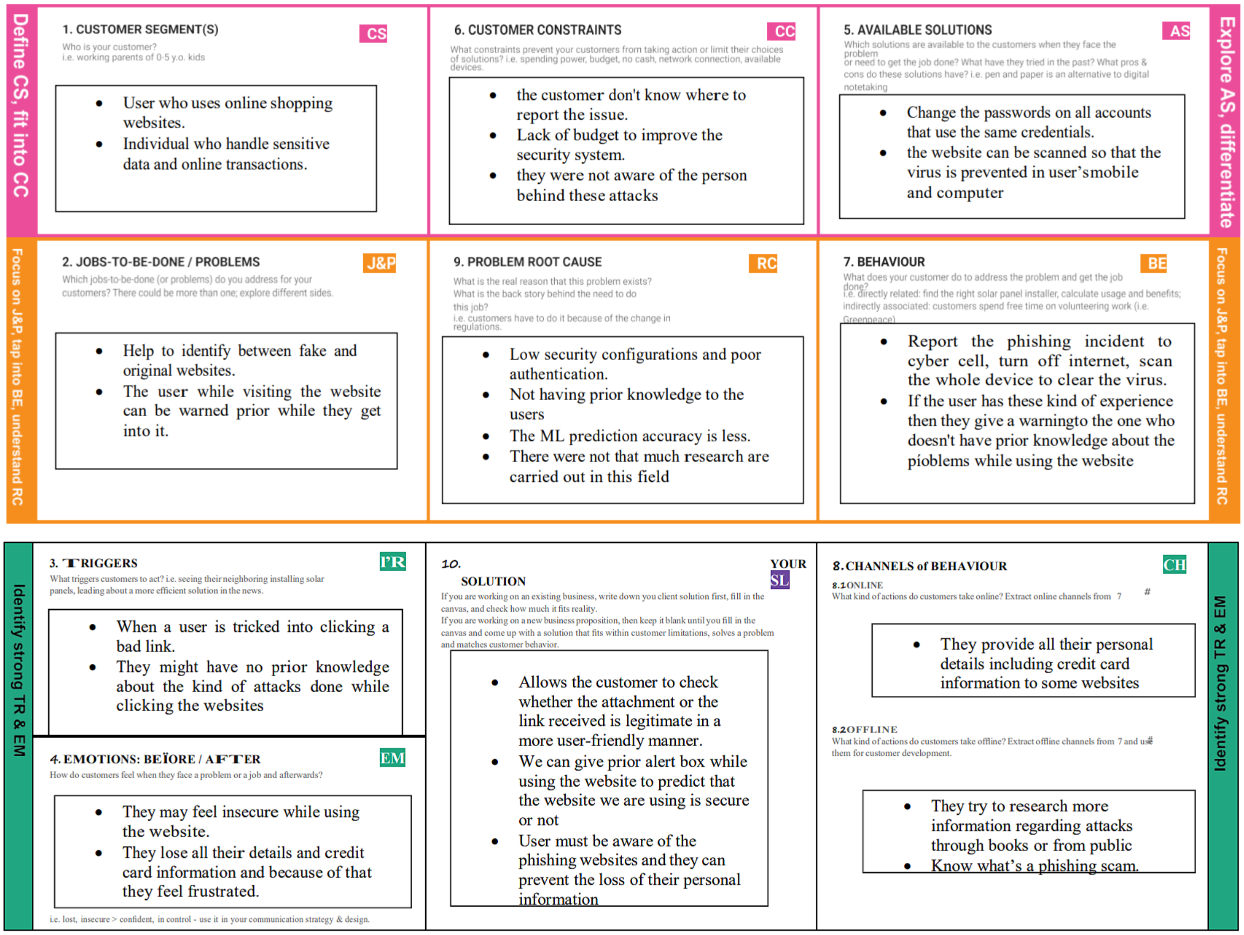
Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

3.3 PROPOSED SOLUTIONS

S.No	Parameter	Description
1.	Problem Statement (Problem to be solved)	A phishing website is a fake website, with domain similar in name and appearance to an official website. They're made in order to fool someone into believing it is legitimate. And then extract login credentials or confidential information such as credit card details from victims to perform malicious activities
2.	Idea/ Solution description	Using data visualization, machine learning algorithm, the user will find the Legitimate websites for their transaction from this project. Keep all systems current with the latest security patches and updates. We implemented classification algorithm and techniques to extract the phishing data sets criteria (URL and Domain Identity, security and encryption) to classify their legitimacy. Then detect whether the website is a phishing site or not
3.	Novelty/ Uniqueness	One of the major contributions of this project is the selection of different new features, which are capable enough to detect 0-h attacks, and these features do not depend on any third-party services. In particular, we extract character level Term Frequency Inverse Document Frequency (TF-IDF) features from noisy parts of HTML and plaintext of the given webpage
4.	Social Impact/ Customer Satisfaction	Since Data mining algorithm used in this system, it provides better performance as compared to other traditional classifications algorithms. With the help of this system user can also purchase products online without any hesitation. The accuracy of phishing site identification is around 89%.
5.	Business Model (Revenue Model)	A free web application system that can be used by many any person, E-commerce website or other websites in order to perform protected transaction. It's quick , free and trustworthy.
6.	Scalability of the Solution	Additional features such as reporting the phishing website can be made. Also this web application can be made as a chrome extension so that users can securely surf through websites without any hesitation.

3.4 PROBLEM SOLUTION FIT

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer's problem. When we enter the stage of Problem-Solution-Fit assessment we do not engage in testing some specific product ideas, — it is yet too early for a product. It's all about testing whether the kind of solution itself looks attractive. It's enough to describe the value proposition you came up within a few sentences without features and technical details. Keep it as simple as possible.


Define CS, fit into CC
Explore AS, differentiate
Focus on J&P, tap into BE, understand RC
Focus on J&P, tap into BE, understand RC
Identify strong TR & EM
Identify strong TR & EM

1. CUSTOMER SEGMENT(S) CS

Who is your customer?
i.e. working parents of 0-5 y.o. kids

- User who uses online shopping websites.
- Individual who handle sensitive data and online transactions.

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.

- the customer don't know where to report the issue.
- Lack of budget to improve the security system.
- they were not aware of the person behind these attacks

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem
or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking

- Change the passwords on all accounts that use the same credentials.
- the website can be scanned so that the virus is prevented in user'smobile and computer

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

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

- Help to identify between fake and original websites.
- The user while visiting the website can be warned prior while they get into it.

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists?
What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.

- Low security configurations and poor authentication.
- Not having prior knowledge to the users
- The ML prediction accuracy is less.
- There were not that much research are carried out in this field

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

- Report the phishing incident to cyber cell, turn off internet, scan the whole device to clear the virus.
- If the user has these kind of experience then they give a warningto the one who doesn't have prior knowledge about the píoblems while using the website

3. TRIGGERS TR

What triggers customers to act? i.e. seeing their neighboring installing solar panels, leading about a more efficient solution in the news.

- When a user is tricked into clicking a bad link.
- They might have no prior knowledge about the kind of attacks done while clicking the websites

4. EMOTIONS: BEÏORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?

- They may feel insecure while using the website.
- They lose all their details and credit card information and because of that they feel frustrated.

i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

10. YOUR SOLUTION SL

If you are working on an existing business, write down you client solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behavior.

- Allows the customer to check whether the attachment or the link received is legitimate in a more user-friendly manner.
- We can give prior alert box while using the website to predict that the website we are using is secure or not
- User must be aware of the phishing websites and they can prevent the loss of their personal information

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE
What kind of actions do customers take online? Extract online channels from 7 #

- They provide all their personal details including credit card information to some websites

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from 7 and use # them for customer development.

- They try to research more information regarding attacks through books or from public
- Know what's a phishing scam.

REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

Functional requirements are product features or functions that developers must implement to enable users to accomplish their tasks. So, it’s important to make them clear both for the development team and the stakeholders. Generally, functional requirements describe system behaviour under specific conditions. For example: The system sends an approval request after the user enters personal information. A search feature allows a user to hunt among various invoices if they want to credit an issued invoice. The system sends a confirmation email when a new user account is created.

FR No	**Functional Requirement (Epic)**	**Sub Requirement (Story / Sub-Task)**
FR-1	User Registration	Registration through Form.
FR-2	User Confirmation	Confirmation via Email.
FR-3	User Authentication	Authentication via Password.
FR-4	User Input	User input an URL to check it is legal or phishing site.
FR-5	Website Comparison	Model comparing the entered URL with the help of Blacklist and Whitelist.
FR-6	Feature extraction	After comparing, if none found on comparison the it extracts feature using heuristic and visual similarity approach.
FR-7	Prediction	Model Predicts the URL using Machine Learning algorithm such as Logistic Regression, KNN.
FR-8	Classifier	Model sends output to classifier and it produce final result.
FR-9	Announcement	Model the displays whether the website is a legal or phishing site.
FR-10	Events	Model needs the capability of retrieving and displaying accurate resultfor a website.

4.2 NON-FUNCTIONAL REQUIREMENT

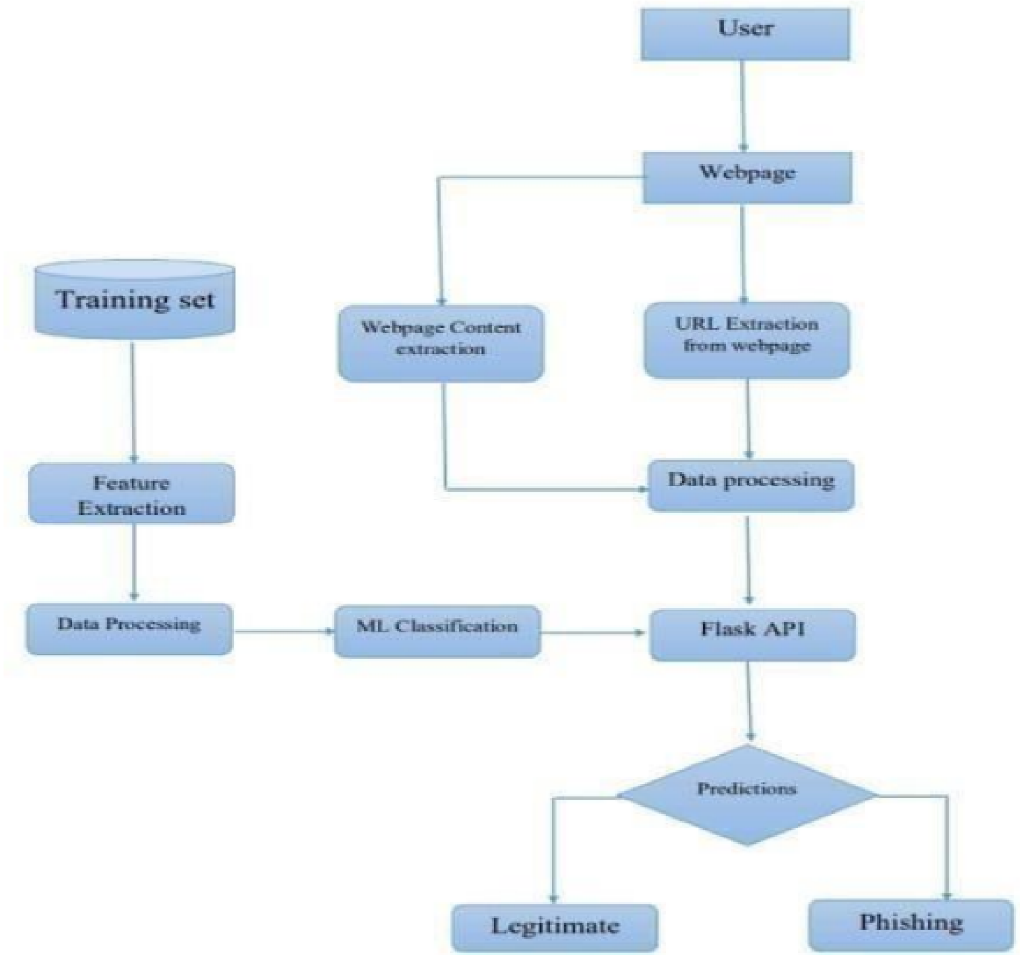
Non-functional Requirements (NFRs) define ‘how’ systems do what they do. This includes characteristics such as their performance, security, maintainability scalability, and ease of use. Essentially, they provide the proper checks and balances to the functional requirements. NFRs are essential. But they’re also a matter of balance and compromise. While you don’t want to forget about them until it’s too late, you also don’t want to focus too much on NFRs at the expense of your costs, schedule, and resources.

FR No	Non-Functional Requirement	Description
NFR-1	Usability	A set of specifications that describe the system’s operation capabilities and constraints and attempt to improve its functionality.
NFR-2	Security	Assuring all data inside the system or its part will be protected against malware attacks or unauthorized access.
NFR-3	Reliability	This approach gives more accuracy then existing system.
NFR-4	Performance	Performance Parameters for the proposed system gives accurate predicted value which is compared to the existing system.
NFR-5	Availability	The system is accessible by user at any time using web browser.
NFR-6	Scalability	The design will be suitable and performs with full efficiency according to rising demands.

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. A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination. Data flowcharts can range from simple, even hand-drawn process overviews, to in-depth, multi-level DFDs that dig progressively deeper into how the data is handled.

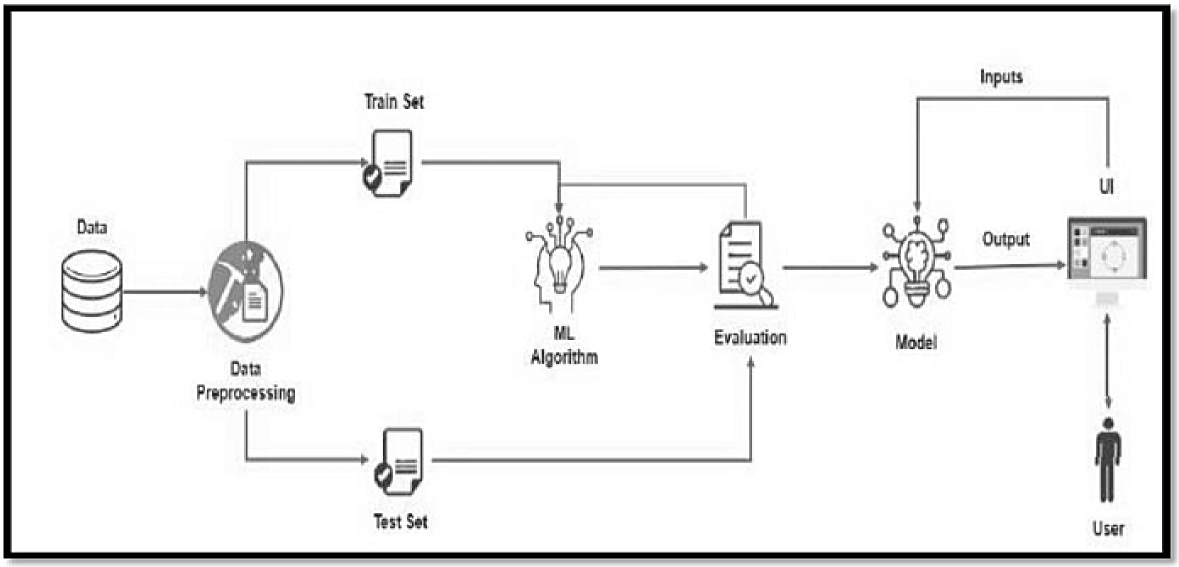


5.2 SOLUTIONS & TECHNICAL ARCHITECTURE

Solution architecture is comprised of several complex processes and sub-processes. It plays a central role in an organization’s efforts to introduce and successfully implement new technology solutions. In the first step, solution architecture specialists closely look at how the different elements of business, information, and technology can be applied to solve a specific problem. Next, they propose a combination of building blocks that provides the best possible fix. This process is very detail-oriented and serves as a connecting piece between enterprise architecture and technical architecture. After solution architects have designed a solution for an existing problem, it is their job to manage the tasks and activities that are involved with its successful implementation.

The technical architect’s main task is to realize particular technical implementation processes. As this requires a high level of in-depth expertise, technical architects usually specialize in one single technology. Technical architects can be in charge of leading large teams of developers and technical professionals. They act as technical project managers who define the structure of a specific system and oversee the related IT assignments. Out of all IT architects, they are the closest to an organization’s end-user. Thus, they have to ensure that the technology is not only delivered in a timely manner but fully functional for the end-user.

MODEL FOR WEB PHISHING DETECTION



5.3 USER STORIES

A user story is the smallest unit of work in an agile framework. It’s an end goal, not a feature, expressed from the software user’s perspective. A user story is an informal, general explanation of a software feature written from the perspective of the end user or customer. The purpose of a user story is to articulate how a piece of work will deliver a particular value back to the customer. Note that "customers" don't have to be external end users in the traditional sense, they can also be internal customers or colleagues within your organization who depend on your team. User stories are a few sentences in simple language that outline the desired outcome. They don't go into detail. Requirements are added later, once agreed upon by the team.

Sprint	Functioal Requirement (Epic)	User Story Number	User Story/ Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can registerfor the application by entering my Phone number/Gmail, Username password, and confirming my password.	2	High	Sagar M
Sprint-1	Registration	USN-2	As a user, I will receive confirmation in phone or Gmail once I have registered for the application	1	High	Sethu K
Sprint-1	Login	USN-3	As a user, I can log into the application by entering email & password	1	High	Sathish G
Sprint-1	Verification	USN-4	As a user, I can verify the registration and login of the user	1	High	Yuvan Jeyan G
Sprint-2	Dataset	USN-5	Collect number of datasets and get	2	Medium	Sagar M

	Collect		accuracy			
Sprint-2	Pre-processing	USN-5	The dataset is extracted	2	High	Sathish G
Sprint-2	Train the model	USN-6	Train the model.	4	High	Sethu K
Sprint-2	Test the model	USN-7	Test the model	6	High	Yuvan Jeyan G
Sprint-3	Detection	USN-8	Load the trained model.	3	High	Sagar M
Sprint-3	Detection	USN-9	Prediction of legitimate site	5	Medium	Sathish G
Sprint-3	Detection	USN-10	classify it by using a trained model to predict the output	8	High	Sethu K
Sprint-4	Detection	USN-11	Alerts the user about the legitimate site	7	High	Yuvan Jeyan G
Sprint-4	Detection	USN-12	As a User, I can detect the phished site	3	Medium	Sagar M
Sprint-4	Logout	USN-13	As a User, I can logout the application.	2	Low	Sathish G

PROJECT PLANNING & SCHEDULING

6.1 SPRINT PLANNING & ESTIMATION

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.

In scrum, the sprint is a set period of time where all the work is done. However, before you can leap into action you have to set up the sprint. You need to decide on how long the time box is going to be, the sprint goal, and where you're going to start. The sprint planning session kicks off the sprint by setting the agenda and focus. If done correctly, it also creates an environment where the team is motivated, challenged, and can be successful. Bad sprint plans can derail the team by setting unrealistic expectations.

- **The What** – The product owner describes the objective(or goal) of the sprint and what backlog items contribute to that goal. The scrum team decides what can be done in the coming sprint and what they will do during the sprint to make that happen.
- **The How** – The development team plans the work necessary to deliver the sprint goal. Ultimately, the resulting sprint plan is a negotiation between the development team and product owner based on value and effort.
- **The Who** – You cannot do sprint planning without the product owner or the development team. The product owner defines the goal based on the value that they seek. The development team needs to understand how they can or cannot deliver that goal. If either is missing from this event it makes planning the sprint almost impossible.
- **The Inputs** – A great starting point for the sprint plan is the product backlog as it provides a list of ‘stuff’ that could potentially be part of the current sprint. The team should also look at the existing work done in the increment and have a view to capacity
- **The Outputs** – The most important outcome for the sprint planning meeting is that the team can describe the goal of the sprint and how it will start working toward that goal. This is made visible in the sprint backlog.

Sprint planning requires some level of estimation. The team needs to define what can or cannot be done in the sprint: estimated effort vs capacity. Estimation is often confused with commitments. Estimates are by their very nature forecasts based on the knowledge at hand. Techniques such as story points or t-shirt sizing add value to the process by giving the team a different way of looking at the problem. They are not, however, magical tools that can find out the truth when there is none to be found. The more unknowns, the less likely the estimate will be correct.

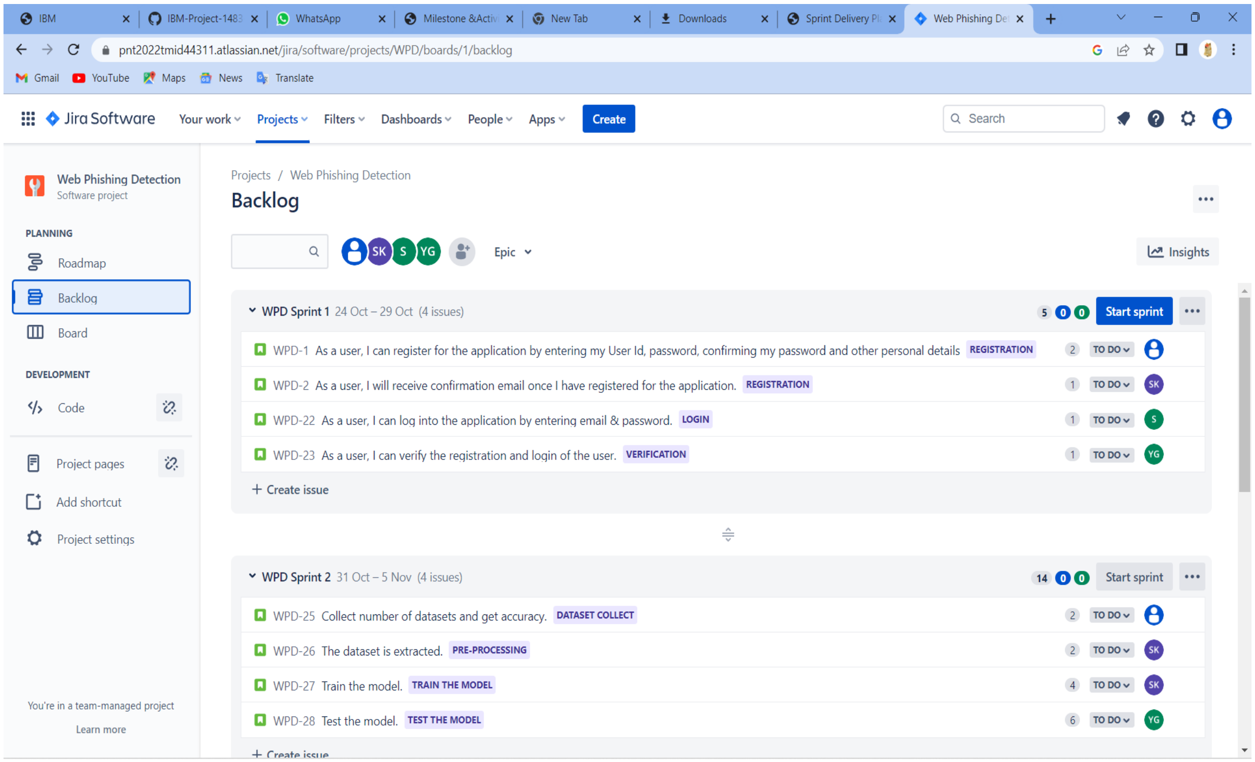
Good estimation requires a trust-based environment where information is given freely, and assumptions are discussed in the pursuit of learning and improvement. If estimates are used in a negative, confrontational way after the work is completed, then it’s likely that future estimates will be either be much bigger to ensure they never are wrong again or the time taken to create them will be much longer as the team second guesses itself worrying about the implications of getting them wrong.

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

6.3 REPORT FROM JIRA

Jira is a software platform for agile development and customer support. It's one of the most popular platforms for agile development — which makes it a go-to for many software development teams.



Projects / Web Phishing Detection

Backlog

WPD Sprint 1 24 Oct – 29 Oct (4 issues) 5 0 0 Start sprint

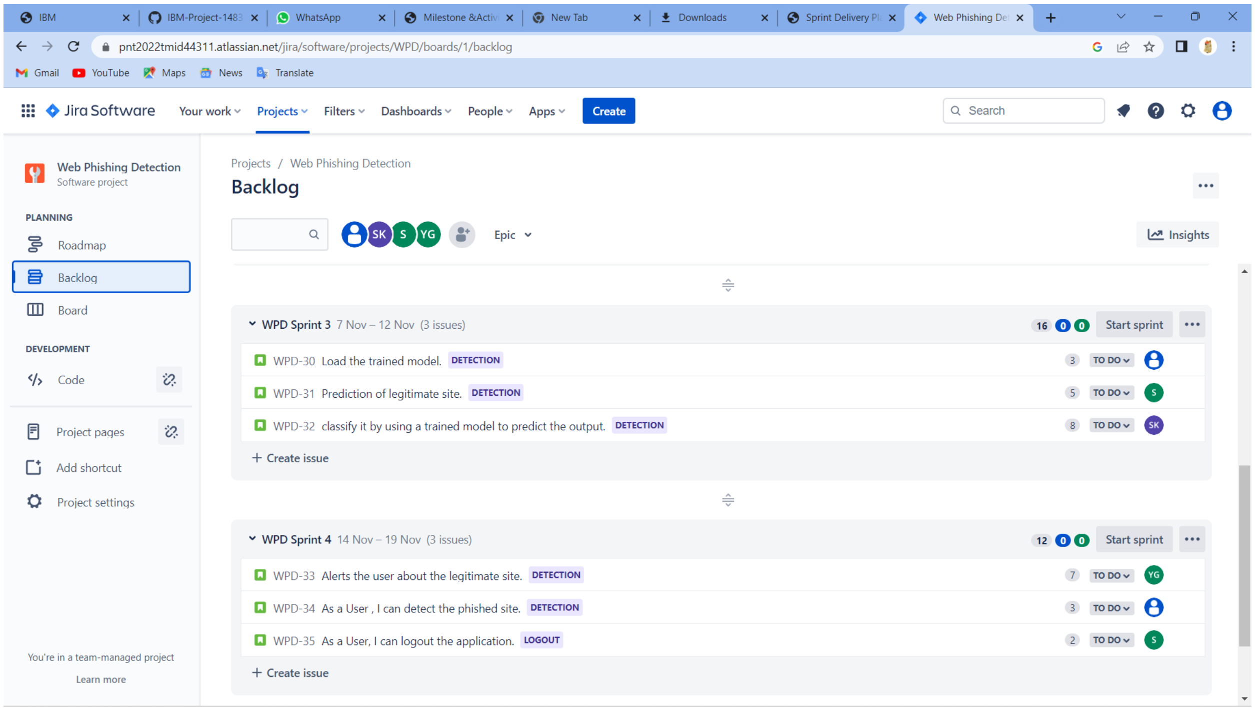
- WPD-1 As a user, I can register for the application by entering my User Id, password, confirming my password and other personal details REGISTRATION 2 TO DO
- WPD-2 As a user, I will receive confirmation email once I have registered for the application. REGISTRATION 1 TO DO SK
- WPD-22 As a user, I can log into the application by entering email & password. LOGIN 1 TO DO S
- WPD-23 As a user, I can verify the registration and login of the user. VERIFICATION 1 TO DO YG

+ Create issue

WPD Sprint 2 31 Oct – 5 Nov (4 issues) 14 0 0 Start sprint

- WPD-25 Collect number of datasets and get accuracy. DATASET COLLECT 2 TO DO
- WPD-26 The dataset is extracted. PRE-PROCESSING 2 TO DO SK
- WPD-27 Train the model. TRAIN THE MODEL 4 TO DO SK
- WPD-28 Test the model. TEST THE MODEL 6 TO DO YG

+ Create issue



Projects / Web Phishing Detection

Backlog

WPD Sprint 3 7 Nov – 12 Nov (3 issues) 16 0 0 Start sprint

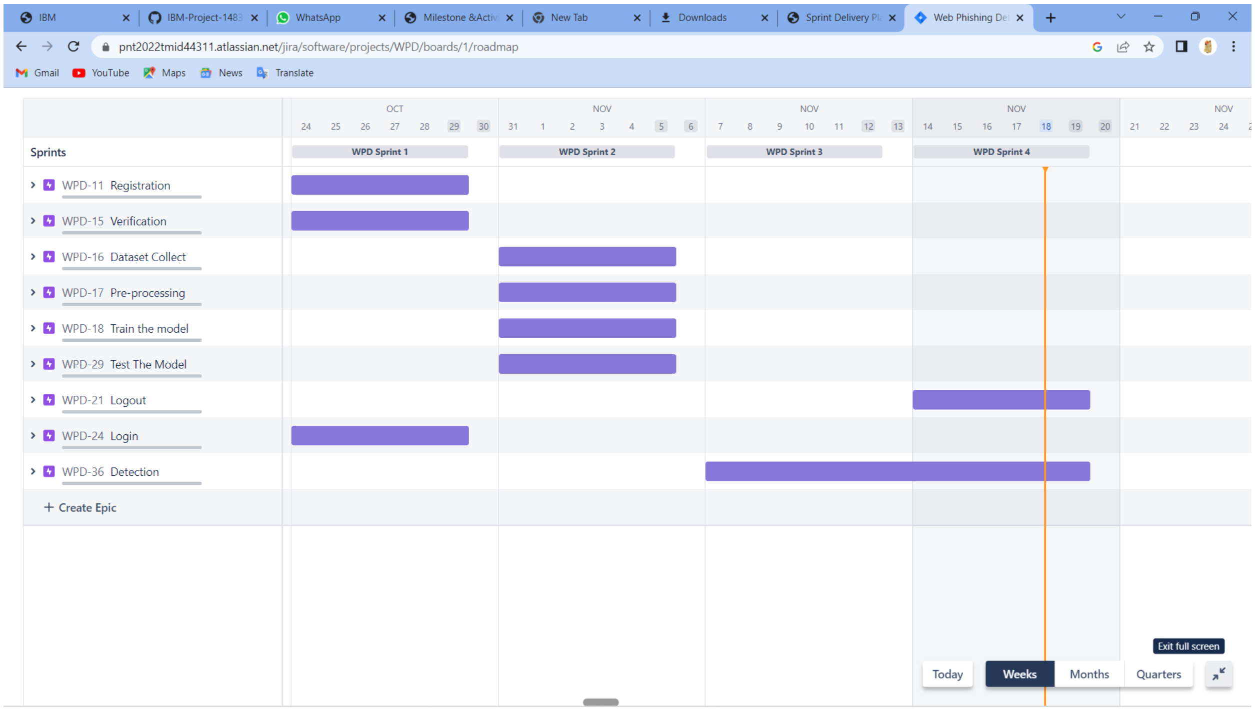
- WPD-30 Load the trained model. DETECTION 3 TO DO
- WPD-31 Prediction of legitimate site. DETECTION 5 TO DO S
- WPD-32 classify it by using a trained model to predict the output. DETECTION 8 TO DO SK

+ Create issue

WPD Sprint 4 14 Nov – 19 Nov (3 issues) 12 0 0 Start sprint

- WPD-33 Alerts the user about the legitimate site. DETECTION 7 TO DO YG
- WPD-34 As a User , I can detect the phished site. DETECTION 3 TO DO
- WPD-35 As a User, I can logout the application. LOGOUT 2 TO DO S

+ Create issue



Sprints

- WPD-11 Registration
- WPD-15 Verification
- WPD-16 Dataset Collect
- WPD-17 Pre-processing
- WPD-18 Train the model
- WPD-29 Test The Model
- WPD-21 Logout
- WPD-24 Login
- WPD-36 Detection

+ Create Epic

CODING & SOLUTIONING

7.1 FEATURE 1

home.html

```
<!doctype html>
<html lang="en">
<head>
  <link rel="stylesheet" type="text/css"
href="{{url_for('static',filename='css/style.css')}}">
  <meta charset="utf-8">
  <meta name="viewport" content="width=device-width, initial-scale=1">
  <title>Home</title>

  <style>



    </style>
<link href="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-Zenh87qX5JnK2Jl0vWa8Ck2rdkQ2Bzep5IDxbcnCeuOxjzrPF/et3URy9Bv1WTRi" crossorigin="anonymous">
</head>


<body class="bg-co"  >
  <div class="bg-nav text-light d-flex flex-column flex-md-row align-items-center pb-3 mb-4 border-bottom">
    <h5 class="my-0 mr-md-auto font-weight-bold mt-3" style="font-size:20px;opacity: 0.5; font-family: Georgia, serif; font-weight: bold;  padding-left: 50px;">URL Prediction</h5>
    <nav class="d-inline-flex mt-2 mt-md-0 ms-md-auto ">
      <a style="font-family: Georgia, serif;font-weight: bold;margin-right:20px; color:black;"  href="/predicturl">Predict URL</a>
      <a style="font-family: Georgia, serif;font-weight: bold;margin-right:20px; color:black;" href="/addurl">Add url</a>
      <a style="font-family: Georgia, serif;font-weight: bold; margin-right:20px; color:black" href="/about">About</a>
    </nav>
 </div>
 <div class="container bg-co">
  <div class="row">
      <div class= "col-md-6">
          <h1 style="font-size:60px; font-weight: bold;font-family: Georgia, serif; color:black;">
          THE MOST RELAIABLE WAY TO PREDICT THE FUTURE IS TO CREATE IT
          </h1>
          <b><h5 style="text-align: right;font-family: Georgia, serif;color:black; font-weight: bold;color: rgba(0, 0, 0, 0.705);">- ABRAHAM LINCOLN</h5></b>
      </div>

      <div class="col-md-6 " >

      <div  style="margin-top: 120px;margin-left: 190px;">

              <form class="form"  action="/predicturl">

                <center><img src="{{url_for('static', filename='Code.jpg')}}"
style="height: 40%; width: 40%; opacity: 0.5; " class="img img-responsive img-circle  mx-
auto d-block" /><br>

                  <input type="submit" style="background-color: rgba(0, 0, 0,
0.801);color: white; font-weight: bolder;"class="btn  btn-lg btn-block mx-auto "
value="PREDICT YOUR URL">

                </center>

              </form>

          </div>

          </div>

      </div>

  </div>

</div>


    <script
src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.6/dist/umd/popper.min.js"
integrity="sha384-oBqDVmMz9ATKxIep9tiCxS/Z9fNfEXiDAYTujMAeBAsjFuCZSmKbSSUnQlmh/jp3"
crossorigin="anonymous"></script>

    <script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.min.js"
integrity="sha384-IDwe1+LCz02ROU9k972gdyvl+AESN10+x7tBKgc9I5HFtuNz0wWnPclzo6p9vxnk"
crossorigin="anonymous"></script>

    <script
src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.bundle.min.js"
integrity="sha384-OERcA2EqjJCMA+/3y+gxIOqMEjwtxJY7qPCqsdltbNJuaOe923+mo//f6V8Qbsw3"
crossorigin="anonymous"></script>

  </body>

</html>
```

EXPLANATION

URL is the first thing to analyse a website to decide whether it is a phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

- Digit count in the URL
- Total length of URL
- Checking whether the URL is Typo squatted or not. (google.com → goggle.com)
- Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)
- Number of subdomains in URL
- Is Top Level Domain (TLD) one of the commonly used one?

7.2 FEATURE 2

predict1.html

```
<!doctype html>
<html lang="en">
<head>
   <link rel="stylesheet" type="text/css" href= "{{url_for('static',filename='css/style.css')}}">
  <meta charset="utf-8">
  <meta name="viewport" content="width=device-width, initial-scale=1">
  <title>URL Prediction</title>
<script>
function clearInput() {
  document.getElementById("Form").reset();
}
</script>
<link href="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-Zenh87qX5JnK2Jl0vWa8Ck2rdkQ2Bzep5IDxbcnCeuOxjzrPF/et3URy9Bv1WTRi" crossorigin="anonymous">
</head>
<body class="bg-co">
  <div class="bg-nav text-light d-flex flex-column flex-md-row align-items-center pb-3 mb-4 border-bottom">
           <h5 class="my-0 mr-md-auto font-weight-bold mt-3" style="font-size:20px;opacity: 0.5; font-family: Georgia, serif; font-weight: bold; padding-left: 50px;">URL Prediction</h5>
    <nav class="d-inline-flex mt-2 mt-md-0 ms-md-auto ">
         <a class="me-3 py-2 text-light text-decoration-none mt-3" style="font-family: Georgia, serif;font-weight: bold;margin-right:20px; " href="/">Home</a>
         <a class="me-3 py-2 text-light text-decoration-none mt-3" style="font-
```

```
family: Georgia, serif;font-weight: bold;margin-right:20px; " href="/addurl">Add url</a>
        <a class="py-2 text-light text-decoration-none mt-3" style="font-family: Georgia, serif;font-weight: bold; margin-right:20px;" href="/about">About</a>
    </nav>
 </div>
  <div class="bg-co" style="margin-top: 120px;  margin-left: 400px;margin-right: 400px;">
   <div class="card-body">
      <form id="Form" action="/predict" method='post'class="form">
             <img src="{{url_for('static', filename='code.png')}}" style="height: 20%; width: 20%;" class="img img-responsive img-circle  mx-auto d-block" /><br>
          <label><b>Enter URL to predict</b></label><br>
                  <input type="text" name="url" id= "myText" placeholder="Ex : https://abcde.com/" class="form-control" required><br>
               <div class="w3-bar "><center>
                         <input type="submit"  style="background-color: black; font-weight: bold; color: white;" class="btn "  value="Predict URL" >
                         <input type= "button" style="background-color: black; font-weight: bold; color: white;" class=" btn " value= "Clear" onclick= "clearInput()"></center>
              </div>
      </form>
    </div>
    </div>
<script src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.6/dist/umd/popper.min.js" integrity="sha384-oBqDVmMz9ATKxIep9tiCxS/Z9fNfEXiDAYTujMAeBAsjFuCZSmKbSSUnQlmh/jp3" crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.min.js" integrity="sha384-IDwe1+LCz02ROU9k972gdyvl+AESN10+x7tBKgc9I5HFtuNz0wWnPclzo6p9vxnk" crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.2/dist/js/bootstrap.bundle.min.js" integrity="sha384-OERcA2EqjJCMA+/3y+gxIOqMEjwtxJY7qPCqsdltbNJuaOe923+mo//f6V8Qbsw3" crossorigin="anonymous"></script>
  </body>
</html>
```

EXPLANATION

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how much reliable a web site is. Some of Page-Based Features are given below.

- Global PageRank
- Country PageRank
- Position at the Alexa Top 1 Million Site
- Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.
- Estimated Number of Visits for the domain on a daily, weekly, or monthly basis
- Average Pageviews per visit
- Average Visit Duration
- Web traffic share per country
- Count of reference from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

TESTING

8.1 TEST CASE

Test Scenario	Expected Result
Verify user is able to enter the URL in the form	Result of classification will be displayed
Verify the UI elements in the form	Application should show below UI elements: a.input form box c.submit button d.services offered e.team
Verify user is able to see an alert when nothing is entered in the textbox	Alert of incomplete input
Verify user is able to see the result when URL is entered in the textbox	Result of classification will be displayed
Verify user is able to enter their name, email and query message in the form	Details are stored in the database

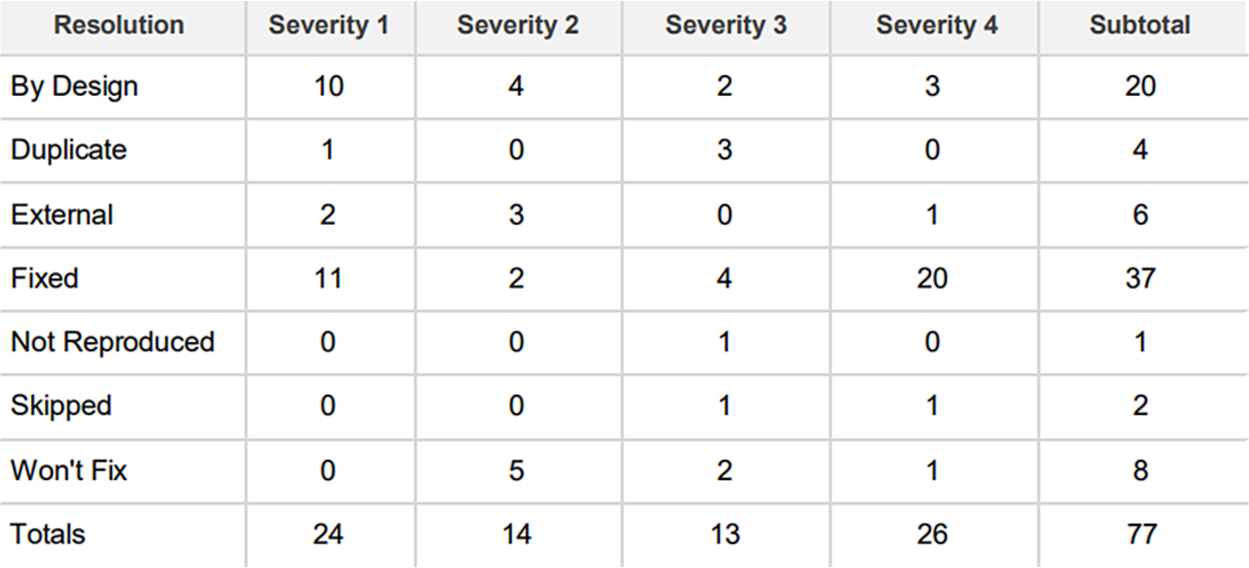
8.2 USER ACCEPTANCE TESTING

1. Purpose of Document

The purpose of this documentis to briefly explain the test coverageand open issues ofthe Web PhishingDetection project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis

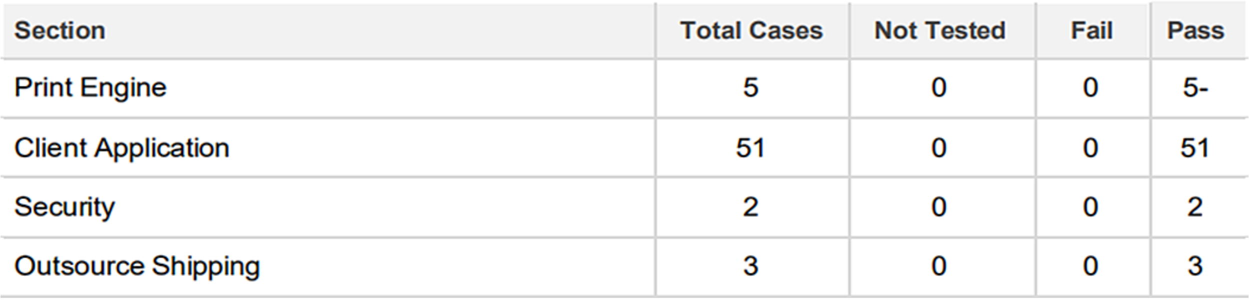
This reportshows the numberof resolved or closed bugs at each severity level,and how they wereresolvedalysis



Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

3. Test Case Analysis

This report shows the number of test cases that have passed, failed,and untested

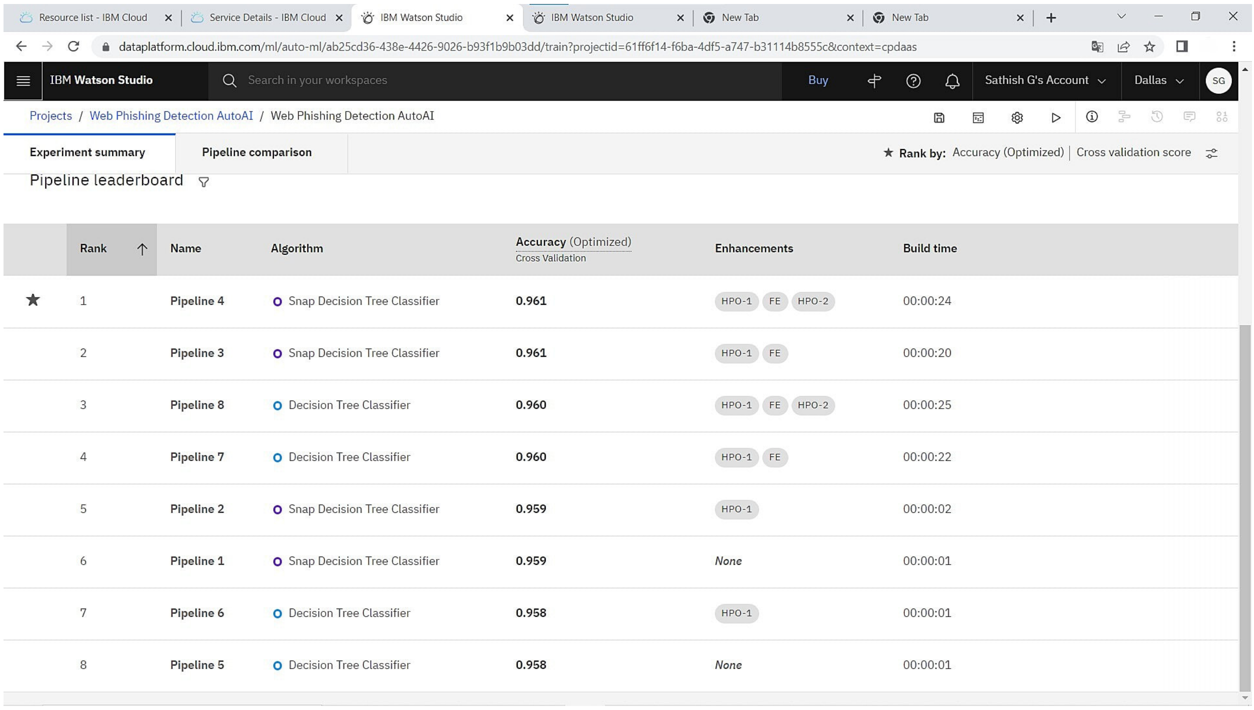


Section	Total Cases	Not Tested	Fail	Pass
Print Engine	5	0	0	5-
Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3

Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

RESULTS

9.1 PERFORMANCE METRICS


Resource list - IBM Cloud
Service Details - IBM Cloud
IBM Watson Studio
IBM Watson Studio
New Tab
New Tab
dataplatform.cloud.ibm.com/ml/auto-ml/ab25cd36-438e-4426-9026-b93f1b9b03dd/train?projectid=61ff6f14-f6ba-4df5-a747-b31114b8555c&context=cpdaas
IBM Watson Studio
Search in your workspaces
Buy
Sathish G's Account
Dallas
SG
Projects / Web Phishing Detection AutoAI / Web Phishing Detection AutoAI
Experiment summary
Pipeline comparison

★ Rank by: Accuracy (Optimized) | Cross validation score

Pipeline leaderboard

	Rank	Name	Algorithm	Accuracy (Optimized) Cross Validation	Enhancements	Build time
★	1	Pipeline 4	Snap Decision Tree Classifier	0.961	HPO-1 FE HPO-2	00:00:24
	2	Pipeline 3	Snap Decision Tree Classifier	0.961	HPO-1 FE	00:00:20
	3	Pipeline 8	Decision Tree Classifier	0.960	HPO-1 FE HPO-2	00:00:25
	4	Pipeline 7	Decision Tree Classifier	0.960	HPO-1 FE	00:00:22
	5	Pipeline 2	Snap Decision Tree Classifier	0.959	HPO-1	00:00:02
	6	Pipeline 1	Snap Decision Tree Classifier	0.959	*None*	00:00:01
	7	Pipeline 6	Decision Tree Classifier	0.958	HPO-1	00:00:01
	8	Pipeline 5	Decision Tree Classifier	0.958	*None*	00:00:01

ADVANTAGE & DISADVANTAGE

ADVANTAGES :

- Measure the degrees of corporate and employee vulnerability
- Eliminate the cyber threat risk level
- Increase user alertness to phishing risks
- Instill a cyber security culture and create cyber security heroes
- Change behavior to eliminate the automatic trust response
- Deploy targeted anti-phishing solutions
- Protect valuable corporate and personal data
- Meet industry compliance obligations
- Assess the impacts of cyber security awareness training
- Segment phishing simulation

DISADVANTAGES:

- Loss of data
- Damaged Reputation
- Direct Monetry Loss
- Loss Of Productivity
- Loss of Customers
- Financial Penalties

CONCLUSION

Phishing detection is now an area of great interest among the researchers due to its significance in protecting privacy and providing security. There are many methods that perform phishing detection by classification of websites using trained machine learning models. URL based analysis increases the speed of detection. Furthermore, by applying feature selection algorithms and dimensionality reduction techniques, we can reduce the number of features and remove irrelevant data. There are many machine learning algorithms that perform classification with good performance measures. In this paper, we have done a study of the process of phishing detection and the phishing detection schemes in the recent research literature. This will serve as a guide for new researchers to understand the process and to develop more accurate phishing detection systems.

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.

The widely publicized Gmail phishing scam earlier this year is just one example of a modern threat that affected users on a large scale. In this case, users were sent an email that appeared legitimate and directed them to an actual Google page. While most phishing scams rely on pushing users to a malicious domain, this particular attack simply led unsuspecting individuals to granting broad permissions to a malicious application. Hackers could then see victims' contacts, read their emails, have insight into the users' locations, and see files created in G Suite.

The Gmail phishing attack shows us just how advanced these techniques have become – it was difficult to detect and difficult to prevent. A critical takeaway is that the attack was able to clear the psychological trust hurdle. Users were tricked into giving permissions to a third party application because they trusted it; they believed the application to be a Google-approved service. A minute change in how the application domain was disguised successfully convinced users that the application was trustworthy.

This is the future of phishing. The ability to spoof cloud apps while masking the true identity of the sender in order to steal personal information – an alarming trend given the rapid increase of cloud adoption in verticals around the world.

Traditional Phishing

Traditional phishing was rather simplistic in execution and relied on the user's lack of knowledge. For example, social engineering driven by phone calls and emails wherein malicious actors would pose as government agents or corporate customer service representatives. Many targets of these attacks – elderly and young internet users alike – would readily provide any and all information to avoid the threat of legal action, penalties, and account shutdowns.

There are two key reasons why traditional attacks have become less effective: advances in detection and the increase in awareness among the average user. Major email providers, for example, alert users when a message is deemed suspicious or the source domain is not as it seems. What's more, users are more computer savvy than ever and know not to trust inbound messages that request personal information. That said, while each major breach prompts reaction from users to update their privacy settings and login credentials, targeted attacks are and will remain relentless.

APPENDIX

SOURCE CODE

app.py

```python
from flask import Flask, request, render_template,session, url_for,redirect,flash
import pickle
import inputScript
from passlib.hash import  pbkdf2_sha256
import json
app = Flask(__name__,template_folder='templates')
model = pickle.load(open('Website_dt.pkl','rb'))

@app.route("/")
def helloworld():
    return render_template("/home.html")

@app.route("/predicturl")
def predicturl():
    return render_template("/predict1.html")

@app.route("/predict" ,methods=["POST","GET"] )
def predict():
    if request.method == 'POST':
        url = request.form['url']
        checkprediction = inputScript.main(url)
        print(url)
        print(checkprediction)
        prediction = model.predict(checkprediction)
        print(prediction)
        output=prediction[0]
        if output==1 :
            return render_template("/output1.html")
        elif output==-1 :
            return render_template("/output.html")

@app.route("/project_details")
def support():
    return render_template("/project_details.html")

@app.route("/addurl")
def addurl():
    return render_template("/addurl.html")

@app.route("/about")
def about():
    return render_template("/about.html")


if __name__ =="__main__":
    app.run(debug=True,host='0.0.0.0',port=2000)

```

inputScript.py

```
import regex
from tldextract import extract
import ssl
import socket
from bs4 import BeautifulSoup
import urllib.request
import whois
import datetime


def url_having_ip(url):
    symbol = regex.findall(r'(http((s)?)://)((((\d)+).)*)((\w)+)(/((\w)+))?',url)
    if(len(symbol)!=0):
        having_ip = 1 #phishing
    else:
        having_ip = -1 #legitimate
    return(having_ip)
    return 0


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


def url_short(url):
    #ongoing
    return 0
```

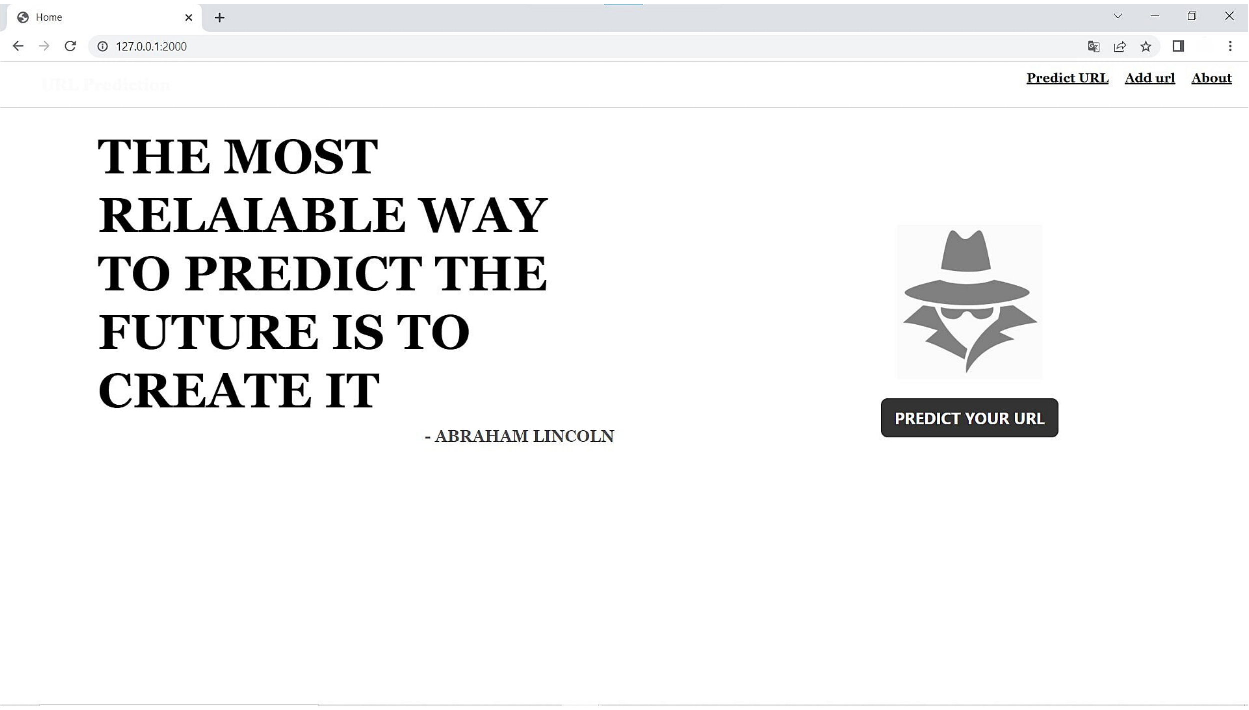
```

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

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

def sub_domain(url):
    subDomain, domain, suffix = extract(url)
    if(subDomain.count('.')==0):
        return -1
    elif(subDomain.count('.')==1):
        return 0
    else:
        return 1

def SSLfinal_State(url):
    try:
#check wheather contains https
        if(regex.search('^https',url)):
            usehttps = 1
        else:
            usehttps = 0
#getting the certificate issuer to later compare with trusted issuer
        #getting host name
        subDomain, domain, suffix = extract(url)
        host_name = domain + "." + suffix
        context = ssl.create_default_context()
        sct = context.wrap_socket(socket.socket(), server_hostname = host_name)
        sct.connect((host_name, 443))
        certificate = sct.getpeercert()
        issuer = dict(x[0] for x in certificate['issuer'])
        certificate_Auth = str(issuer['commonName'])
        certificate_Auth = certificate_Auth.split()
        if(certificate_Auth[0] == "Network" or certificate_Auth == "Deutsche"):
            certificate_Auth = certificate_Auth[0] + " " + certificate_Auth[1]
        else:
            certificate_Auth = certificate_Auth[0]
        trusted_Auth = ['Comodo','Symantec','GoDaddy','GlobalSign','DigiCert','StartCom','Entrust','Verizon','Trustwave','Unizeto','Buypass','QuoVadis','Deutsche Telekom','Network Solutions','SwissSign','IdenTrust','Secom','TWCA','GeoTrust','Thawte','Doster','VeriSign']
#getting age of certificate
        startingDate = str(certificate['notBefore'])
        endingDate = str(certificate['notAfter'])
        startingYear = int(startingDate.split()[3])
        endingYear = int(endingDate.split()[3])
        Age_of_certificate = endingYear-startingYear
#checking final conditions
        if((usehttps==1) and (certificate_Auth in trusted_Auth) and (Age_of_certificate>=1) ):
            return -1 #legitimate
        elif((usehttps==1) and (certificate_Auth not in trusted_Auth)):
            return 0 #suspicious
        else:
            return 1 #phishing
    except Exception as e:
        return 1

def domain_registration(url):
    try:
        w = whois.whois(url)
        updated = w.updated_date
        exp = w.expiration_date
        length = (exp[0]-updated[0]).days
        if(length<=365):
            return 1
        else:
            return -1
    except:
        return 0

def favicon(url):
```

```python
    #ongoing
    return 0

def port(url):
    #ongoing
    return 0

def https_token(url):
    subDomain, domain, suffix = extract(url)
    host =subDomain +'.' + domain + '.' + suffix
    if(host.count('https')): #attacker can trick by putting https in domain part
        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
        elif(0.22<=avg<=0.61):
            return 0
        else:
            return 1
    except:
        return 0


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):
    #ongoing
    return 0

```

```python
def email_submit(url):
    try:
        opener = urllib.request.urlopen(url).read()
        soup = BeautifulSoup(opener, 'lxml')
        if(soup.find('mailto:')):
            return 1
        else:
            return -1
    except:
        return 0

def abnormal_url(url):
    #ongoing
    return 0

def redirect(url):
    #ongoing
    return 0

def on_mouseover(url):
    #ongoing
    return 0

def rightClick(url):
    #ongoing
    return 0

def popup(url):
    #ongoing
    return 0

def iframe(url):
    #ongoing
    return 0

def age_of_domain(url):
    try:
        w = whois.whois(url)
        start_date = w.creation_date
        current_date = datetime.datetime.now()
        age =(current_date-start_date[0]).days
        if(age>=180):
            return -1
        else:
            return 1
    except Exception as e:
        print(e)
        return 0
def dns(url):
    #ongoing
    return 0

def web_traffic(url):
    #ongoing
    return 0

```

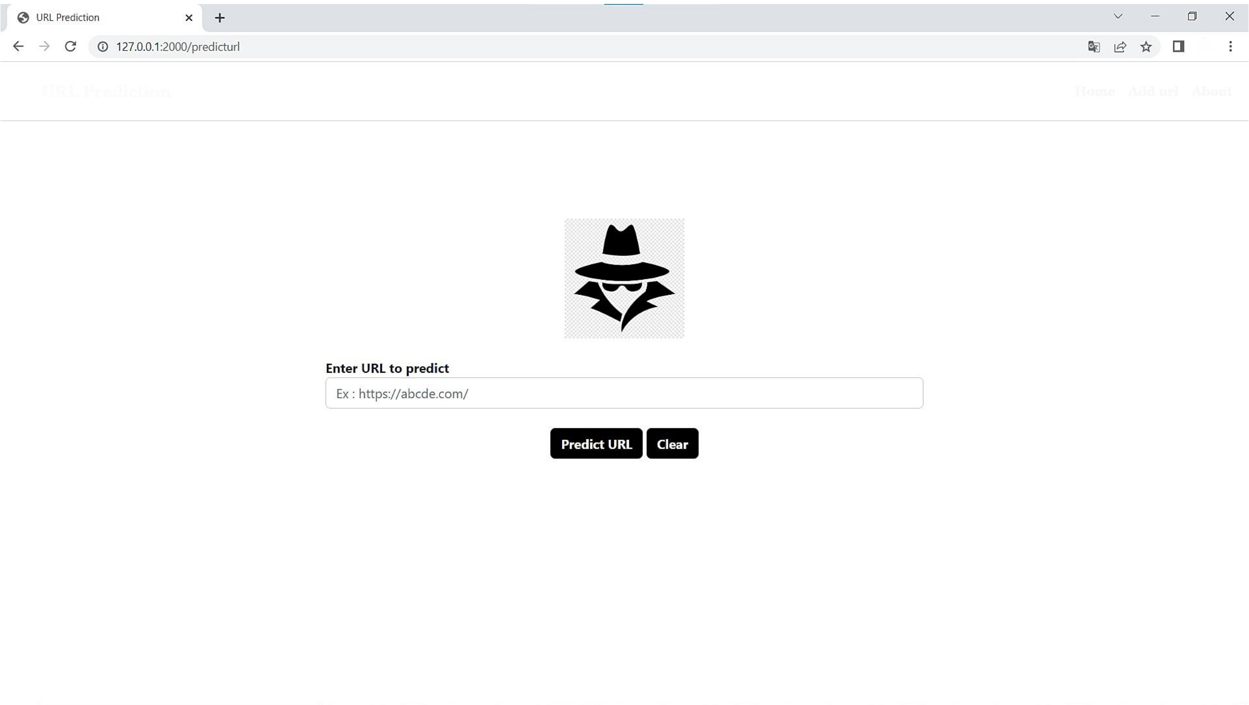
```
def page_rank(url):
    #ongoing
    return 0

def google_index(url):
    #ongoing
    return 0


def links_pointing(url):
    #ongoing
    return 0

def statistical(url):
    #ongoing
    return 0

def main(url):


    check = [[url_having_ip(url),url_length(url),url_short(url),having_at_symbol(url),
    doubleSlash(url),prefix_suffix(url),sub_domain(url),SSLfinal_State(url),
    domain_registration(url),favicon(url),port(url),https_token(url),request_url(url),
    url_of_anchor(url),Links_in_tags(url),sfh(url),email_submit(url),abnormal_url(url),
redirect(url),on_mouseover(url),rightClick(url),popup(url),iframe(url),
age_of_domain(url),dns(url),web_traffic(url),page_rank(url),google_index(url),
              links_pointing(url),statistical(url)]]
    return check
```

SCREENSHOTS



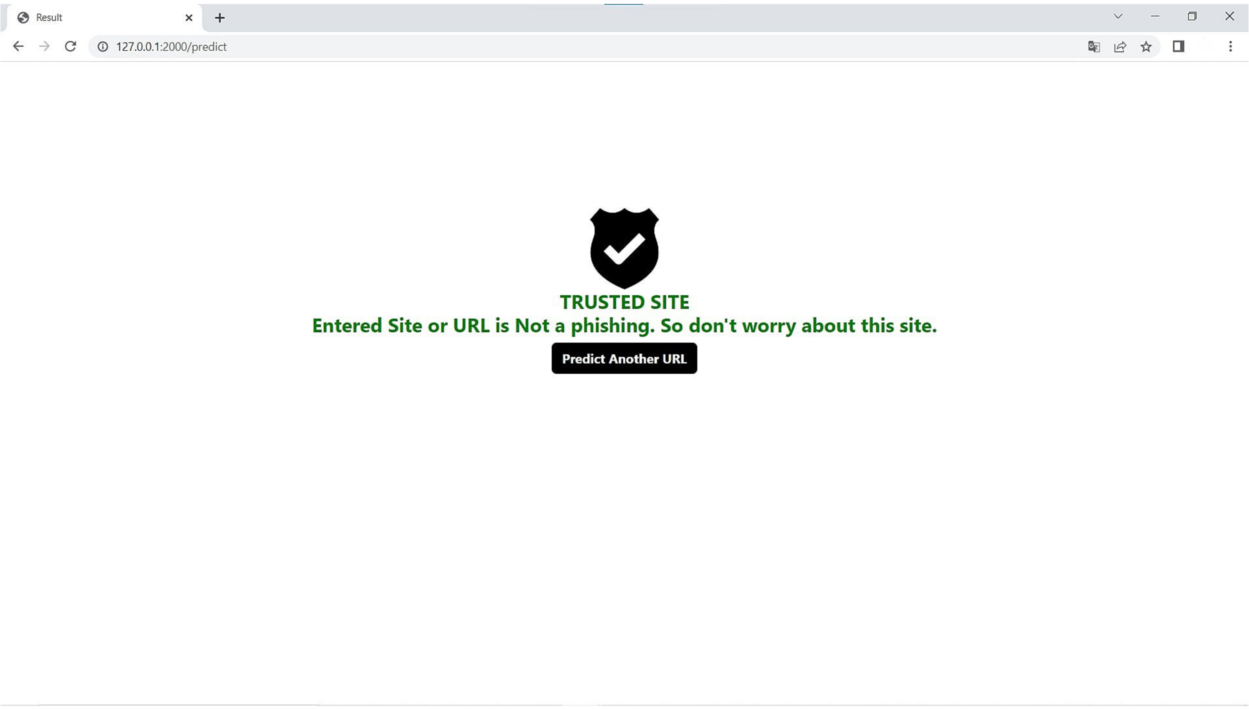
THE MOST RELAIABLE WAY TO PREDICT THE FUTURE IS TO CREATE IT

- ABRAHAM LINCOLN



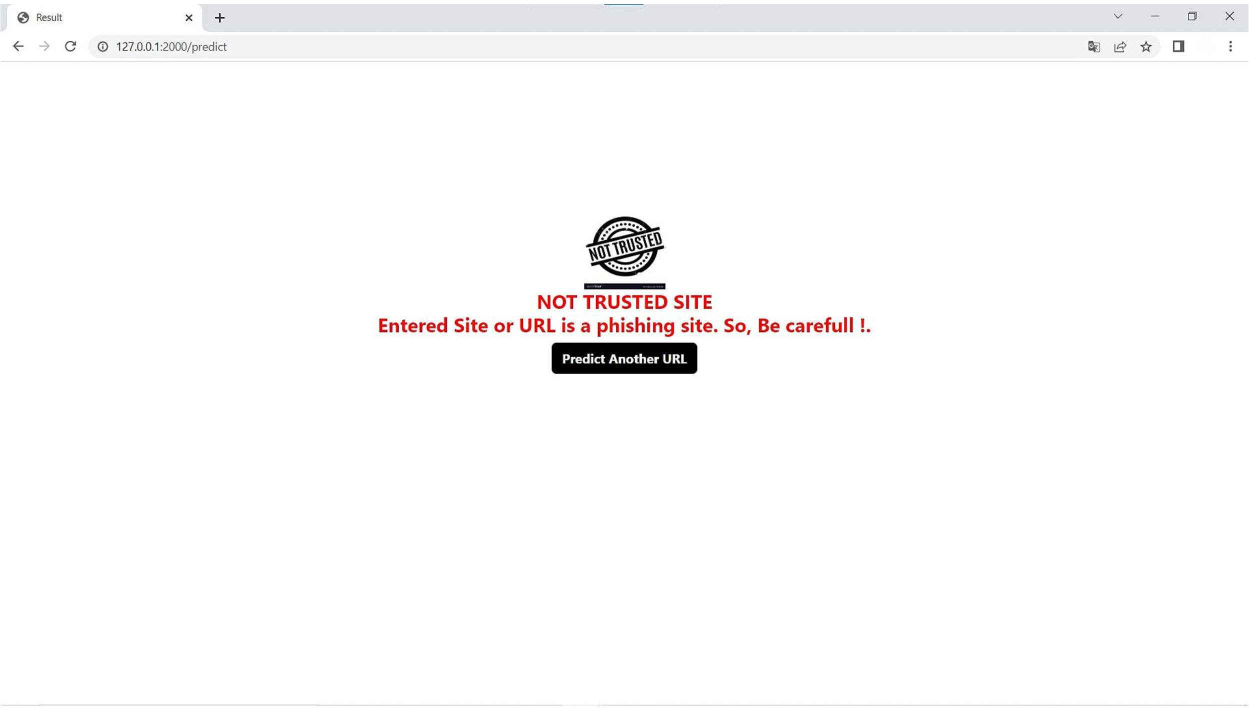
Enter URL to predict

Ex : https://abcde.com/



TRUSTED SITE

Entered Site or URL is Not a phishing. So don't worry about this site.



NOT TRUSTED SITE

Entered Site or URL is a phishing site. So, Be carefull !.

GIT HUB & PROJECT DEMO LINK

- ➤ https://github.com/IBM-EPBL/IBM-Project-1483-1658390493