IBM PROJECT REPORT

WEB PHISHING DETECTION

TEAM ID	PNT2022TMID03803
TEAM LEAD	ARUNERAJ J
TEAM MEMBER 1	PREM KUMAR A R
TEAM MEMBER 2	SANTHOSH S
TEAM MEMBER 3	SYED ABDUL AZEEM A

1. INTRODUCTION

Nowadays Phishing becomes the main area of concern for security researchers because it is not difficult to create a fake website that looks so close to a legitimate website. Experts can identify fake websites but not all users can identify fake websites and such users become the victim of phishing attacks. The main aim of the attacker is to steal banks' account credentials. Phishing attacks are becoming successful because of a 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 website act as a sort of real memory that the last word objective is to accumulate unstable info, for example, passwords, account focal points, or MasterCard numbers. The phishing activity in early 2016 was the highest ever recorded since it began monitoring in 2004. The total number of phishing attacks in 2016 was 1,220,523. This was a 65 percent increase over 2015. In the fourth quarter of 2004, there were 1,609 phishing attacks per month. In the fourth quarter of 2016, there was an average of 92,564 phishing attacks per month, an increase of 5,753% over 12 years According to the 3rd Microsoft Computing Safer Index Report released in February 2014, the annual worldwide impact of phishing could be as high as $5 billion.

With the prevalence of networks, phishing has become one of the most serious security threats in modern society, thus making detecting and defending against web phishing an urgent and essential research task. Web phishing detection is crucial for both private users and enterprises. This means that their square measure some contrary to phishing programming and techniques for recognizing potential phishing tries in messages and characteristic phishing substances on locales, phishes think about new and crossbreed procedures to bypass the open programming and frameworks. Some possible solutions to combat phishing were created, including specific legislation and technologies. From a technical point of view, list and white list, detection based on Uniform Resource Locator (URL) features detection based on web content and detection based on machine learning. The anti-phishing way using a blacklist may be an easy way, but it cannot find new phishing websites.

The detection of URLs is to analyze the features of URL. The URL of phishing websites may Wireless Communications and Mobile Computing 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. Phishing may be a fraud framework that uses a mixture of social designing which is an additional, advancement to sensitive and personal data, for 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 from true blue sources like money-connected institutions, online business goals, etc. to draw in customers to go to phony destinations through joins given within the phishing websites.

1.1 Project Overview

Phishing is a common attack on credulous people by making them disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin personal information like user names, passwords, and online banking transactions. Phishers use websites that are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing Mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning. Phishing is popular among attackers since it is easier to trick someone into clicking a malicious link that seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate content. Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques.

1.2 Purpose

In the effort of developing the proposed system, a project methodology has to be selected and defined, as to generate a practicable development environment and realistic schedule. Thus, this project will be done using the Agile Unified Process (AUP) Lifecycle for its abridged development period and flexible process A baseline of hardware and software requirements is set. This is to ensure the operating system platform is capable to handle and perform the development of the system. The software that is used to develop the system is using Microsoft Visual Studio 2010 Ultimate. Microsoft Visual Studio 2010 Ultimate generates the system C# was the most appropriate language to run the program. MySQL stocks up data and implements database in this system.

MySQL builds in the database in the Microsoft Visual Studio 2010 Ultimate. In hardware, at least 4GB RAM is required in the laptop/PC to build the system. This ensures a smooth process during the development. Admin can filter which URLs are blacklisted and which are not blacklisted by copying and pasting the URLs at the "Site" row. Admin can classify blacklisted URLs as 1 and not blacklisted URL as 0. Admin can also edit, update and delete the sites once the URLs have been added. As per the overall discussion, the proposed project has been successfully coded and developed and has met the expectations made during the project proposal phase. From the various results aggregated during the various tests, the system can be concluded that it has been successfully developed to its specifications. The system interface has been successfully designed, the functions are coded into functions and the different requirements are met.

2. LITERATURE SURVEY

- **H. Huang** et al., (2009) proposed frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages normally keep their CSS vogue like their objective pages.
- **S. Marchal** et al., (2017) proposed this technique to differentiate Phishing websites depending on the examination of authentic site server log knowledge. An application Off-the hook application or identification of phishing websites. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, nice language freedom, speed of selection, flexibility to dynamic phishing, and flexibility to advancement in phishing ways.
- **Mustafa Aydin** et al (2019) proposed a classification algorithm for phishing website detection by extracting websites' URL features and analyzing subset-based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites. The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses Alphanumeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis, and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others are based on third parties services.
- **Samuel Marchal** et al (2017) present Phish Storm, an automated phishing detection system that can analyze in real-time any URL to identify potential phishing sites. Phish storm is proposed as an automated real-time URL phishing Ness rating system to protect users against phishing content. Phish Storm provides a phishing Ness score for URLs and can act as a Website reputation rating system.
- **Fadi Thabtah** et al. experimentally compared large numbers of ML techniques on real phishing datasets and concerning different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more appropriate as anti-phishing solutions.
- **Muhammadet Baykara** et al. proposed an application which is known as Anti Phishing Simulator, it gives information about the detection problem of phishing and how to detect phishing emails. Spam emails are added to the database by the Bayesian algorithm. Phishing attackers use JavaScript to place a legitimate URL of URL onto the browser’s address bar. The recommended approach in the study is to use the text of the e-mail as a keyword only to perform complex word processing.

2.1 Existing System

The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts their accuracy. The existing system has good accuracy but it is still not the best as phishing attack is very crucial, and we have to find the best solution to eliminate this. In the currently existing system, only one machine learning algorithm is used to predict the accuracy, using only one algorithm is not a good approach to improving the prediction accuracy. Each of the algorithms explained in the earlier chapter has some disadvantages hence it is not recommended to use one machine learning algorithm to further improve the accuracy.

2.2 References

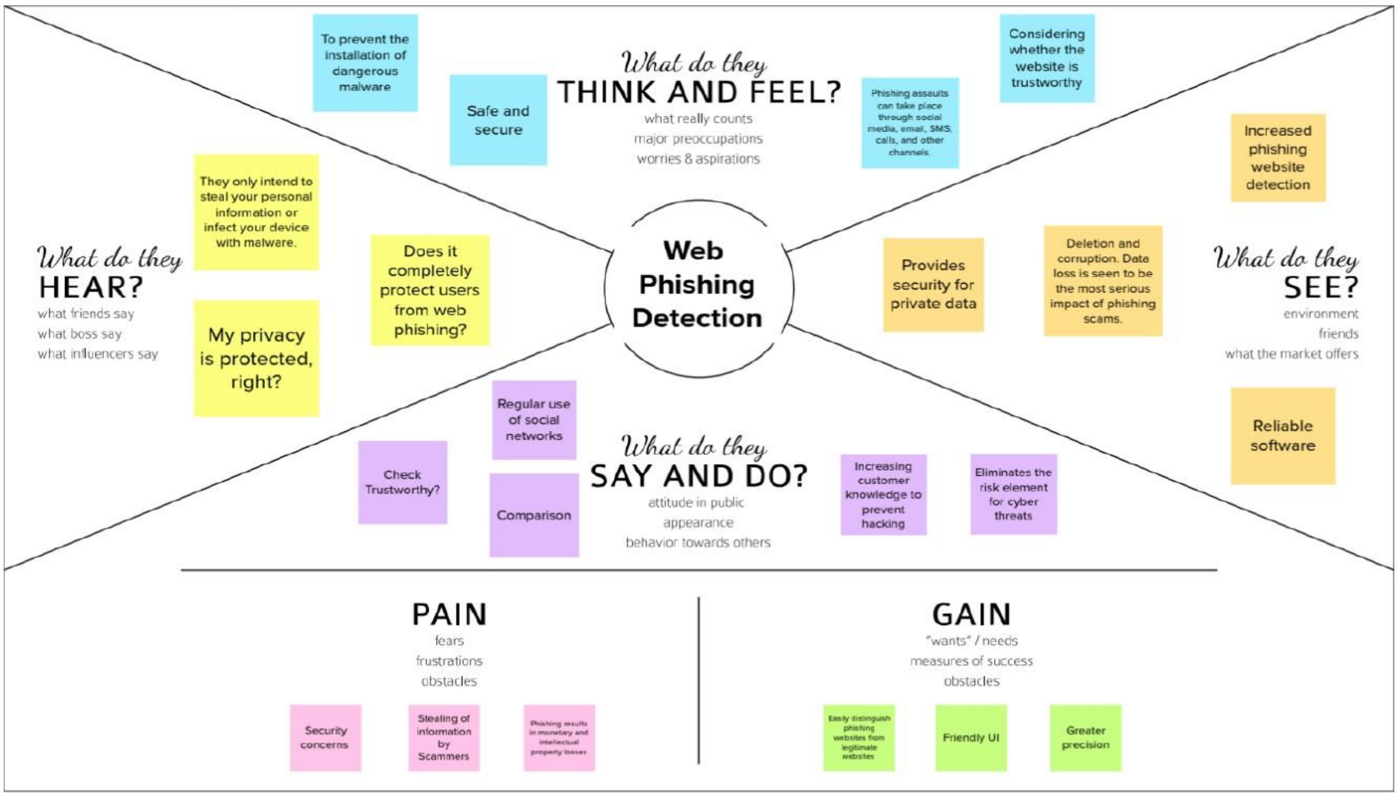
- Matthew Dunlop, Stephen Groat, David Shelly (2010) " Gold Phish: Using Images for Content-Based Phishing Analysis”
- Rishikesh Mahajan (2018) “Phishing Website Detection using Machine Learning Algorithms”
- Purvi Pujara, M. B.Chaudhari (2018) “Phishing Website Detection using Machine Learning: A Review”
- David G. Dobolyi, Ahmed Abbasi (2016) “Phish Monger: A Free and Open-Source Public Archive of Real-World Phishing Websites”
- Satish.S, Suresh Babu. K (2013) “Phishing Websites Detection Based on Web Source Code and URL In the Webpage”
- Purvi Pujara, M. B. Chaudhari (2018) “Phishing Website Detection using Machine Learning: A Review”
- Satish.S, Suresh Babu. K (2013) “Phishing Websites Detection Based on Web Source Code and URL In the Webpage”
- Tenzin Dakpa, Peter Augustine (2017) “Study of Phishing Attacks and Preventions”
- Ping Yi (2018) “Web Phishing Detection Using a Deep Learning Framework”
- Jalil Nourmohammadi Khiarak (2017) “What is Machine Learning”

2.3 Problem Statement Definition

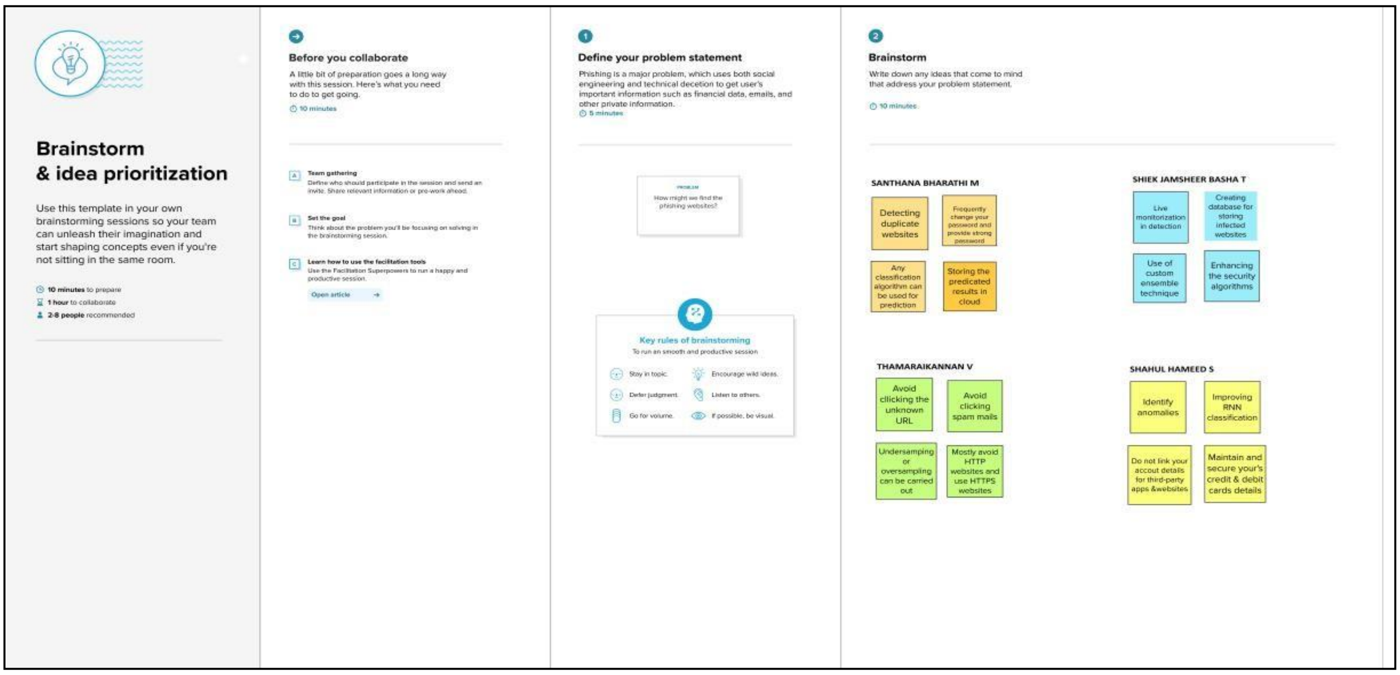
Therefore, a combination of the two can increase the accuracy while implementing different error detection methods. Since scams are widespread and nobody wants to fall for web phishing, regular users who search the internet for information need a way to make sure the links they click are safe. Most phishing attacks exhibit the application of social engineering tactics. Social engineering is the most significant factor that leads to malicious hacking crimes since 99% of cyber-attacks need some level of human intervention to execute. Whereas it may be easy to blame the system users, it is also important to note that phishing has become increasingly sophisticated. It is increasingly becoming challenging to note new attacks since the perpetrators make the email look as coming from a trusted source.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas


What do they THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
To prevent the installation of dangerous malware
Safe and secure
Phishing assaults can take place through social media, email, SMS, calls, and other channels.
Considering whether the website is trustworthy
What do they HEAR?
what friends say
what boss say
what influencers say
They only intend to steal your personal information or infect your device with malware.
My privacy is protected, right?
Does it completely protect users from web phishing?
Web Phishing Detection
What do they SEE?
environment
friends
what the market offers
Increased phishing website detection
Provides security for private data
Deletion and corruption. Data loss is seen to be the most serious impact of phishing scams.
Reliable software
What do they SAY AND DO?
attitude in public
appearance
behavior towards others
Regular use of social networks
Check Trustworthy?
Comparison
Increasing customer knowledge to prevent hacking
Eliminates the risk element for cyber threats
PAIN
fears
frustrations
obstacles
Security concerns
Stealing of information by Scammers
Phishing results in monetary and intellectual property losses
GAIN
"wants" / needs
measures of success
obstacles
Easily distinguish phishing websites from legitimate websites
Friendly UI
Greater precision

3.2 Ideation & Brainstorming



Brainstorm & idea prioritization

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

10 minutes to prepare
1 hour to collaborate
2-8 people recommended

Before you collaborate

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

10 minutes

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

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

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

Open article

1 **Define your problem statement**

Phishing is a major problem, which uses both social engineering and technical decetion to get user's important information such as financial data, emails, and other private information.

5 minutes

PROBLEM
How might we find the phishing websites?

Key rules of brainstorming
To run an smooth and productive session

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

2 **Brainstorm**

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

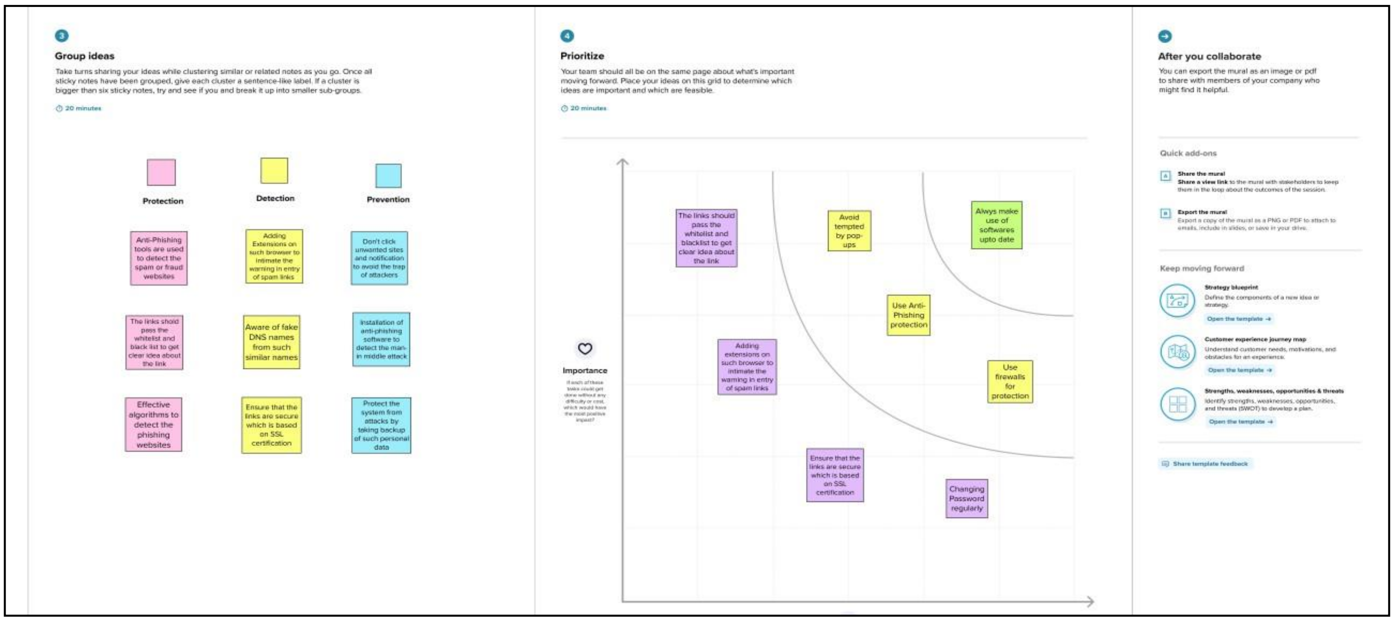
10 minutes

SANTHANA BHARATHI M
- Detecting duplicate websites
- Frequently change your password and provide strong password
- Any classification algorithm can be used for prediction
- Storing the predicated results in cloud

SHIEK JAMSHEER BASHA T
- Live monitorization in detection
- Creating database for storing infected websites
- Use of custom ensemble technique
- Enhancing the security algorithms

THAMARAIKANNAN V
- Avoid clicking the unknown URL
- Avoid clicking spam mails
- Undersamping or oversampling can be carried out
- Mostly avoid HTTP websites and use HTTPS websites

SHAHUL HAMEED S
- Identify anomalies
- Improving RNN classification
- Do not link your accout details for third-party apps &websites
- Maintain and secure your's credit & debit cards details


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.
20 minutes

Protection	Detection	Prevention
Anti-Phishing tools are used to detect the spam or fraud websites	Adding Extensions on such browser to intimate the warning in entry of spam links	Don't click unwanted sites and notification to avoid the trap of attackers
The links shold pass the whitelist and black list to get clear idea about the link	Aware of fake DNS names from such similar names	Installation of anti-phishing software to detect the man-in middle attack
Effective algorithms to detect the phishing websites	Ensure that the links are secure which is based on SSL certification	Protect the system from attacks by taking backup of such personal data

4
Prioritize
Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.
20 minutes

Importance
If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

The links should pass the whitelist and blacklist to get clear idea about the link
Avoid tempted by pop-ups
Alwys make use of softwares upto date
Use Anti-Phishing protection
Adding extensions on such browser to intimate the warning in entry of spam links
Use firewalls for protection
Ensure that the links are secure which is based on SSL certification
Changing Password regularly

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

Quick add-ons
Share the mural
Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.
Export the mural
Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.

Keep moving forward
Strategy blueprint
Define the components of a new idea or strategy.
Open the template →
Customer experience journey map
Understand customer needs, motivations, and obstacles for an experience.
Open the template →
Strengths, weaknesses, opportunities & threats
Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.
Open the template →
Share template feedback

3.3 Proposed Solution

We propose that there are three ways in which the solution to phishing can be approached: detect phishing attacks before they reach the user, detect once the user has reached the phishing site, or train users to detect or prevent them by themselves. Therefore, a combination of the two can increase the accuracy while implementing different error detection methods. Since scams are widespread and nobody wants to fall for web phishing, regular users who search the internet for information need a way to make sure the links they click are safe.

3.4 Problem Solution Fit

Phishing is a type of social engineering attack often used to steal user data. Phishing attacks are becoming more and more sophisticated, and our algorithms are suffering to keep up with this level of sophistication. They have a low detection rate and high false alarm especially when novel phishing approaches are used. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier.

Moreover, a comprehensive blacklist can ensure a perfect up-to-date database. Various other techniques such as page content inspection algorithms have been used to combat the false negatives but as each algorithm uses a different approach, their accuracy varies.

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

These are the requirements that the end user specifically demands as basic facilities that the system should offer. All these functionalities need to be necessarily incorporated into the system as a part of the contract. These are represented or stated in the form of input to be given to the system, the operation performed and the output expected. They are the requirements stated by the user which one can see directly in the final product, unlike the non-functional requirements.

4.2 Non-Functional requirements

These are the quality constraints that the system must satisfy according to the project contract. The priority or extent to which these factors are implemented varies from one project to another. They are also called non-behavioural requirements.

They deal with issues like:

• Portability

• Security

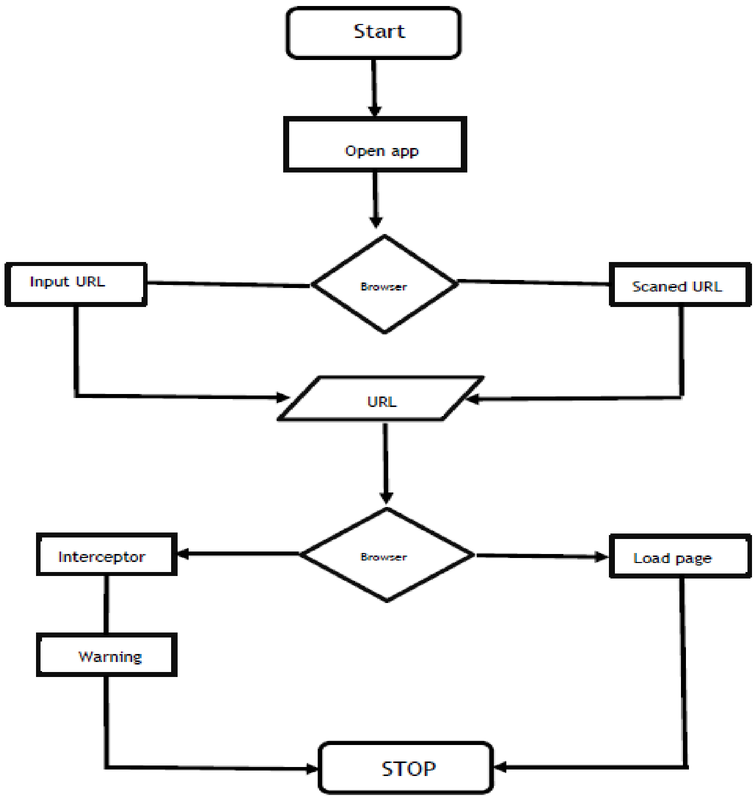
• Maintainability

• Reliability

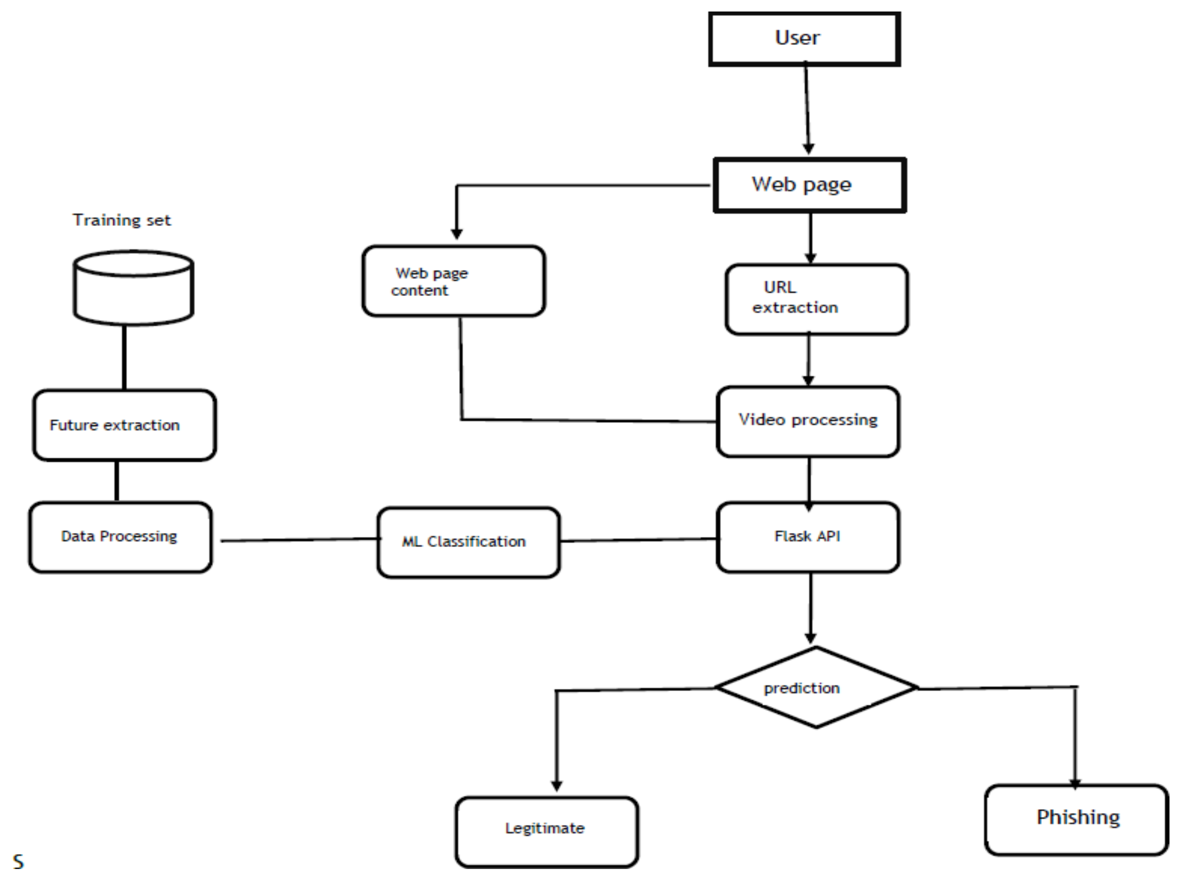
• Scalability

5. PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer	Registration	USN-1	I can sign up for the application as a user by providing my email address, password, and password confirmation.	I can access my dashboard and account.	High	Sprint-1
		USN-2	I will receive a confirmation email after I register as a user for the application.	I can get a confirmation email and confirm it.	High	Sprint-1

		USN-3	I can sign up for the application as a user through LinkedIn.	I may sign up and access the dashboard using my LinkedIn account.	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can register & access the dashboard with Gmail Login	Medium	Sprint-1
	Login	USN-5	I can access the application as a user by providing my email address and password.	I can access my dashboard or account	High	Sprint-1
	Dashboard	USN-6	As a user, I paste the Link that needs to be Verified as a Phishing site or not	I can paste the Link into the Textbox	High	Sprint-1
		USN-7	As a user, I can see the Result	I can view it as a Safe Site	High	Sprint-1

6. PROJECT PLANNING & SCHEDULING

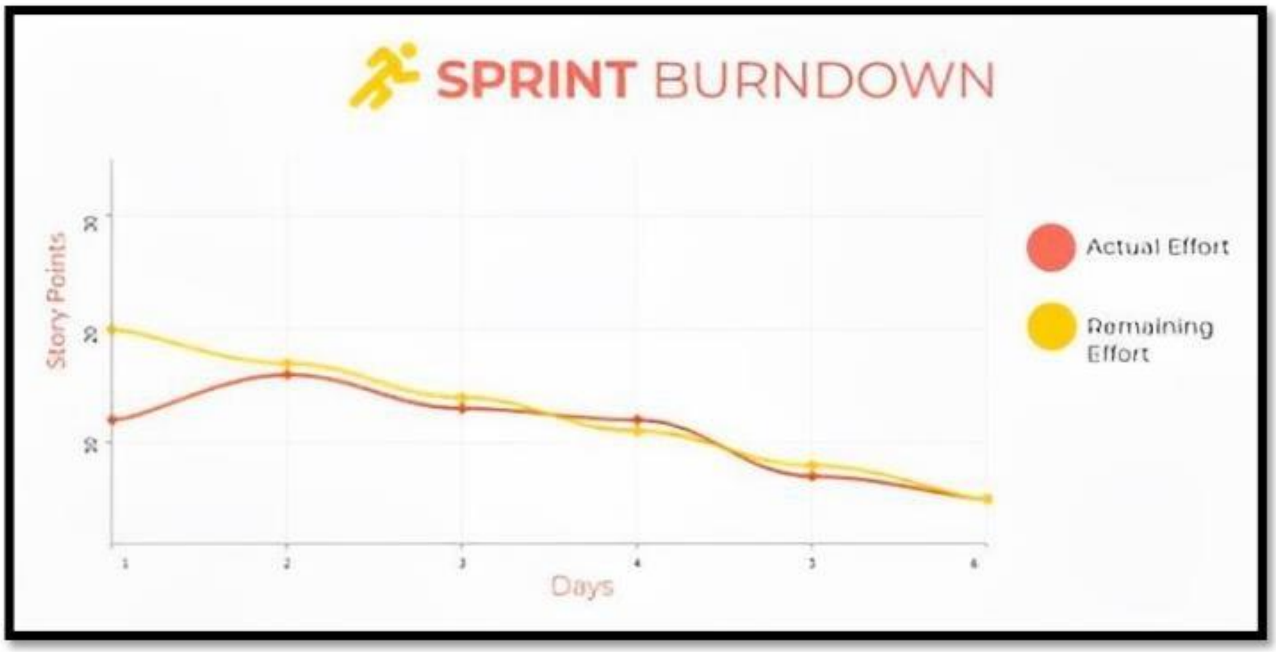
6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, and password, and confirming my password.	2	High	Aruneraj J Prem Kumar A R Santhosh S Syed Abdul Azeem A
Sprint-1		USN-2	As a user, I will receive a confirmation email once I have registered for the application	2	High	Aruneraj J Prem Kumar A R Santhosh S Syed Abdul Azeem A
Sprint-1		USN-3	As a user, I can register for the application through LinkedIn	2	Low	Aruneraj J Prem Kumar A R Santhosh S Syed Abdul Azeem A
Sprint-1		USN-4	As a user, I can register for the application through Gmail	4	Medium	Aruneraj J Prem Kumar A R Santhosh S Syed Abdul Azeem A

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date	Story Points Completed	Sprint Release Date
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 Reports from JIRA



7. CODING & SOLUTIONING

7.1 Feature 1

In the SelectBest () function in the Sklearn module, the parameter k can be passed in, and the top k features with the best performance can be output through chi-square detection. At the same time, these features are independent of each other to the greatest extent, and the discrimination is good. Among them, the first six features with the best classification effect are (text features after word segmentation are not considered. WHOIS information The first is to classify the best-performing WHOIS information. The registration date of generally benign webpages is relatively long, and malicious webpages often need to be re-registered due to reasons such as being destroyed by security practitioners. The registration date is often the nearest, so the distinction is better and the classification effect is improved.

7.2 Feature 2

Next is the length of the URL which ranks second in performance. The distribution of the length of 7030 URLs in the training sample according to the labels are labeled blue represents benign URLs, and red represents malicious URLs. It can be seen that the length of all URLs is concentrated in 0 ~ 200 characters, of which benign URLs are distributed in the range of 25 ~ 30, while malicious URLs are distributed in the range of 40 ~ 75. Although there is a partial intersection between the two, the distinguishing effect is still obvious, you can see the length of the malicious URL Relatively speaking, it will be longer than a benign URL, so in terms of discrimination, URL length is a relatively good feature.

8. TESTING

8.1 Test Cases

The test data come from ISP and are composed of two data sets. The small data set includes real traffic flow for 40 minutes. The big data set includes real traffic flow for 24 hours. After pre-treatment, we get record sum, unique IP, unique AD, and unique URL This paper belongs to a classical binary classification model application. In the binary classification model, the results are usually marked as Positive (P) or Negative (N). In this paper, the corresponding node is either a phishing site or not a phishing site

- True Positive (TP) : is P and the classification is also P
- False Positive (FP) : is N and the classification is also P
- True Negative (TN) : is N and the classification is also N
- False Negative (FN) : is actually and the classification is also N

The above classification data can generate four categories of evaluation criteria with details as follows:

- Accuracy (ACC) : $ACC = (TP+TN) / (TP+TN+FP+ FN)$
- True Positive Rate (TPR, Recall) : $TPR = TP/ (TP + FN)$
- False Positive Rate (FPR, Fall-Out) : $FPR = FP/ (FP + TN)$
- Positive Predictive Value (PPV, Precision): $PPV = TP/ (TP + FP)$

8.2 User Acceptance Testing

There are three parameters to affect the accuracy. They are the number N of the DBN layer, the number T of iterations per layer, and the number of nodes in hidden layers. L. McAfee shows that when the number of iterations and the number of hidden layer nodes exceed a certain threshold, the precision of the algorithm will reach a higher level. With the number of iterations or hidden layer nodes increasing, the detection rate will be a small drop. The reason may be overfitting. Therefore, we first set the larger number of iterations T = 1000 and hidden layer nodes, such as $layers = \{top = 100, hidden = 50, . . ., 50, \mathrm{V}isible = 10\}$. TPR is related to the number of layers. When the number of layers is 2, TPR gets the top level at about 89%. As the number of layers increases, TPR decreases a little. The reason is that too many layers will lead to overfitting. Therefore, the best number of layers is two layers. TPR is related to the number of iterations. The results show that when the number of iterations is at 200, the detection rate is above 80%. The highest detection rate achieves at about 250 iterations. After that, the accuracy of the algorithm decreases with the increase in the number of iterations. Moreover, the more iterations of each layer, the longer the algorithm's overall run time.

9. RESULTS

After reviewing and researching appropriate monitoring tools, the proposed system has been identified and chosen to address the complexity of monitoring requirements for the current situation. This software is designed to show awareness of the extensive level of its functionality and features that can be displayed in the monitoring era. The system fosters many features in comparison to other software. Its unique features such as capturing blacklisted URLs from the browser directly to verify the validity of the website, notifying users on blacklisted websites while they are trying to access them through pop-ups, and also notifying through email. This system will assist the user to be alert when they are trying to access a blacklisted website.

9.1 Performance Metrics

TPR is related to the number of hidden units. The results show that TPR increases significantly to above 85% when the number of hidden units gets 20. The detection rate does not change much under 30 hidden units. And when it gets to 40 hidden nodes, the detection rate again significantly increases and reaches nearly 90%. Since then, as the number of nodes increases, the detection rate under 80 hidden units is slightly higher than 90%. But the overall detection rate does not significantly change, after more than 40 hidden nodes. As the number of hidden layers nodes increase, the running time also significantly increases. Therefore, the number of hidden units should be 40. Table 2 shows TPR between BP and no BP. We find that fine-tuning in BP does not improve the TPR but reduces the detection rate and increases running time. Therefore, we do not use BP in detection. After training and getting the parameters in the small data set, we used it to detect the phishing websites in the big data set. The results show that there were 17672 nodes in phishing websites, and the detection rate was 89.2%. The FPR was 0.6%. Because the big data set cannot be fully calibrated, the results are only reference significance.

10. ADVANTAGES & DISADVANTAGES

Feature extraction, that's an Attribute extension that allows us to create more columns from URLs. Finally, we use a classifier algorithm to train our models. They take advantage of the obtained classified dataset. The remainder of our classified data would be used to validate the models. ML algorithms have been used to identify pre-processed data. The classifier utilized had been Random Forest.

Successful phishing attack scares customers away from a business. A UK survey revealed that more than half of consumers stop patronizing a hacked organization for several months after a data breach. Some 41% of customers no longer patronize businesses that got their data leaked. This effect could haunt an organization for a long time.

11. CONCLUSION

Thus to summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks. We have tested two machine learning algorithms on the ‘Phishing Websites Dataset’ and reviewed their results. We then selected the best algorithm based on its performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using the Random Forest algorithm with an accuracy of 97.31%. For future enhancements, we intend to build the phishing detection system as a scalable web service that will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

12. FUTURE SCOPE

Although the use of URL lexical features alone has been shown to result in high accuracy (∼97%), phishers have learned how to make predicting a URL destination difficult by carefully manipulating the URL to evade detection. Therefore, combining these features with others, such as the host, is the most effective approach. For future enhancements, we intend to build the phishing detection system as a scalable web service that will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

13. APPENDIX

```
#Importing required libraries
from flask import Flask, request, render_template
import numpy as np
import pandas as pd
from sklearn import metrics
import warnings
import pickle
warnings.filterwarnings('ignore')
from feature import FeatureExtraction

import requests

# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud
account.
API_KEY =  "V7p-6PyIwKuNRoU4L2a38uPSAmc8QzciVBPNZbhI2hne"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":
API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

app = Flask(_name_)

@app.route("/", methods=["GET", "POST"])
def index():
    if request.method == "POST":
        url =  request.form["url"]
        obj = FeatureExtraction(url)
        x = np.array(obj.getFeaturesList()).reshape(1,30)

        # NOTE: manually define and pass the array(s) of values to be scored in the next line
        payload_scoring = {"input_data": [{"fields":
[['f0','f1','f2','f3','f4','f5','f6','f7','f8','f9','f10','f11','f12','f13','f14','f15','f16','f17','f18','f19','f20','f21','f22','f23','f2
4','f25','f26','f27','f28','f29']], "values": [[-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-
1]]}]}
```

```
        response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/7bc163a4-
37f7-4ec6-b0bf-63d1e40a85e4/predictions?version=2022-11-16', json=payload_scoring,
        headers={'Authorization': 'Bearer ' + mltoken})
        #print("Scoring response")
        #print(response_scoring.json())
        pred=response_scoring.json()
        output=pred['predictions'][0]['values'][0][0]

        y_pred =output
        #1 is safe
        #-1 is unsafe
        y_pro_phishing = gbc.predict_proba(x)[0,0]
        y_pro_non_phishing = gbc.predict_proba(x)[0,1]
        if(y_pred ==1 ):
            pred = "It is gg{0:.2f} %safe to go ".format(y_pro_phishing*100)
        return render_template('index.html',xx =round(y_pro_non_phishing,2),url=url )
    return render_template("index.html", xx =-1)
if__name__== "_main_":
    app.run(debug=True)
    # NOTE: manually define and pass the array(s) of values to be scored in the next line
    payload_scoring = {"input_data": [{"fields":
[['f0','f1','f2','f3','f4','f5','f6','f7','f8','f9','f10','f11','f12','f13','f14','f15','f16','f17','f18','f19','f20','f21','f22','f23','f2
4','f25','f26','f27','f28','f29']], "values": [[-1,1,1,1,-1,-1,-1,-1,-1,1,1,-1,1,-1,1,-1,-1,-1,0,1,1,1,1,-1,-1,-1,-1,1,1,-
1]]}]}

    response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/7bc163a4-37f7-
4ec6-b0bf-63d1e40a85e4/predictions?version=2022-11-16', json=payload_scoring,
    headers={'Authorization': 'Bearer ' + mltoken})
    #print("Scoring response")
    #print(response_scoring.json())
    pred=response_scoring.json()
    output=pred['predictions'][0]['values'][0][0]
```

GitHub Link: https://github.com/IBM-EPBL/IBM-Project-28393-1660111365

Demo Video Link: IBM-Project-28393-1660111365/Project Demo Video.mp4 at main · IBM-EPBL/IBM-Project-28393-1660111365 (github.com)