WEB PHISHING DETECTION

A PROJECT REPORT

Submitted by

SARANAMALAR V

SRI HARI R

SHARMILA TAJ S

THANIGAVEL M

VISHNU DAS P M

of

ELECTRONICS AND
COMMUNICATION
ENGINEERING

SONA COLLEGE OF
TECHNOLOGY

SALEM

CHAPTER 1
INTRODUCTION

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels.
Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which 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 contents.
In this article I 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.

1.1 PROJECT OVERVIEW :

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this notebook is to collect data & extract the selctive features form the URLs.

1.2 PURPOSE

The main reason is the lack of awareness of users. But security defenders must take precautions to prevent users from confronting these harmful sites. Preventing these huge costs can start with making people conscious in addition to building strong

security mechanisms which are able to detect and prevent phishing domains from reaching the user.

CHAPTER 2

LITERATURE SURVEY :

2.1 EXISTING PROBLEM

Many users unwittingly click phishing domains every day and every hour. The attackers are targeting both the users and the companies. According to the 3rd Microsoft Computing Safer Index Report, released in February 2014, the annual worldwide impact of phishing could be very high as $5 billion.

2.2 REFERENCES

1.Anti-Phishing Working Group (APWG), https://docs.apwg.org//reports/apwg_trends_report_q4_2019.pdf

2.Jain A.K., Gupta B.B. “PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning”, Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, doi: 10.1007/978-981-10-8536-9_44.

3.Purbay M., Kumar D, “Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection”, Lecture Notes in Electrical Engineering, vol. 683, 2021, doi: 10.1007/978-981-15-6840-4_40.

4.Gandotra E., Gupta D, “An Efficient Approach for Phishing Detection using Machine Learning”, Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12.

5.Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, “URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection”, Conference’17,

Washington, DC, USA, arXiv:1802.03162, July 2017.

6.Hong J., Kim T., Liu J., Park N., Kim SW, “Phishing URL Detection with Lexical Features and Blacklisted Domains”, Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.

7.J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, “Phishing Website Classification and Detection Using Machine Learning,” 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

8.Hassan Y.A. and Abdelfettah B, “Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

9.Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1;83:246–67.

10.Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020 Sep;9(9):1514.

2.3 PROBLEM STATEMENT DEFINITION:

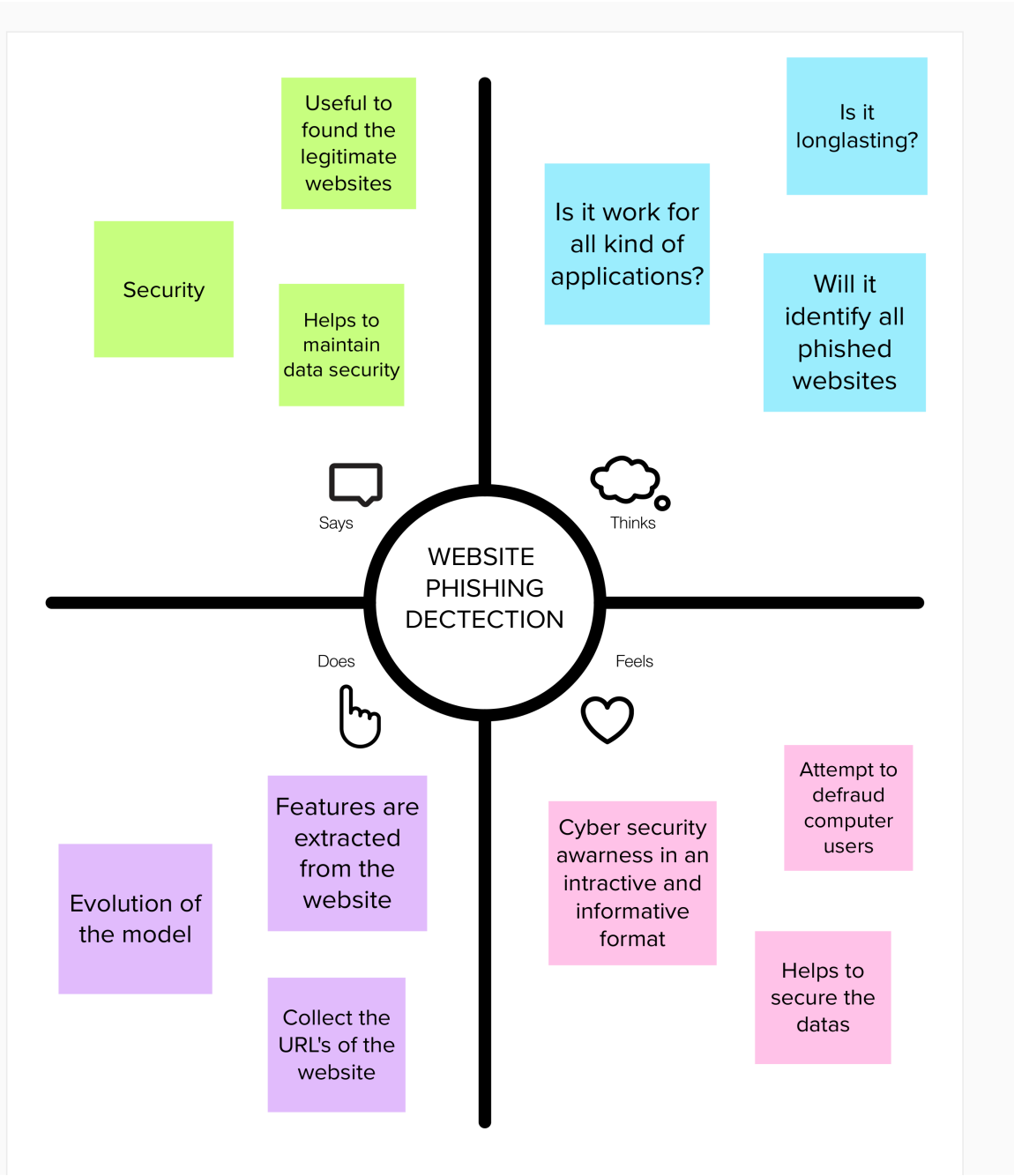
In recent years, advancements in Internet and cloud technologies have led to a significant increase in electronic trading in which consumers make online purchases and transactions. This growth leads to unauthorized access to users’ sensitive information and damages the resources of an enterprise. Phishing is one of the familiar attacks that trick users to access malicious content and gain their information. In terms of website interface and uniform resource locator (URL), most phishing webpages look identical to the actual webpages. Various strategies for detecting phishing websites, such as blacklist, heuristic, Etc., have been suggested. However, due to inefficient security technologies, there is an exponential increase in the number of victims. The anonymous and uncontrollable framework of the Internet is more vulnerable to phishing attacks. Existing research works show that the performance of the phishing detection

system is limited. There is a demand for an intelligent technique to protect users from the cyber-attacks. In this study, the author proposed a URL detection technique based on machine learning approaches. A recurrent neural network method is employed to detect phishing URL. Researcher evaluated the proposed method with 7900 malicious and 5800 legitimate sites, respectively. The experiments' outcome shows that the proposed method's performance is better than the recent approaches in malicious URL detection.

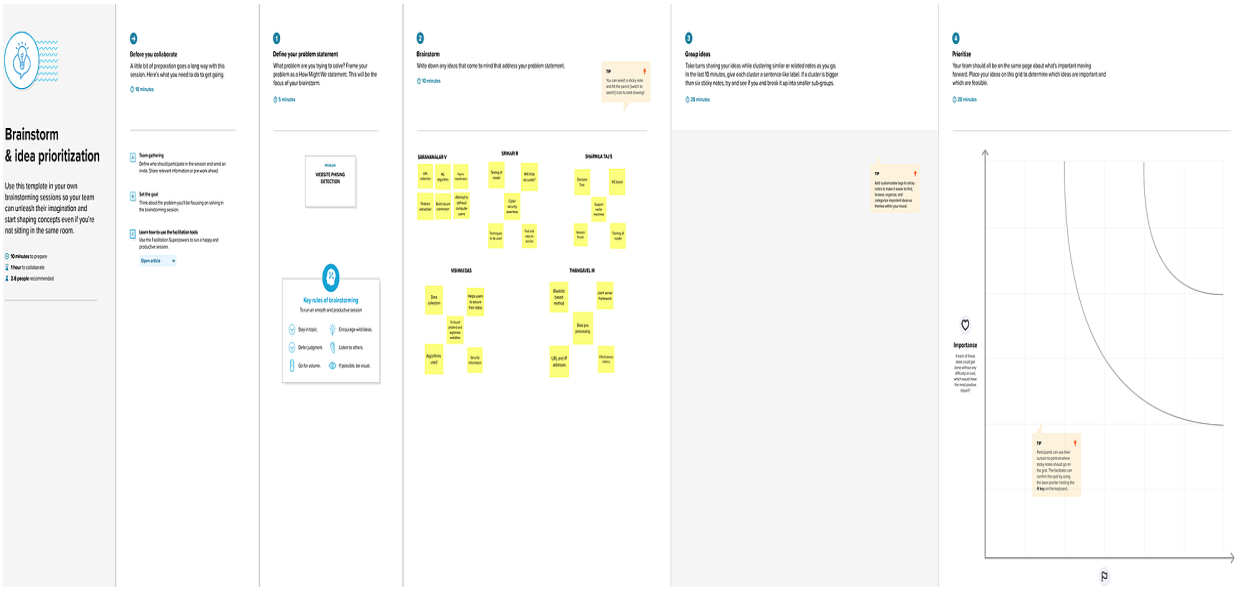
CHAPTER 3

IDEATION & PROPOSED SOLUTION:

3.1 EMPATHY MAP CANVAS:



3.2 IDEATION & BRAINSTORMING:



3.3 PROPOSED SOLUTION:

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.

3.4 PROBLEM SOLUTION FIT:

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database. Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement the vulnerabilities of the stale lists. Moreover, page content inspection algorithms each have different approach to phishing website detection with varying degrees of accuracy. Therefore, ensemble can be seen to be a better solution as it can combine the similarity in accuracy and different error-detection rate properties in selected algorithms.
2.	Idea / Solution description	A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.

3.	Novelty / Uniqueness	Before stating the ML model training, the data is split into 80-20 i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.
4.	Social Impact / Customer Satisfaction	The website shows information regarding the services provided by us. It also contains information regarding ill- practices occurring in todays technological world. The website is created with an opinion such that people are not only able to distinguish between legitimate and fraudulent website, but also become aware of the mal-practices occrring in current world. They can stay away from the people trying to exploit ones personal information, like email address, password, debit card numbers, credit card details, CVV, bank account numbers, and the list goes on.
5.	Business model	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 host, is the most effective approach .
6.	Scalability of solution	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.

CHAPTER 4

REQUIREMENT ANALYSIS:

4.1 FUNCTIONAL REQUIREMENT:

Following are the functional requirements of the proposed solution.

FR NO	Functional Requirements	Classification
FR-1	Fetch Electronic Mail Messages	Core
FR-2	Extract URLS	Core
FR-3	Extract Header Information	Core
FR-4	Classify Email	Core
FR-5	Static or Dynamic (Inbox)	Core
FR-6	Provide User Feedback	Core

4.2 NON FUNCTIONAL REQUIREMENTS:

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

FR NO	Non-Functional Requirements	Description
NFR-1	Usability	System is easy to configure and is efficient in carrying out user tasks.
NFR-2	Availability	System is available to work asrequired whenit is required.
NFR-3	Reliability	System will perform the tasks it wasdesigned to do.
NFR-4	Performance	System will perform tasks in a fashion that complies withpredetermined criteria.
NFR-5	Security	System will protect all data manipulated internally from unauthorized accessand threats.
NFR-6	Scalability	System will appropriately handleincreasing and decreasing workloads.

CHAPTER 5

PROJECT DESIGN

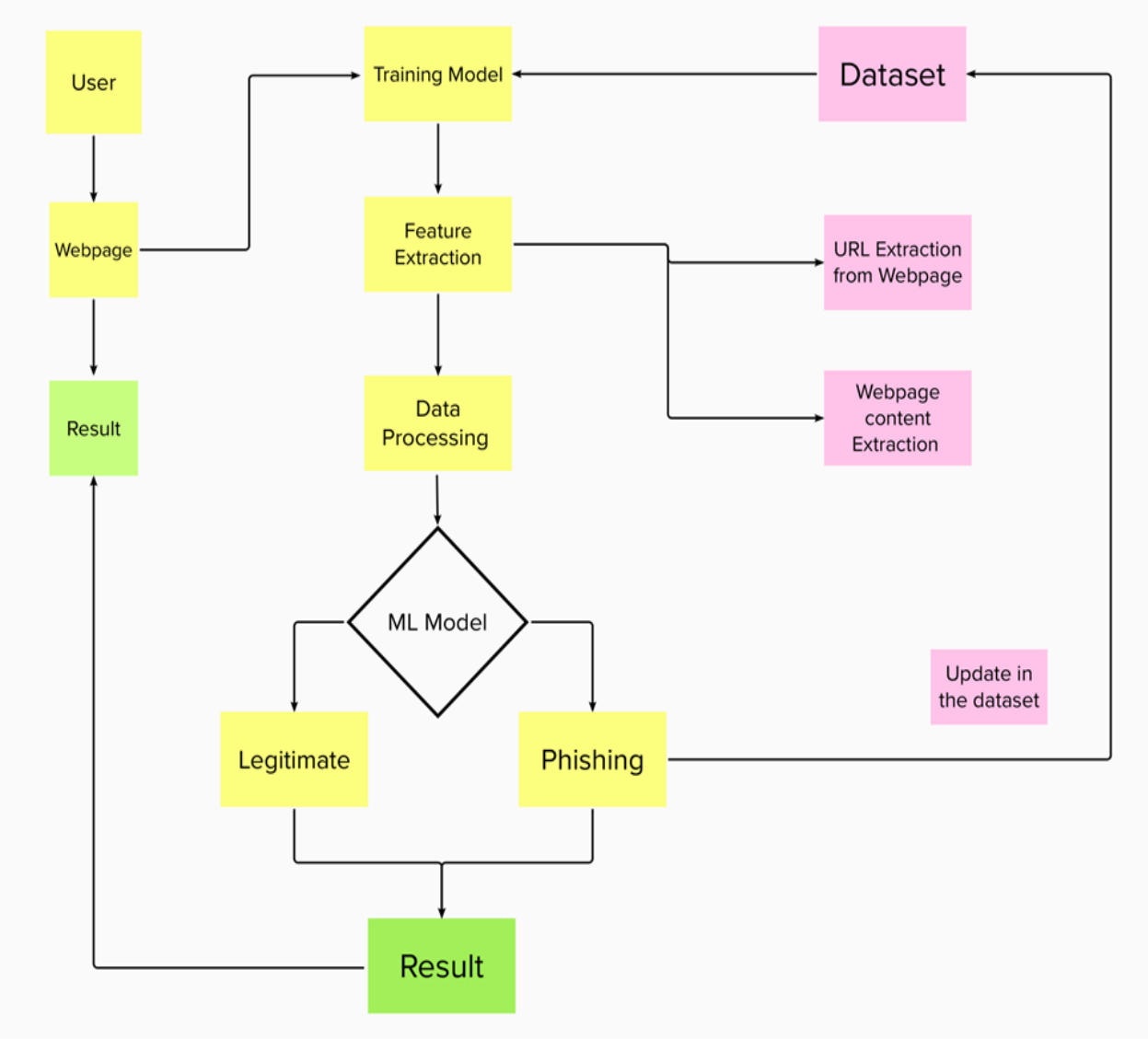
A popular and common method for cybersecurity threats is phishing URLs. Phishing is a type of cybercrime that aims to convince its victims to provide the attacker access to their private and sensitive information. The attacker wants to obtain personalinformation such user names, passwords, financial account information, information from socialnetworking sites, and addresses.

Then, these confidential login information is frequently exploited for nefarious purposes including fraud, infamy, profit, reputation damage, and many other unlawful acts. This paper provides a thorough analysisof the various systems currentlyin use for phishing website detection. The technique described here makes use of advanced machine learning to classify webpages as phishing or starting with greater precision and accuracy.

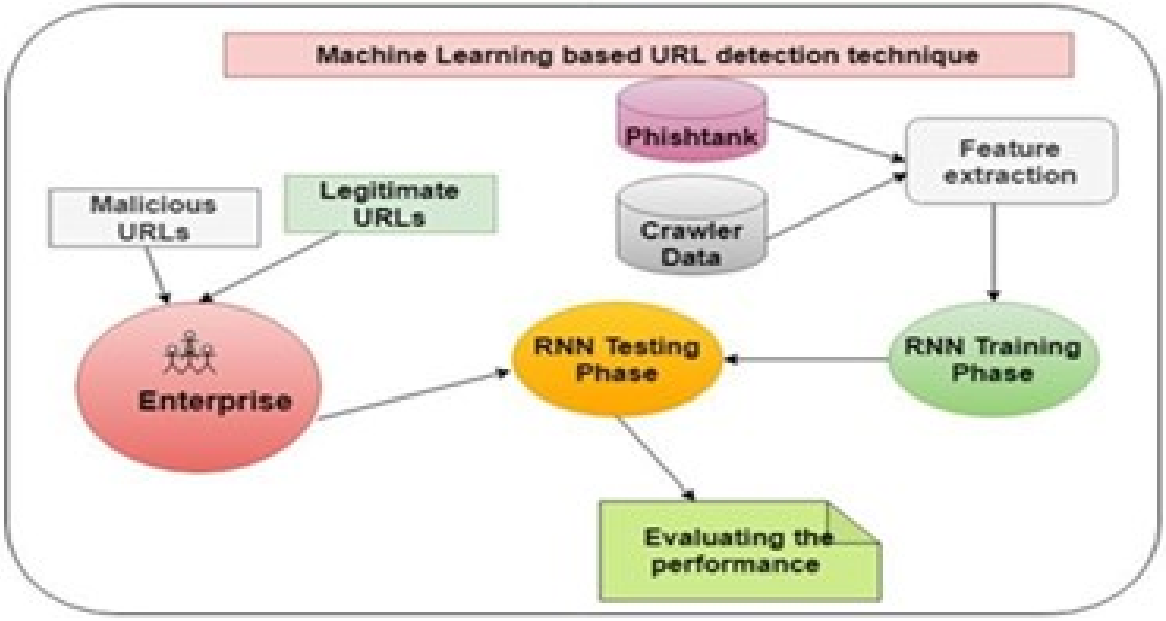
Due to the anonymity offeredby the internet and the rapid growthof online transactions, hackers try to trick end users by using techniques like phishing, SQL injection, malware, man-in-the-middle attacks, domain name system tunnelling, ransomware, web trojans, and so on.

Phishing is said to be the most misleading attackamong all of these. Usually,the goal is to enticepeople to divulgesensitive data like system loginsor financial information.

5.1 DATA FLOW DIAGRAMS:



5.2 SOLUTION & TECHNICAL ARCHITECTURE:



The Deliverable shall include the architectural diagramas below and the information as per the table1 & table 2

Table-1 : Components & Technologies:

S.No	Component	Description	Technology

1.	User Interface	How user interacts with application e.g.Web UI, Mobile App, Chatbot etc.	HTML, CSS, JavaScript
2.	Application Logic forlogic	Logic for a processin the application	Flask login(Python)
3.	Cloud Database	Database Serviceon Cloud	IBM Watson
4.	File Storage	File storagerequirements	MongoDB
5.	Machine Learning Model	Purpose of Machine Learning Model	Logistic Regression,Decision Tree
6.	Infrastructure (Server / Cloud)	Application Deployment on Local System/ Cloud LocalServer Configuration: Cloud Server Configuration :	Local, Render, IBM Cloud

Table-2: Application Characteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Sckit Learn package in Python that deals withML algorithms	Machine Learning
2.	Security Implementations	Typosquatting, Cybersquatting	Cybersecurity
3.	Scalable Architecture	Justify the scalability of architecture (3 – tier, Micro-services)	Technology used
4.	Availability	It can balancethe load traffic among the serversto help improve uptime. Can scaleapplications by adding or removing servers, with minimal disruption to traffic flows.	IBM Cloud Load Balancers
5.	Performance	It provides performance feedback such as page size andhow long it takes to load a page, and canshow the impact new features have on the performance of the site.	Blacklists/whitelists, Natural language Processing, Visual similarity, rules, machine learning techniques, etc

5.3 USER STORIES

Use the below templateto list all the user stories for the product.

User Type	Functional Requirement(Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobileuser)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation emailonceI have registered for the application	I can receive confirmationemail &click confirm	High	Sprint-1
		USN-3	As a user,I can register for the application through Facebook	I can register & access the dashboard withFacebook Login	Low	Sprint-2
		USN-4	As a user,I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user,I can log into the application by entering email &password		High	Sprint-1
	Dashboard					

Customer (Webuser)	User Input	USN-1	As a user, I can enterthe required URLin theboxwhile awaiting validation.	I can access the website without any problem	High	Sprint-1
Customer Care Executive	Feature Extraction	USN-1	In the eventthat nothing is discovered during comparison, we can extractfeatures using aheuristic and a visualsimilarity technique.	As a userI can have comparis on between websites for security	High	Sprint-1
Administrator	Prediction	USN-1	The model will use machinelearni ng algorithms like a logistics regression and KNNto forecast the URLs of the websites.	I can accurately forecast the specific algorithms in this way.	High	Sprint-1
	Classifer	USN-2	To create the final product, I will now feed all of the model outputto classifier.	I'll use thisto identify the appropriate classifier for generating the outcome.	Medi um	Sprint-2

CHAPTER 6
PROJECT PLANNING &SCHEDULING:

6.1 Sprint Planning & Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	StoryPoints	Priority	TeamMembers
Sprint-1	Login	USN-1	As a user,I can navigate into the website.	1	High	Amala
Sprint-1	Dashboard	USN-2	As a user, I will input any site's URL in the form to checkits genuineness.	1	High	Annie
Sprint-1		USN-3	As a user, Ican see theoutput.	2	High	Akshaya
Sprint-2	Backend	USN-4	As an admin, if a new URL isfound, I canaddthe new state into the database.	3	Medium	Shekinah

Sprint-3	Report	USN-5	As a user, Ican ask myqueries andreport suspicious sitesin the report box.	1	Low	Akshaya
Sprint-4		USN-6	As an admin, I can take actions to the queries asked by the user.	2	Low	Shekinah

6.2 Sprint DeliverySchedule:

Sprint	Total StoryPoints	Duration	SprintStart Date	Sprint EndDate (Planned)	Story PointsCompleted(as on Planned End Date)	Sprint ReleaseDate (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

CHAPTER 7
CODING & SOLUTIONING

7.1 Feature 1 – Classification of URL:

The primary feature of this project is to classify the given URL as phishing or benign. Variousclassification algorithms are used to achieve this.
Methodology:

7.11 Data collection:

URL features of legitimate websites and phishing websites were collected. The data set consists of total 11,055 URLs which include 6,157 legitimate URLs and 4,898 phishing URLs. Legitimate URLs arelabelled as “1” and phishing URLs are labelled as “-1”. The features that are present in the data set include:

a. IP Address in URL

b. Length of URL

c. Using URL Shortening Services

d. "@" Symbolin URL

e. Redirection "//" in URL

f. Prefix or Suffix "-" in Domain

g. Having Sub Domain

h. Length of Domain Registration

i. Favicon

j. Port Number

k. HTTPS Token

l. Request URL

m. URL of Anchor

n. Links in Tags

o. SFH

p. Email Submission

q. Abnormal URL

r. Status Bar Customization (on mouse over)

s. Disabling RightClick

t. Presence of Popup Window

u. IFrame Redirection

v. Age of Domain

w. DNS Record

x. Web Traffic

y. Page Rank

z. Google Index

aa. Links pointing to the page

ab. Statistical Report

ac. Result

Using IBM Cloud Storage this data is accessed throughout the project. The code written below isusedto import the dataset.

Data pre-processing and Exploratory Data Analysis:

Few plots and graphs were drawn to find how the data is distributed and the how features arerelated to each other.

Univariate analysis:

Univariate analysis provides an understanding in the characteristics of each feature in the data set.Different characteristics are computed for numerical and categorical data. For the numerical features characteristics are standard deviation, skewness, kurtosis, percentile, interquartile range (IQR) and range. For the categorical features characteristics are count, cardinality, list of unique values, top and freq.

CHAPTER 8
TESTING

8.1 Test Cases:

Test case ID	Feature Type	Component	Test Scenario	Steps To Execute	Test Data	Expected Result	Actual Result	Status
DashBoard_TC_OO1	Functional	Home Page	Verify user is able to enter the URL in the form	1. Open Hook Phish website 2. Enter a URL and click submit	https://google.com/	Result of classification willbe displayed	Working as expected	Pass
DashBoard_TC_OO2	UI	Home Page	Verify the UI elements in the form	1. Enter URL and clickgo 2. The services and teams' sectio	https://google.com/	Application should show below UI elements: a. input form b. submit button c. services	Working as expected	Pass

				ns are visible 3. Enter a URL and click sub mit		d. team		
DashBoard _TC_OO3	Func ti onal	Ho me page	Verify user is able to see an alert when nothing is entered in the textbox	1. Enter URL and clickgo 2. Enter nothi ng and click submit 3. An alert is display ed to provide proper input		Alert of incomplete input	Working as expect ed	Pass
DashBoard _TC_OO4	Func ti onal	Ho me page	Verify user is able to see the result when URL is entered in the textbox	1. Enter URL and clickgo 2. Enter any URL and click submit 3. The result of the classi	https://go ogle.com/	Result of classification willbe displayed	Working as expect ed	Pass

				ficati on is displa yed.				
Report_TC_ OO1	Func ti onal	Repo rt page	Verify user is able to enter their name, email and query message inthe form	1. Enter URL and clickgo 2. Cli ck on repo rt butt on 3. Enter Valid name, email and query in the form 4. Click on sub mit butt on	**Name:** Alex **Email:**alex123 @gmail.co m **Query:** Hey! I need to check if a website is legitimate	Details are storedin the database	Working as expect ed	Pass

8.2 User Acceptance Testing:

Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

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

Test Case Analysis:

This report showsthe 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

CHAPTER 9
RESULTS

9.1 PERFORMANCE METRICES:

Performance Evaluation:

In [71]:

```
#computing the accuracy of the model performance
acc_train_svm = accuracy_score(y_train,y_train_svm)
acc_test_svm = accuracy_score(y_test,y_test_svm)

print("SVM: Accuracy on training Data: {:.3f}".format(acc_train_svm))
print("SVM : Accuracy on test Data: {:.3f}".format(acc_test_svm))
```

```
SVM: Accuracy on training Data: 0.899
SVM : Accuracy on test Data: 0.892
```

CHAPTER 10
ADVANTAGES & DISADVANTAGES

Phishing email is one of the major issues of the web nowadays ensuing in monetary

losses for organizations and individual users. Varied approaches are developed to filter phishing emails. The current paper focuses on machine learning applications used to detect and predict phishing emails.

No	Techniques Used	Advantages	Disadvantages
1	*Methods based on Bag-of-Words model*	-Build secure connection between user's mail transfer Agent (MTA) and mail user agent (MUA)	-Time consuming - huge number of features -consuming memory
2	*Compared multi Classifiers algorithms*	-Provide clear idea about the effective level of each classifier on phishing email detection	Non standard classifier
3	*hybrid system*	-High level of accuracy by take the advantages of many classifiers	-Time consuming because this technique has many layers to make the final result
4	*Classifiers Model-Based Features*	- High level of accuracy - create new type of features like Markov features	-huge number of features -many algorithm for classification which mean time consuming -higher cost -need large mail server and high memory requirement
5	*Clustering of Phishing Email*	-Fast in classification process	-Less accuracy because it depend on unsupervised learning , need feed continuously
6	Evolving Connectionist System (ECOS) for phishing email detection	fast ,less consuming memory, high accuracy, Evolving with time, online working	Need feed continuously

CHAPTER 11
CONCLUSION

With this the objective of this notebook is achieved. We finally extracted 18 features for 10,000 URL which has 5000 phishing & 5000 legitimate URLs.

CHAPTER 12
FUTURE SCOPE

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

CHAPTER 13
APPENDIX

SOURCE CODE:

https://github.com/IBM-EPBL/IBM-Project-26356-1660025547/blob/main/Project%20Development%20phase/Sprint%201/Preprocessing%26ModelBuilding.ipynb

GITHUB LINK:
https://github.com/IBM-EPBL/IBM-Project-26356-1660025547

VEDIO LINK:

https://drive.google.com/file/d/1lqe3g-kfJprygnQpjziy3jRaKJXsAe32/view?usp=share_link