PHISHING WEBSITE DETECTION USING MACHINE LEARNING

ABSTRACT

Phishing are one of the most common and most dangerous attacks among cybercrimes. The aim of these attacks is to steal the information used by individuals and organizations to conduct transactions. Phishing websites contain various hints among their contents and web browser-based information. The purpose of this study is to perform classification and prediction of phishing attacks using random forest SVM and logistics regression based classification for 30 features including phishing website data.

INTRODUCTION

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies.

Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints among their contents and web browser-based information. Individual committing the fraud sends the fake website or e-mail information to the target address as if it comes from an organization, bank or any other reliable source that performs reliable transactions.

Phishing Web sites Features Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below. In the first section we defined rules and we gave equations of web features. We need these equations in order to explain phishing attacks characterization.

MOTIVATION

- Phishing is the most commonly used social engineering and cyber attack.
- Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently.
- In order to avoid getting phishing,
  - User should have awareness about phishing websites.
  - Have a blacklist of phishing websites which requires the knowledge of website being detected as phishing.
  - Detect them in their early appearance , using machine learning and deep neural network algorithms.
- Of the above three, the machine learning based method is proven to be most effective than the other methods.
- Even then, online users are still being trapped into revealing sensitive information in phishing websites.

SYSTEM ANALYSIS

PROBLEM STATEMENT

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 as become easier.

1.How to process raw data set for phishing detection?

2.How to increase detection rate in phishing websites algorithms?

3.How to reduces false negative rate in phishing websites algorithm?

4.What are the best compositions of classifiers that can give a good detection rate of phishing website?

OBJECTIVE

A phishing websites 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 data set 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.

EXISTING SYSTEM

Phishing websites mostly get the e-banking sites and attack their passwords, credit card number, bank account and personal details of the user. He says it’s a “New Internet Crime”. Comparing with the forma like virus and hacking the phishing is mostly popular now days. In this they introduce a risk assessment model with the help of the fuzzy rule and classification algorithm.

DRAWBACKS

- As the social phishing attacks underscore the dangers of the public it takes all the personal information’s and need to adequate counter measures.
- In existing methods they fail to find the phishing websites, but they tried it to a mark upto 50% still they can’t succeed.

PROPOSED SYSTEM

In this study, we implement different classification algorithm like svm , random forest and logistic regression based classification was performed for the following 30 features extracted based on the features of websites in UC Irvine Machine Learning Repository. Procedural steps for solving the classification problem presented is as follows:

• **Identification of the problem**

This study attempts to solve the problem as to how phishing analysis data will be classified.

• **Data set**

We scrap the features from different website and created the dataset containing 7000 records the 32 features extracted asked on the features of websites UC lrvine machine learning repository database.

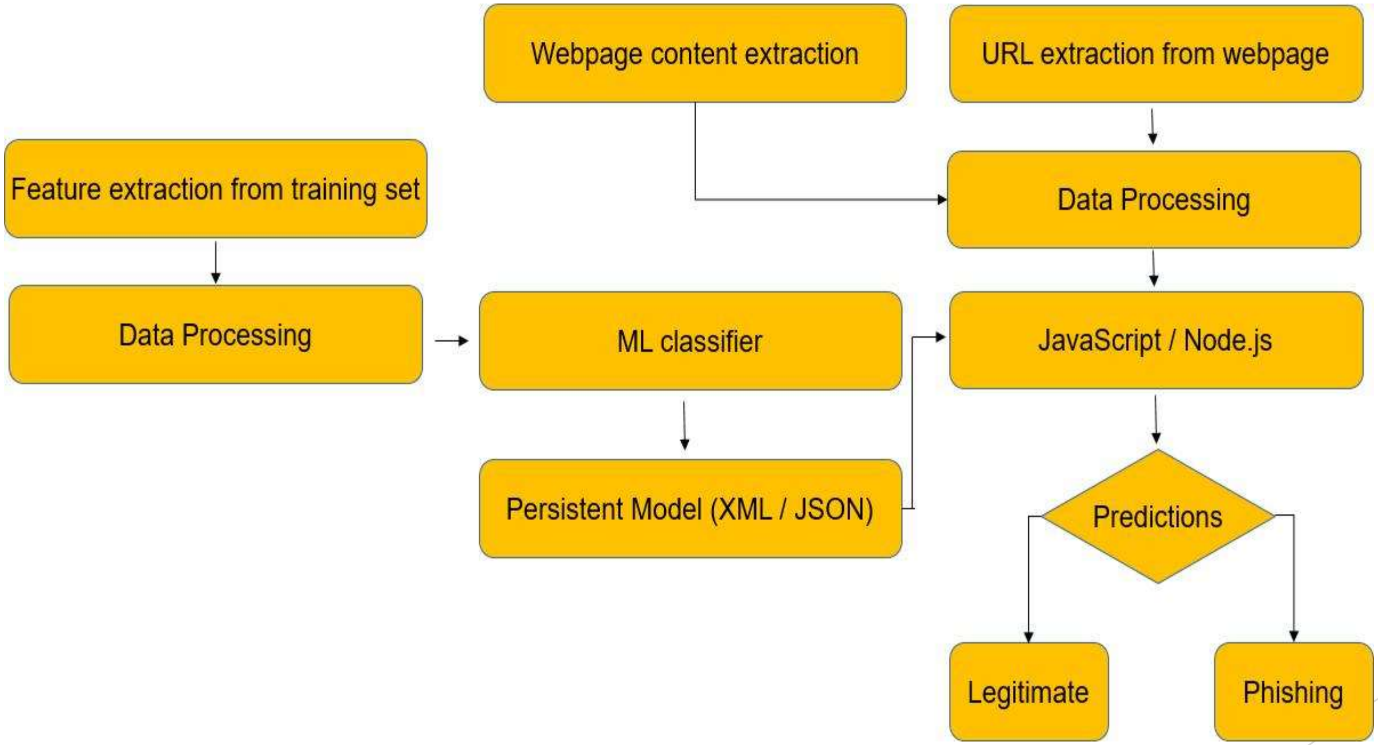
• **Modeling**

After the data is ready to be processed , modeling process for the learning algorithm is initiated. The model is basically the construction of the need for output identified in accordance with the task qualifications.

ADVANTAGES

- This study is considered to be an applicable design in automated systems with high performing classification against the phishing activity of websites.
- Furthermore, in literature comparisons, this study is observed to be high-performing by having a high performance.

ARCHITECTURE


Webpage content extraction
URL extraction from webpage
Feature extraction from training set
Data Processing
Data Processing
ML classifier
JavaScript / Node.js
Persistent Model (XML / JSON)
Predictions
Legitimate
Phishing

IMPLEMENTATION

Below mentioned are the steps involved in the completion of this project:

- Collect dataset containing phishing and legitimate websites from the open source platforms.
- Write a code to extract the required features from the URL database.
- Analyze and preprocess the dataset by using EDA techniques.
- Divide the dataset into training and testing sets.
- Run selected machine learning algorithms like SVM, Random Forest and neural networks on the dataset.
- Write a code for displaying the evaluation result considering accuracy metrics.
- Compare the obtained results for trained models and specify which is better.

UML DIAGRAMS

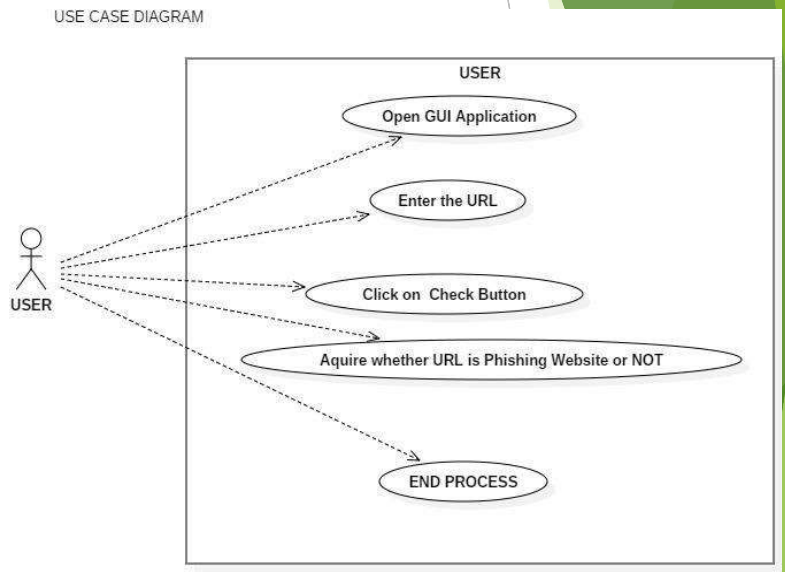
The Unified Modeling Language allows the software engineer to express an analysis model using the modeling notation that is governed by a set of syntactic semantic and pragmatic rules.

A UML system is represented using five different views that describe the system from distinctly different perspective.

1. Use case Diagram
2. Class Diagram
3. Sequence Diagram
4. Activity Diagram

USECASE DIAGRAM

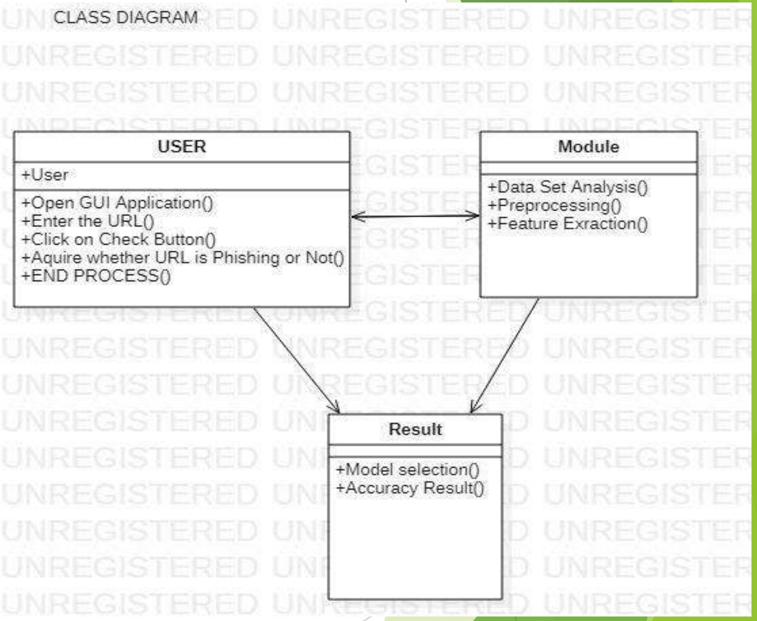
A use case diagram at its simplest is a representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system. This type of diagram is typically used in conjunction with the textual use case and will often be accompanied by other types of diagrams as well.



CLASS DIAGRAM

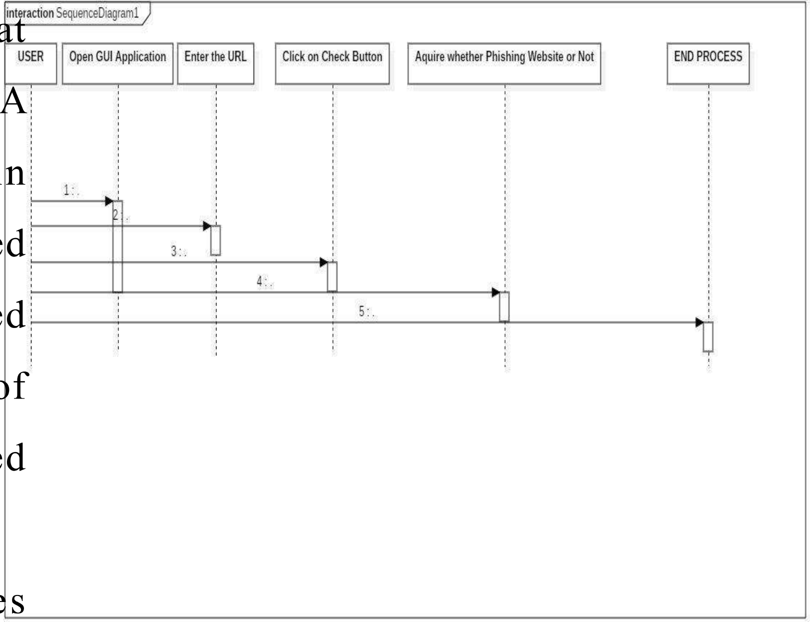
The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for data modeling A class with three sections, in the diagram, classes is represented with boxes which contain three parts:

1. The upper part holds the name of the class
2. The middle part contains the attributes of the class
3. The bottom part gives the methods or operations the class can take or undertake.



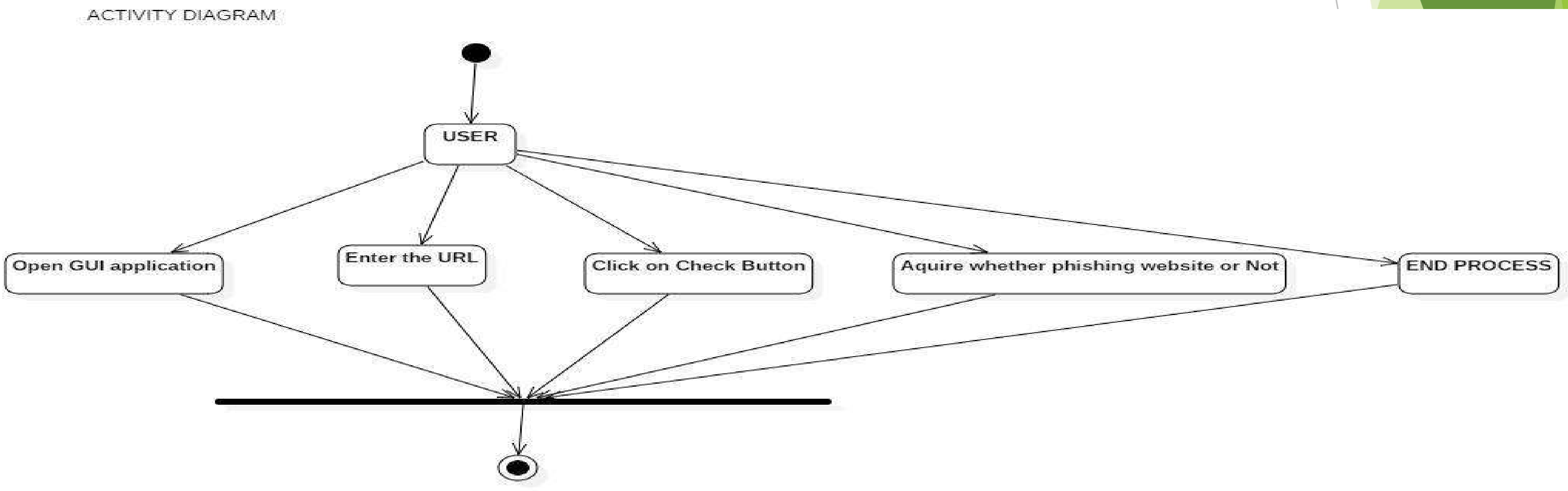
SEQUENCE DIAGRAM

A sequence diagram is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



SYSTEM TESTING

MODULE TESTING

To locate errors, each module is tested individually. This enables us to detect error and correct it without affecting any other modules. Whenever the program is not satisfying the required function, it must be corrected to get the required result. Thus all the modules are individually tested from bottom up starting with the smallest and lowest modules and proceeding to the next level. Each module in the system is tested separately. For example the job classification module is tested separately. This module is tested with different job and its approximate execution time and the result of the test is compared with the results that are prepared manually. Each module in the system is tested separately. In this system the resource classification and job scheduling modules are tested separately and their corresponding results are obtained which reduces the process waiting time.

INTEGRATION TESTING

After the module testing, the integration testing is applied. When linking the modules there may be chance for errors to occur, these errors are corrected by using this testing. In this system all modules are connected and tested. The testing results are very correct. Thus the mapping of jobs with resources is done correctly by the system.

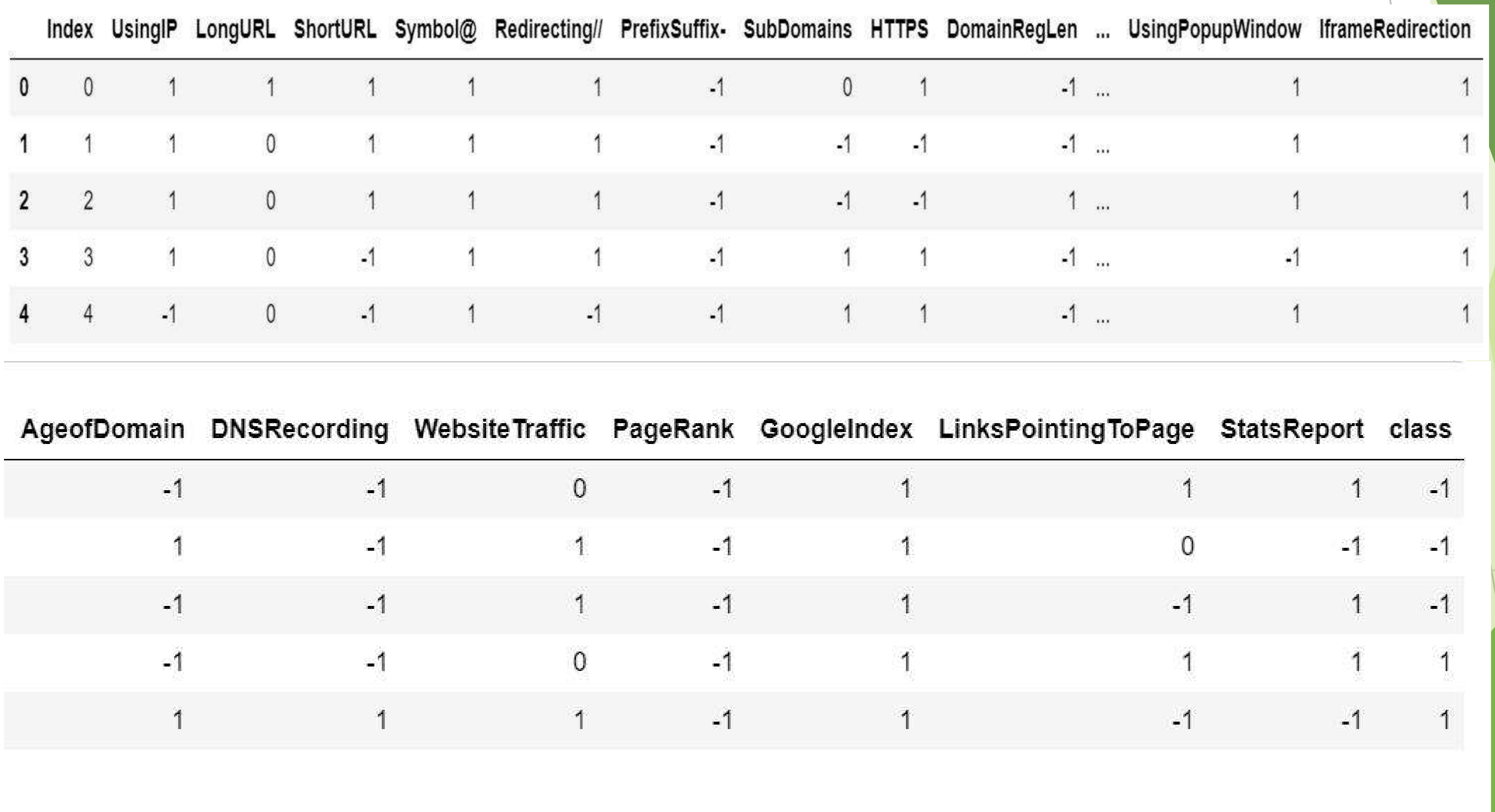
ACCEPTANCE TESTING

When that user fined no major problems with its accuracy, the system passers through a final acceptance test. This test confirms that the system needs the original goals, objectives and requirements established during analysis without actual execution which elimination wastage of time and money acceptance tests on the shoulders of users and management, it is finally acceptable and ready for the operation.

TC ID	Condition Being Tested	Expected Result	Result
1	Check if dataset is taken as input	If taken process next step else throw error	passed
2	check for null values in the dataset	If null values drop the null records	Passed
3	Check for future extraction and save thefile in. pkl format	If file not saved throw the error	passed

Results

Data loading and Pre-processing:

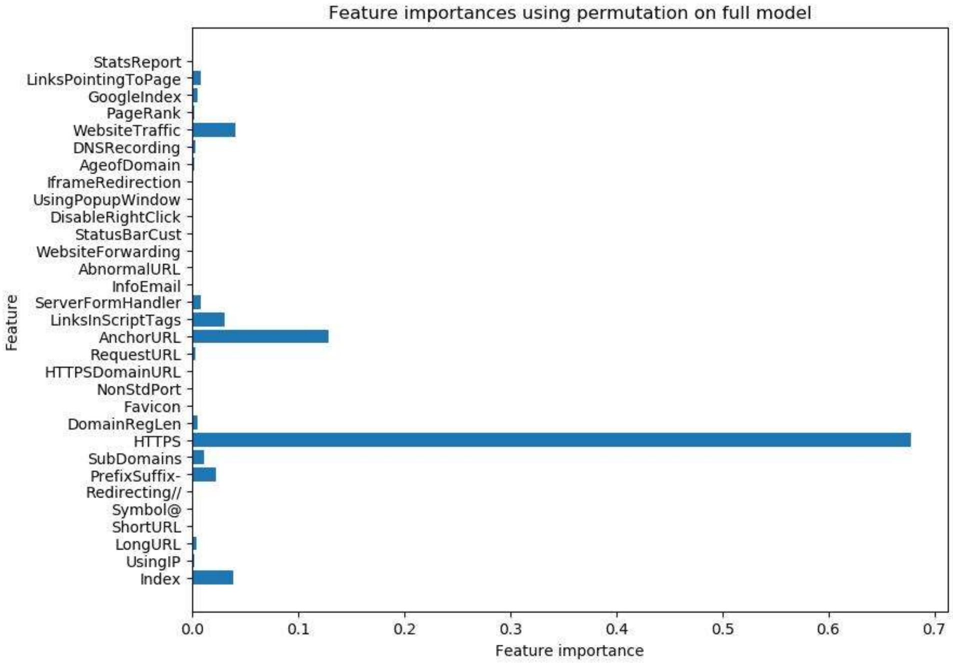


	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection
0	0	1	1	1	1	1	-1	0	1	-1	...	1	1
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	1
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	1
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	1
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1

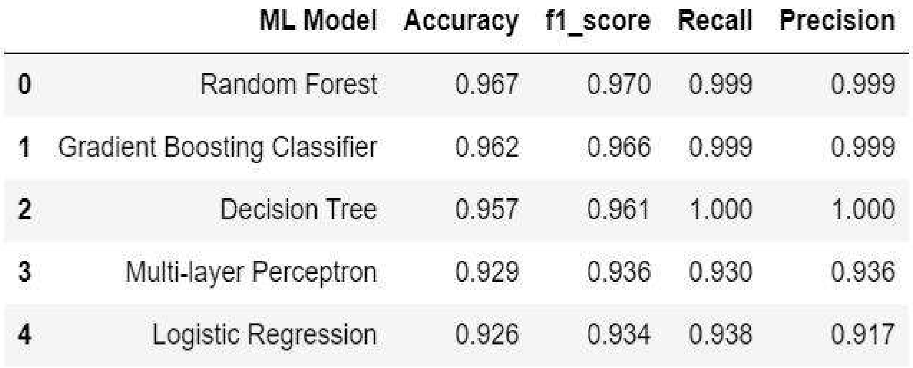
AgeofDomain	DNSRecording	WebsiteTraffic	PageRank	GoogleIndex	LinksPointingToPage	StatsReport	class
-1	-1	0	-1	1	1	1	-1
1	-1	1	-1	1	0	-1	-1
-1	-1	1	-1	1	-1	1	-1
-1	-1	0	-1	1	1	1	1
1	1	1	-1	1	-1	-1	1

Obtaining Feature Importance

The bar graph representation of the feature importance, these features are further utilized to train the ML models for detecting Phishing URL’s.



ML models accuracy

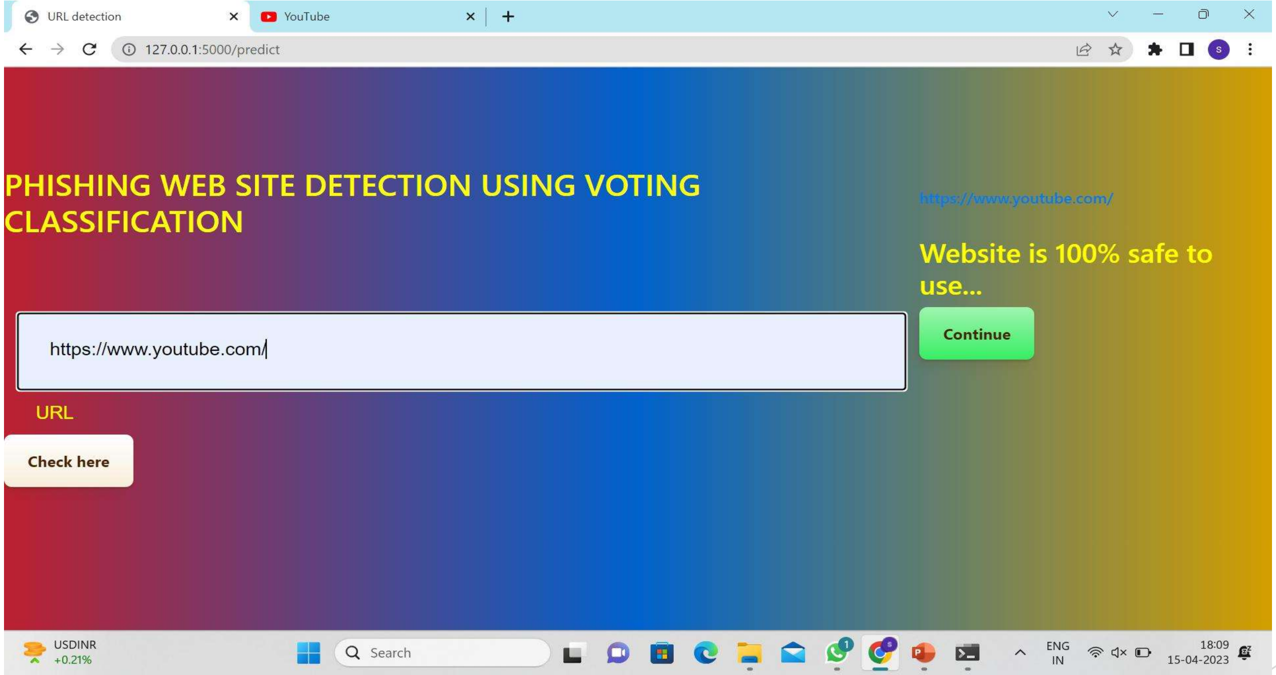


	ML Model	Accuracy	f1_score	Recall	Precision
0	Random Forest	0.967	0.970	0.999	0.999
1	Gradient Boosting Classifier	0.962	0.966	0.999	0.999
2	Decision Tree	0.957	0.961	1.000	1.000
3	Multi-layer Perceptron	0.929	0.936	0.930	0.936
4	Logistic Regression	0.926	0.934	0.938	0.917

Legitimacy of Input Text

Output-1

The manual input text is provided i.e URL and the output is determined to be Phishing as the URL is determined to be having feature of the phished website by the trained ML models.



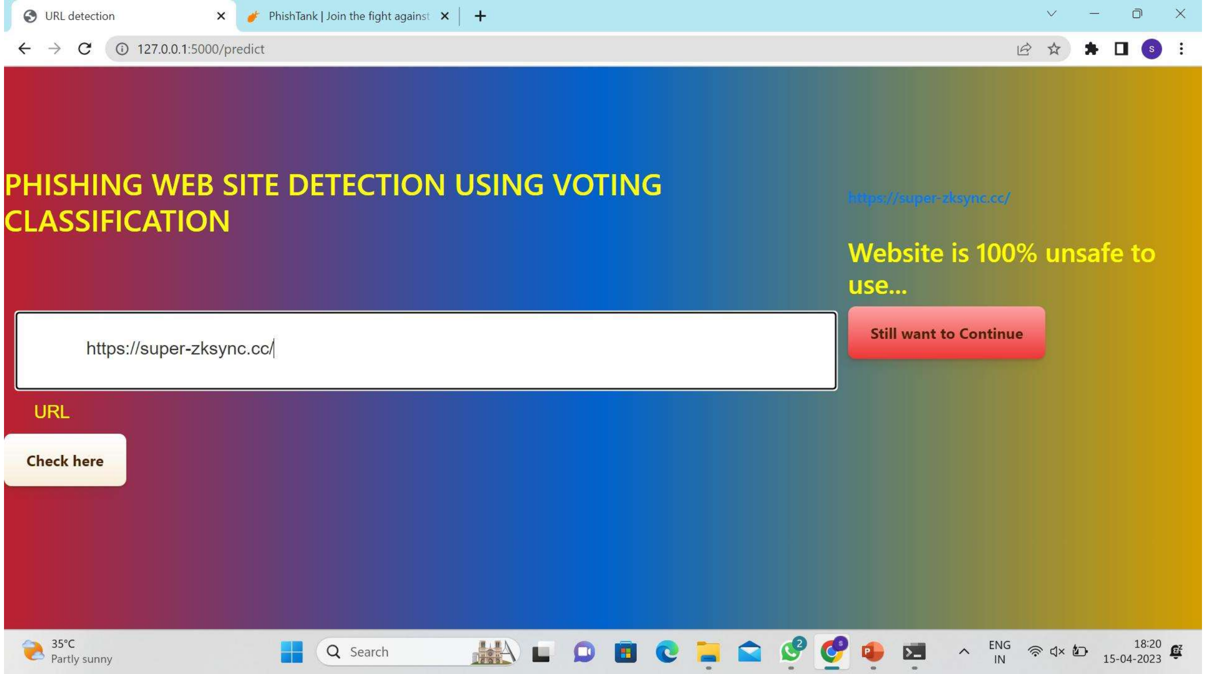
PHISHING WEB SITE DETECTION USING VOTING CLASSIFICATION

https://www.youtube.com/

Website is 100% safe to use...

URL: https://www.youtube.com/

Output-2



PHISHING WEB SITE DETECTION USING VOTING CLASSIFICATION

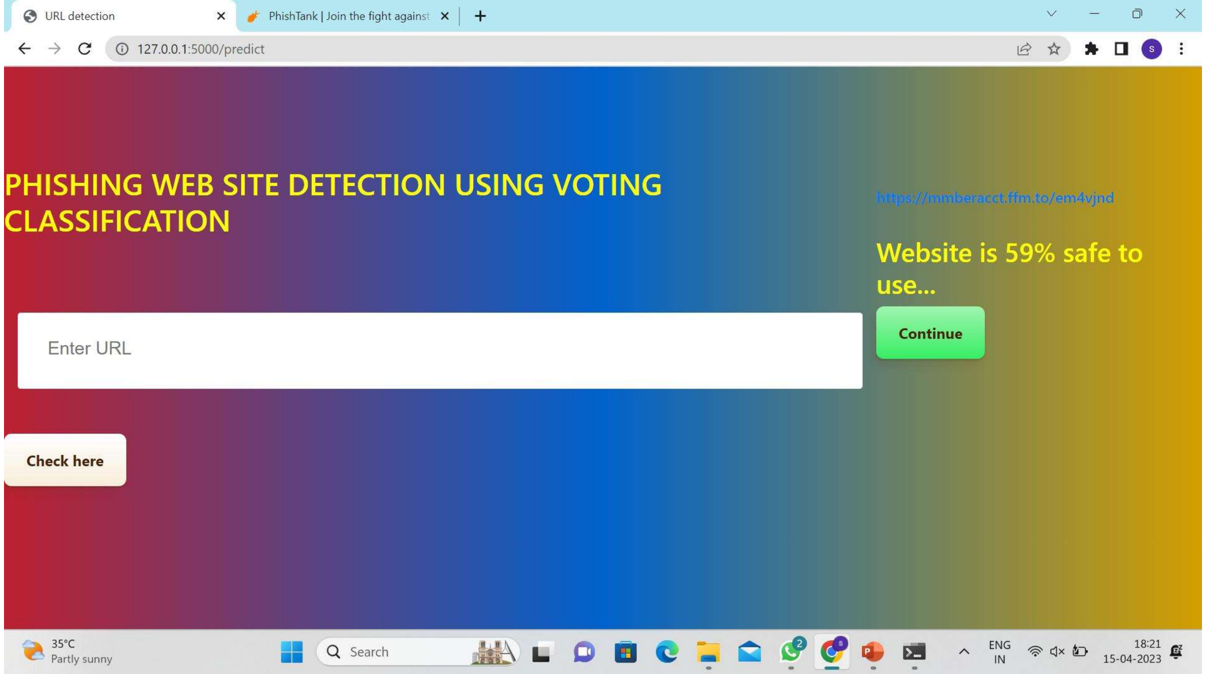
https://super-zksync.cc/

URL

https://super-zksync.cc/

Website is 100% unsafe to use...

Output-3



PHISHING WEB SITE DETECTION USING VOTING CLASSIFICATION

https://mmberacct.ffm.to/em4vjnd

Website is 59% safe to use...

CONCLUSION

Phishing website attacks are a massive challenge for researchers, and they continue to show a rising trend in recent years. Blacklist/whitelist techniques are the traditional way to alleviate such threats. However, these methods fail to detect non-blacklisted phishing websites (i.e., 0-day attacks). As an improvement, machine learning techniques are being used to increase detection efficiency and reduce the misclassification ratio. However, some of them extract features from third-party services, search engines, website traffic, etc., which are complicated and difficult to access. We discussed question generation from text and proposed that, in addition to the typical focus of such work on meaning and understanding, questions can also play an important role for functionally-driven input enhancement. In line with the focus-on-form perspective in Second Language Acquisition research and the notion of structured input activities, such questions help the learner in processing relevant forms and draw form-meaning connections while engaging in a meaning-based activity. We proposed two types of questions designed to provide functionally-driven .