Low Level Design (LLD)

Phishing Domain Detection

Revision Number: 1
Last date of revision: 10/05/2023

Karan Singh

Document Version Control

Date Issued	Version	Description	Author
10th May 2023	1.1	First Draft	Karan Singh

Table of Contents

Document Version Control 1
Abstract 3
1.Introduction 4
1.1 Why this Low-Level Design Document? 4
1.2 Scope 4
1.3 Constraints..**Error! Bookmark not defined.**
1.4 Risks..**Error! Bookmark not defined.**
1.5 Out of Scope ....**Error! Bookmark not defined.**
2.Technical specifications 7
2.1 Dataset Name and Source 7
2.2 Dataset overview 7
2.3 Data Description..8
2.4 Logging..**Error! Bookmark not defined.**
2.5 Database..**Error! Bookmark not defined.**
3. Deployment ....**Error! Bookmark not defined.**

Abstract

Phishing is a type of fraud in which an attacker impersonates a reputable company or person in order to get sensitive information such as login credentials or account information via email or other communication channels. Phishing is popular among attackers because it is easier to persuade someone to click a malicious link that appears to be authentic than it is to break through a computer's protection measures The mail goal of this end to end application is to predict whether the domains are real or malicious using various combination of machine learning algorithms by dividing the dataset into various clusters and applying a suitable algorithm for that particular cluster.

1. Introduction

1.1 Why this Low-Level Design Document?

The purpose of this Low-Level Design (LLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The LLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:

  ○ Security ○ Reliability ○
  Maintainability ○ Portability
  ○ Reusability ○ Application
  compatibility ○ Resource
  utilization ○ Serviceability

1.2 Scope

The LLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The LLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system. This software system will be a Web application this system will be designed to detect malicious websites.

2. System Architecture

The system architecture of the Phishing Domain Detection project consists of the following components:

2.1 Data Preprocessing

The data preprocessing module is responsible for collecting and cleaning the data used to train the machine learning models. This module will retrieve the data from the sources, remove the duplicates, and normalize it for further processing.

2.2 Feature Engineering

The feature engineering module is responsible for extracting and transforming the features from the preprocessed data. This module will create a feature matrix that will be used as input to the machine learning models.

2.3 Machine Learning Models

The machine learning models module is responsible for training and testing the models that will be used for phishing domain detection. This module includes a variety of models, including logistic regression, decision trees, and neural networks.

2.4 Model Evaluation

The model evaluation module is responsible for evaluating the performance of the machine learning models. This module will provide metrics such as precision, recall, and F1 score to assess the model's performance.

3. Software Architecture

The software architecture of the Phishing Domain Detection project is designed to be modular and flexible. The system uses a microservice architecture that allows for easy scalability and maintenance.

The following modules are used in the system:

3.1 Data Collection

The data collection module retrieves the data used for training the machine learning models. This module includes data sources such as WHOIS records and DNS records.

3.2 Data Preprocessing

The data preprocessing module cleans and normalizes the data before it is used for feature engineering.

3.3 Feature Engineering

The feature engineering module extracts and transforms the features from the preprocessed data. This module includes feature extraction techniques such as bag of words and term frequency–inverse document frequency (TF-IDF).

3.4 Machine Learning Models

The machine learning models module includes a variety of models used for phishing domain detection. This module includes models such as Random Forest, decision trees, and neural networks.

3.5 Model Evaluation

The model evaluation module evaluates the performance of the machine learning models. This module provides metrics such as precision, recall

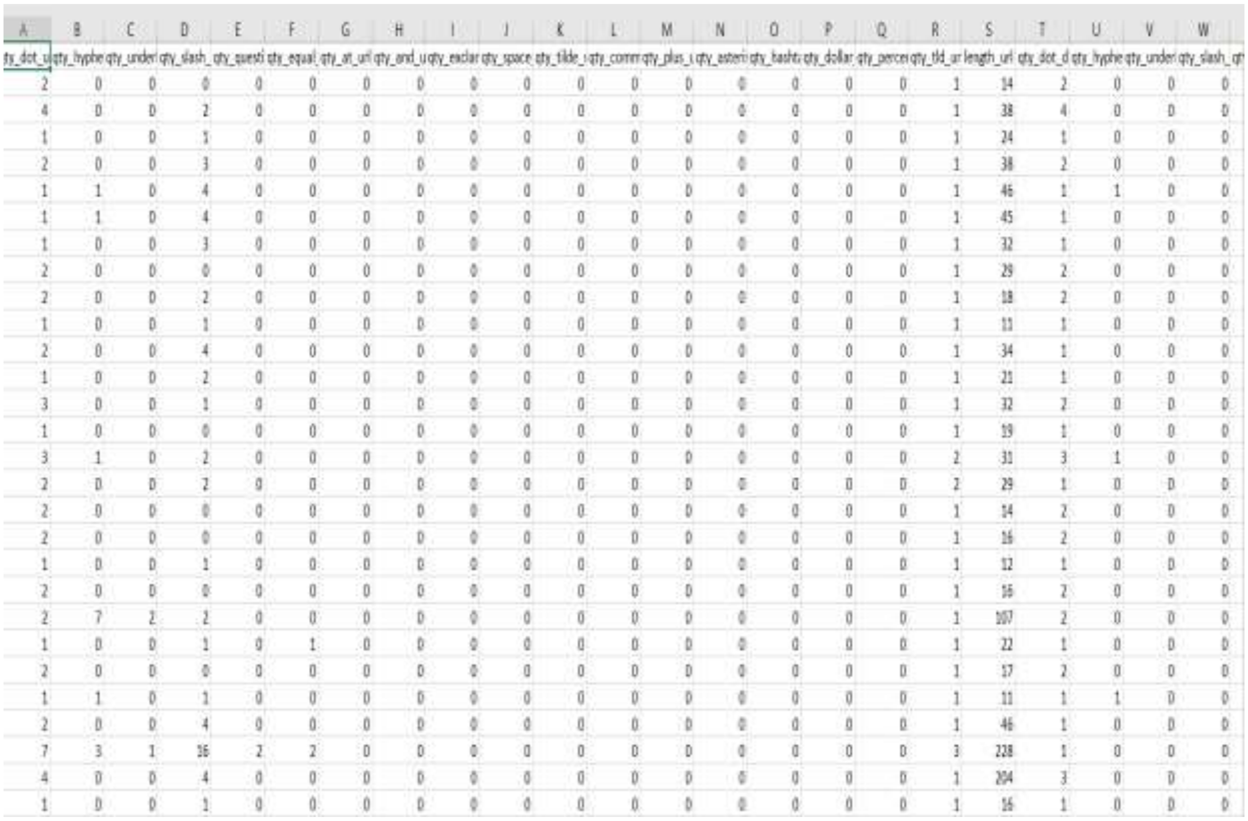
4.Technical specifications

4.1 Dataset Name and Source

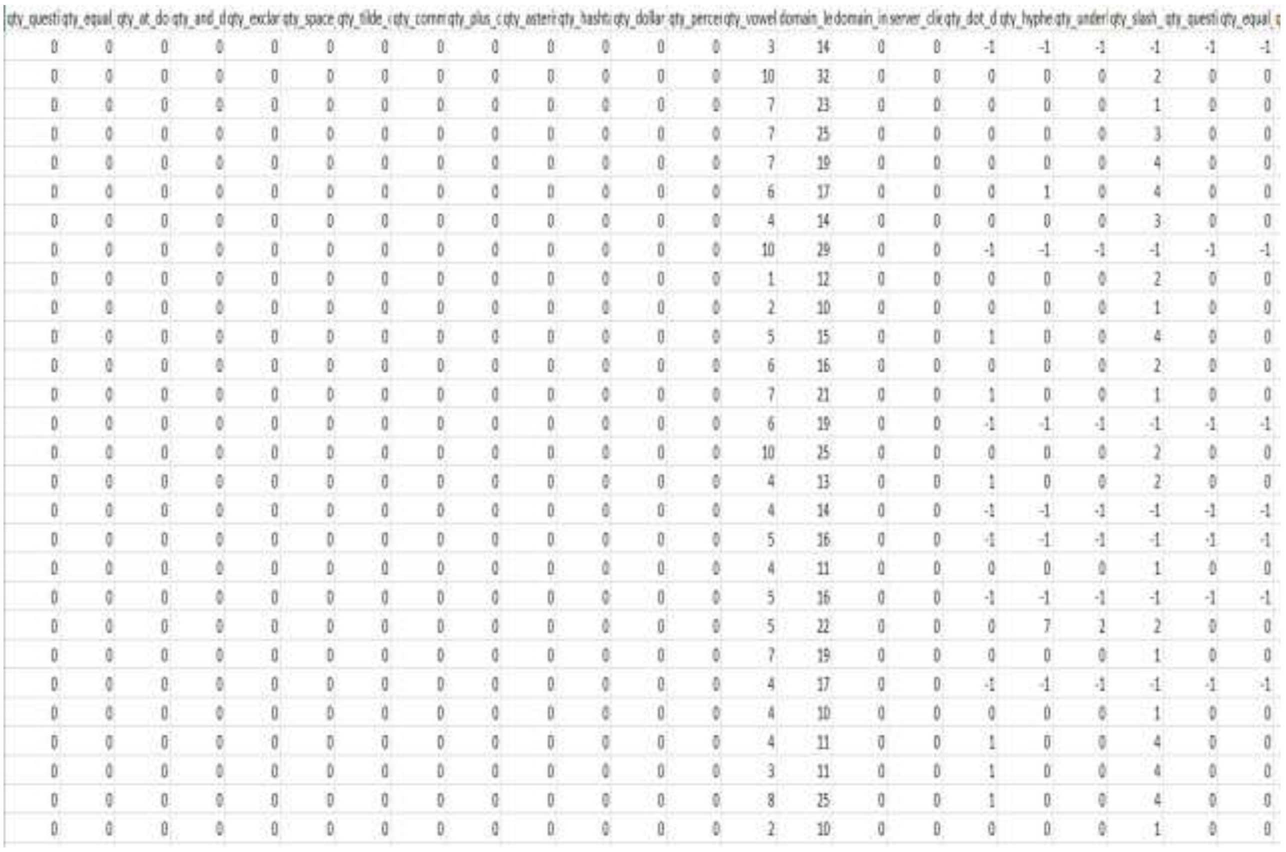
Data Set Name	Finalized	Source
Dataset_small.csv	Yes	Phishing Websites Dataset - Mendeley Data

4.2 Dataset overview

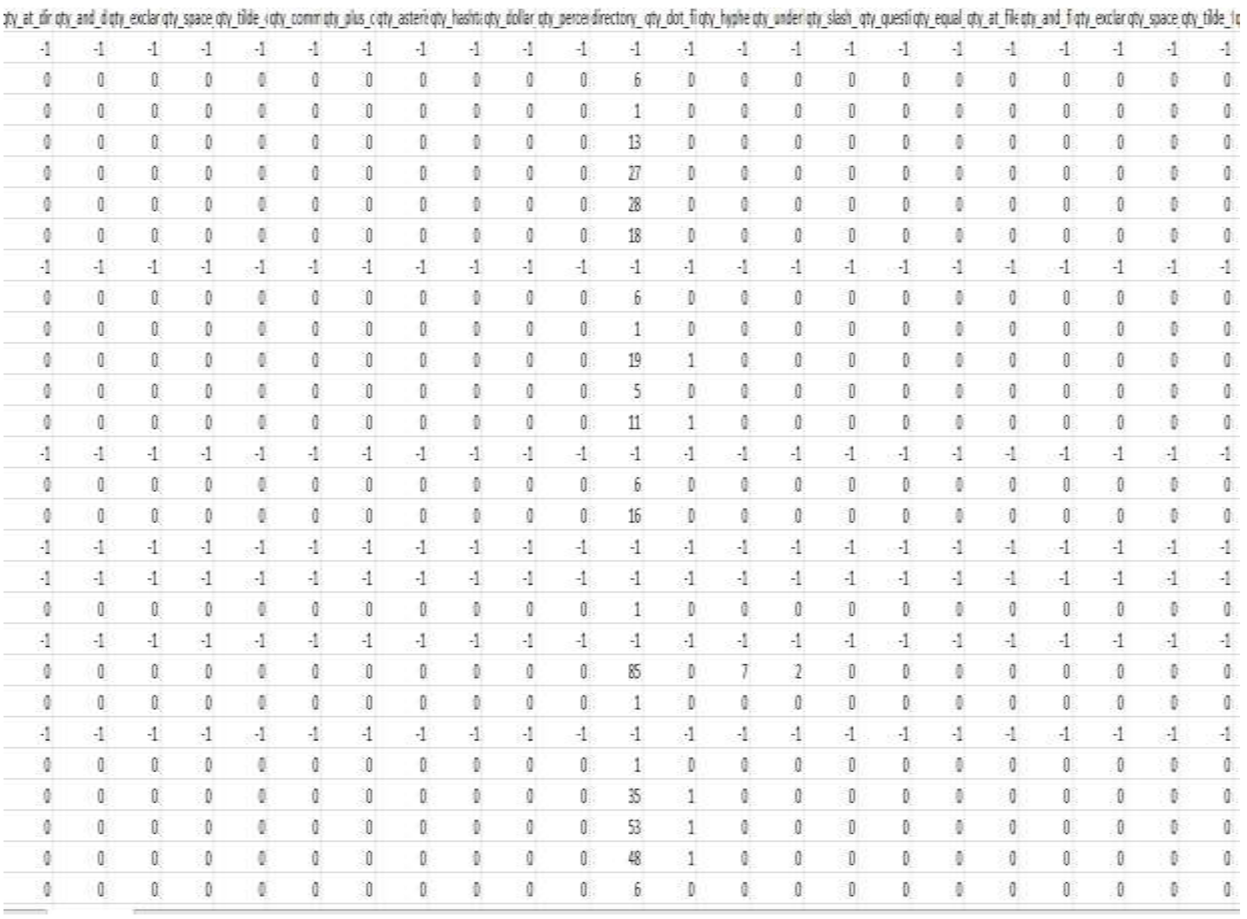
- These data consist of a collection of legitimate as well as phishing website instances. Each website is represented by the set of features which denote, whether website is legitimate or not. Data can serve as an input for machine learning process. In this repository the two variants of the Phishing Dataset are presented. Full variant - dataset_full.csv Short description of the full variant dataset: Total number of instances: 88,647 Number of legitimate website instances (labelled as 0): 58,000 Number of phishing website instances (labelled as 1): 30,647 Total number of features: 111



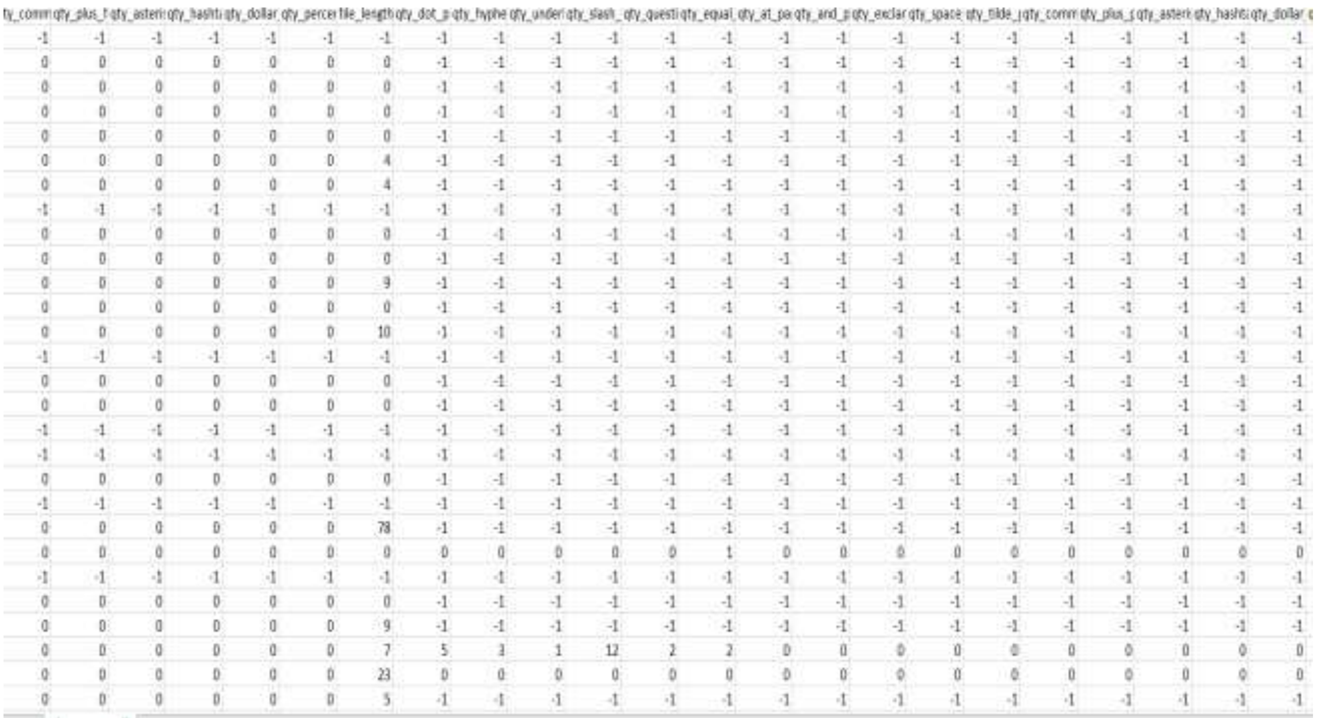
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
ty_dot_u	qty_hyphe	qty_underl	qty_slash_	qty_questi	qty_equal_	qty_at_url	qty_and_u	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_	qty_asteri	qty_hasht	qty_dollar	qty_percer	qty_tld_ur	length_url	qty_dot_d	qty_hyphe	qty_under	qty_slash_
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	2	0	0	0
4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	38	4	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	24	1	0	0	0
2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	38	2	0	0	0
1	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	1	1	0	0
1	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	45	1	0	0	0
1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	32	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	29	2	0	0	0
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	18	2	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	1	0	0	0
2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	34	1	0	0	0
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	1	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	32	2	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	1	0	0	0
3	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	31	3	1	0	0
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	29	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	2	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2	0	0	0
2	7	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	107	2	0	0	0
1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	22	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	2	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	1	1	0	0
2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	1	0	0	0
7	3	1	16	2	2	0	0	0	0	0	0	0	0	0	0	0	3	228	1	0	0	0
4	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	204	3	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	1	0	0	0



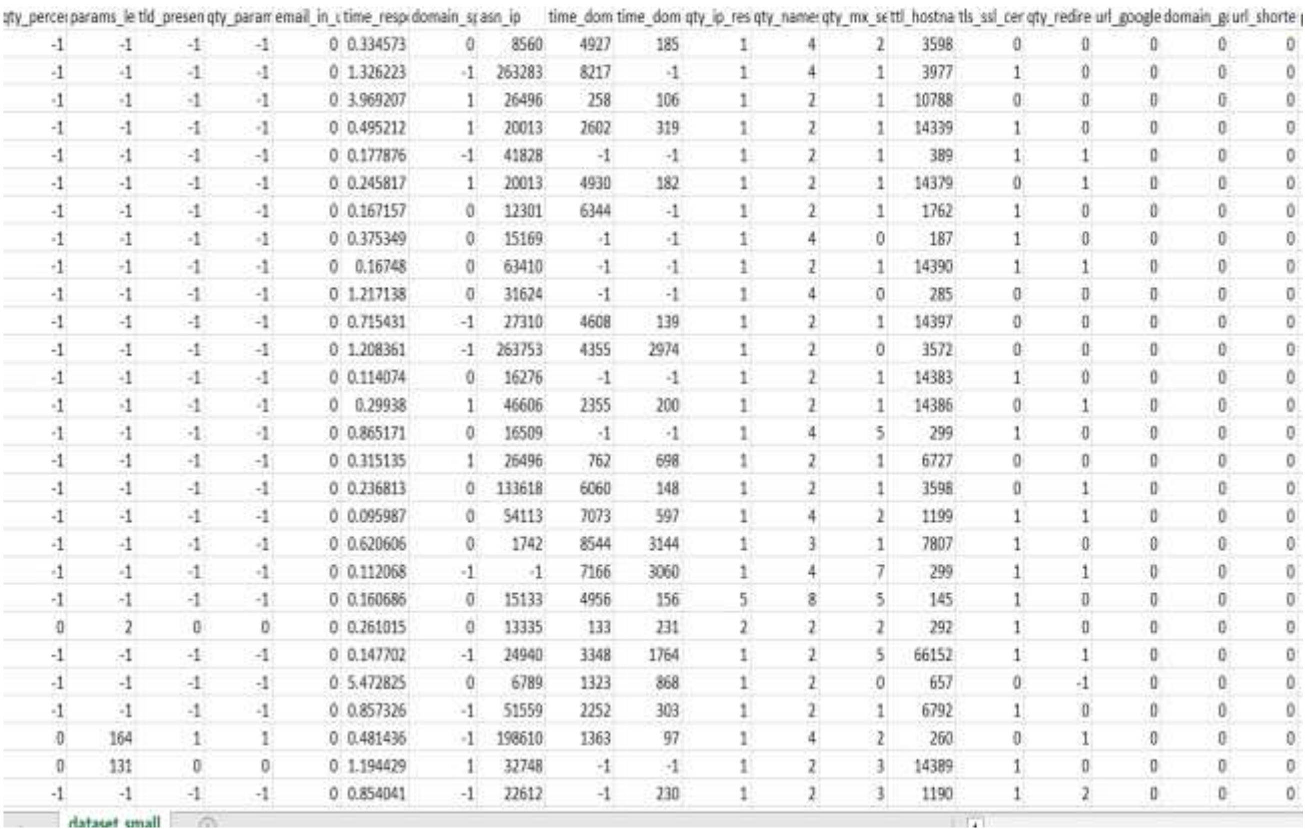
qty_questi	qty_equal_	qty_at_do	qty_and_d	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_c	qty_asteri	qty_hasht	qty_dollar	qty_percei	qty_vowel	domain_le	domain_in	server_clie	qty_dot_d	qty_hyphe	qty_underl	qty_slash_	qty_questi	qty_equal_
0	0	0	0	0	0	0	0	0	0	0	0	0	3	14	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	10	32	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	7	23	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	7	25	0	0	0	0	0	3	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	7	19	0	0	0	0	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	6	17	0	0	0	1	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	14	0	0	0	0	0	3	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	10	29	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	12	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	5	15	0	0	1	0	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	6	16	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	7	21	0	0	1	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	6	19	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	10	25	0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	13	0	0	1	0	0	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	14	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	5	16	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	4	11	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	5	16	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	5	22	0	0	0	7	2	2	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	7	19	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	17	0	0	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	0	4	10	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	4	11	0	0	1	0	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	3	11	0	0	1	0	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	8	25	0	0	1	0	0	4	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	2	10	0	0	0	0	0	1	0	0



qty_at_dir	qty_and_d	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_c	qty_asteri	qty_hasht	qty_dollar	qty_percei	directory_	qty_dot_fi	qty_hyphe	qty_underl	qty_slash_	qty_questi	qty_equal_	qty_at_file	qty_and_fi	qty_exclar	qty_space	qty_tilde_
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	19	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	11	1	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	85	0	7	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	35	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	53	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	48	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0



ty_comm	qty_plus_f	qty_asteri	qty_hasht	qty_dollar	qty_percer	file_length	qty_dot_p	qty_hyphe	qty_underl	qty_slash_	qty_questi	qty_equal_	qty_at_pa	qty_and_p	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_p	qty_asterk	qty_hasht	qty_dollar
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	78	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	7	5	3	1	12	2	2	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1



qty_perce	params_le	tld_presen	qty_param	email_in_	time_resp	domain_sp	asn_ip	time_dom	time_dom	qty_ip_res	qty_name	qty_mx_se	ttl_hostna	tls_ssl_cer	qty_redire	url_google	domain_g	url_shorte
-1	-1	-1	-1	0	0.334573	0	8560	4927	185	1	4	2	3598	0	0	0	0	0
-1	-1	-1	-1	0	1.326223	-1	263283	8217	-1	1	4	1	3977	1	0	0	0	0
-1	-1	-1	-1	0	3.969207	1	26496	258	106	1	2	1	10788	0	0	0	0	0
-1	-1	-1	-1	0	0.495212	1	20013	2602	319	1	2	1	14339	1	0	0	0	0
-1	-1	-1	-1	0	0.177876	-1	41828	-1	-1	1	2	1	389	1	1	0	0	0
-1	-1	-1	-1	0	0.245817	1	20013	4930	182	1	2	1	14379	0	1	0	0	0
-1	-1	-1	-1	0	0.167157	0	12301	6344	-1	1	2	1	1762	1	0	0	0	0
-1	-1	-1	-1	0	0.375349	0	15169	-1	-1	1	4	0	187	1	0	0	0	0
-1	-1	-1	-1	0	0.16748	0	63410	-1	-1	1	2	1	14390	1	1	0	0	0
-1	-1	-1	-1	0	1.217138	0	31624	-1	-1	1	4	0	285	0	0	0	0	0
-1	-1	-1	-1	0	0.715431	-1	27310	4608	139	1	2	1	14397	0	0	0	0	0
-1	-1	-1	-1	0	1.208361	-1	263753	4355	2974	1	2	0	3572	0	0	0	0	0
-1	-1	-1	-1	0	0.114074	0	16276	-1	-1	1	2	1	14383	1	0	0	0	0
-1	-1	-1	-1	0	0.29938	1	46606	2355	200	1	2	1	14386	0	1	0	0	0
-1	-1	-1	-1	0	0.865171	0	16509	-1	-1	1	4	5	299	1	0	0	0	0
-1	-1	-1	-1	0	0.315135	1	26496	762	698	1	2	1	6727	0	0	0	0	0
-1	-1	-1	-1	0	0.236813	0	133618	6060	148	1	2	1	3598	0	1	0	0	0
-1	-1	-1	-1	0	0.095987	0	54113	7073	597	1	4	2	1199	1	1	0	0	0
-1	-1	-1	-1	0	0.620606	0	1742	8544	3144	1	3	1	7807	1	0	0	0	0
-1	-1	-1	-1	0	0.112068	-1	-1	7166	3060	1	4	7	299	1	1	0	0	0
-1	-1	-1	-1	0	0.160686	0	15133	4956	156	5	8	5	145	1	0	0	0	0
0	2	0	0	0	0.261015	0	13335	133	231	2	2	2	292	1	0	0	0	0
-1	-1	-1	-1	0	0.147702	-1	24940	3348	1764	1	2	5	66152	1	1	0	0	0
-1	-1	-1	-1	0	5.472825	0	6789	1323	868	1	2	0	657	0	-1	0	0	0
-1	-1	-1	-1	0	0.857326	-1	51559	2252	303	1	2	1	6792	1	0	0	0	0
0	164	1	1	0	0.481436	-1	198610	1363	97	1	4	2	260	0	1	0	0	0
0	131	0	0	0	1.194429	1	32748	-1	-1	1	2	3	14389	1	0	0	0	0
-1	-1	-1	-1	0	0.854041	-1	22612	-1	230	1	2	3	1190	1	2	0	0	0

dataset_small

4.3 Data Description

The presented dataset was collected and prepared for the purpose of building and evaluating various classification methods for the task of detecting phishing websites based on the uniform resource locator (URL) properties, URL resolving metrics, and external services. The attributes of the prepared dataset can be divided into six groups:

1) attributes based on the whole URL properties presented in table 1.

Nr.	Attribute	Format	Description
1	qty_dot_url	Number of ”.” signs	Numeric
2	qty_hyphen_url	Number of ”-” signs	Numeric
3	qty_underline_url	Number of ”_” signs	Numeric
4	qty_slash_url	Number of ”/” signs	Numeric
5	qty_questionmark_url	Number of ”?” signs	Numeric
6	qty_equal_url	Number of ”=” sings	Numeric
7	qty_at_url	Number of ”@” signs	Numeric
8	qty_and_url	Number of ”&” signs	Numeric
9	qty_exclamation_url	Number of ”!” signs	Numeric
10	qty_space_url	Number of ” ” signs	Numeric
11	qty_tilde_url	Number of ”′˜′signs	Numeric
12	qty_comma_url	Number of ”,” signs	Numeric
13	qty_plus_url	Number of ”+” signs	Numeric
14	qty_asterisk_url	Number of ”*” signs	Numeric
15	qty_hashtag_url	Number of ”#” signs	Numeric
16	qty_dollar_url	Number of ”$” signs	Numeric
17	qty_percent_url	Number of ”%” signs	Numeric
18	qty_tld_url	Top level domain character length	Numeric
19	length_url	Number of characters	Numeric
20	email_in_url	Is email present	Boolean

2) Attributes based on the domain properties presented in Table 2.

Nr	Attribute	Format	Description
1	qty_dot_domain	Number of ”.” signs	Numeric
2	qty_hyphen_domain	Number of ”-” signs	Numeric
3	qty_underline_domain	Number of ”_” signs	Numeric
4	qty_slash_domain	Number of ”/” signs	Numeric
5	qty_questionmark_domain	Number of ”?” signs	Numeric
6	qty_equal_domain	Number of ”=” signs	Numeric
7	qty_at_domain	Number of ”@” signs	Numeric
8	qty_and_domain	Number of ”&” signs	Numeric
9	qty_exclamation_domain	Number of ”!” signs	Numeric
10	qty_space_domain	Number of ” ” signs	Numeric
11	qty_tilde_domain	Number of ”signs	Numeric

12	qty_comma_domain	Number of ”,” signs	Numeric
13	qty_plus_domain	Number of ”+” signs	Numeric
14	qty_asterisk_domain	Number of ”*” signs	Numeric
15	qty_hashtag_domain	Number of ”#” signs	Numeric
16	qty_dollar_domain	Number of ”$” signs	Numeric
17	qty_percent_domain	Number of ”%” signs	Numeric
18	qty_vowels_domain	Number of vowels	Numeric
19	domain_length	Number of domain characters	Numeric
20	domain_in_ip	URL domain in IP address format	Boolean
21	server_client_domain	”server” or ”client” in domain	Boolean

3) attributes based on the URL directory properties presented in Table 3

Nr	Attribute	Format	Description
1	qty_dot_directory	Number of ”.” signs	Numeric
2	qty_hyphen_directory	Number of ”-” signs	Numeric
3	qty_underline_directory	Number of ”_” signs	Numeric
4	qty_slash_directory	Number of ”/” signs	Numeric
5	qty_questionmark_directory	Number of ”?” signs	Numeric
6	qty_equal_directory	Number of ”=” signs	Numeric
7	qty_at_directory	Number of ”@” signs	Numeric
8	qty_and_directory	Number of ”&” signs	Numeric
9	qty_exclamation_directory	Number of ”!” signs	Numeric
10	qty_space_directory	Number of ” ” signs	Numeric
11	qty_tilde_directory	Number of ”signs	Numeric
12	qty_comma_directory	Number of ”,” signs	Numeric
13	qty_plus_directory	Number of ”+” signs	Numeric
14	qty_asterisk_directory	Number of ”*” signs	Numeric
15	qty_hashtag_directory	Number of ”#” signs	Numeric
16	qty_dollar_directory	Number of ”$” signs	Numeric
17	qty_percent_directory	Number of ”%” signs	Numeric
18	directory_length	Number of directory characters	Numeric

4) attributes based on the URL file properties presented in Table 4

Nr	Attribute	Format	Description
1	qty_dot_file	Number of ”.” signs	Numeric
2	qty_hyphen_file	Number of ”-” signs	Numeric
3	qty_underline_file	Number of ”_” signs	Numeric
4	qty_slash_file	Number of ”/” signs	Numeric
5	qty_questionmark_file	Number of ”?” signs	Numeric
6	qty_equal_file	Number of ”=” signs	Numeric
7	qty_at_file	Number of ”@” signs	Numeric
8	qty_and_file	Number of ”&” signs	Numeric
9	qty_exclamation_file	Number of ”!” signs	Numeric
10	qty_space_file	Number of ” ” signs	Numeric
11	qty_tilde_file	Number of ”signs	Numeric
12	qty_comma_file	Number of ”,” signs	Numeric
13	qty_plus_file	Number of ”+” signs	Numeric
14	qty_asterisk_file	Number of ”*” signs	Numeric
15	qty_hashtag_file	Number of ”#” signs	Numeric
16	qty_dollar_file	Number of ”$” signs	Numeric
17	qty_percent_file	Number of ”%” signs	Numeric
18	file_length	Number of file name characters	Numeric

5) attributes based on the URL parameter properties presented in Table 5

Nr	Attribute	Format	Description
1	qty_dot_params	Number of ”.” signs	Numeric
2	qty_hyphen_params	Number of ”-” signs	Numeric
3	qty_underline_params	Number of ”_” signs	Numeric
4	qty_slash_params	Number of ”/” signs	Numeric
5	qty_questionmark_params	Number of ”?” signs	Numeric
6	qty_equal_params	Number of ”=” signs	Numeric
7	qty_at_params	Number of ”@” signs	Numeric
8	qty_and_params	Number of ”&” signs	Numeric
9	qty_exclamation_params	Number of ”!” signs	Numeric
10	qty_space_params	Number of ” ” signs	Numeric
11	qty_tilde_params	Number of ”signs	Numeric
12	qty_comma_params	Number of ”,” signs	Numeric
13	qty_plus_params	Number of ”+” signs	Numeric
14	qty_asterisk_params	Number of ”*” signs	Numeric
15	qty_hashtag_params	Number of ”#” signs	Numeric

16	qty_dollar_params	Number of ”$” signs	Numeric
17	qty_percent_params	Number of ”%” signs	Numeric
18	params_length	Number of parameters characters	Numeric
19	tld_present_params	TLD present in parameters	Boolean
20	qty_params	Number of parameters	Numeric

6. Attributes based on the URL resolving data and external metrics presented in Table 6.

Nr	Attribute	Format	Description
1	time_response	Domain lookup time response	Numeric
2	domain_spf	Domain has SPF	Boolean
3	asn_ip	ASN	Numeric
4	time_domain_activation	Domain activation time (in days)	Numeric
5	time_domain_expiration	Domain expiration time (in days)	Numeric
6	qty_ip_resolved	Number of resolved IPs	Numeric
7	qty_nameservers	Number of resolved NS4	Numeric
8	qty_mx_servers	Number of MX 5servers	Numeric
9	ttl_hostname	Time-To-Live (TTL)	Numeric
10	tls_ssl_certificate	Has valid TLS 6/SSL 7certificate	Numeric
11	qty_redirects	Number of redirects	Boolean
12	url_google_index	Is URL indexed on Google	Numeric
13	domain_google_index	Is domain indexed on Google	Boolean
14	url_shortened	Is URL shortened	Boolean
15	phishing	Is phishing website	Boolean

The first group is based on the values of the attributes on the whole URL string, while the values of the following four groups are based on the particular sub-strings, as presented in Figure 1. The last group attributes are based on the URL resolve metrics as well as on the external services such as Google search index.

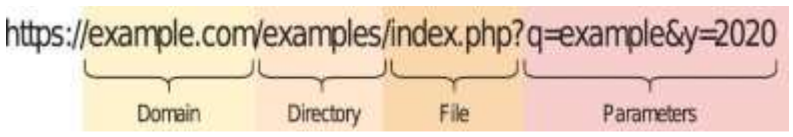


Fig. 1. Separation of the whole URL string into sub-strings.

5. Detailed Design

5.1 Data Collection

The data collection module includes WHOIS records, DNS records, and SSL certificate records. These records are collected using APIs and web scraping techniques. The collected data is stored in a database for further processing.

5.2 Data Preprocessing

The data preprocessing module includes the following submodules:

5.2.1 Data Cleaning

The data cleaning submodule removes the duplicates and inconsistent data.

5.2.2 Data Normalization

The data normalization submodule converts the data into a standard format that can be used for feature engineering.

5.3 Feature Engineering

The feature engineering module includes the following submodules:

5.3.1 Feature Extraction

The feature extraction submodule includes techniques such as bag of words and TF-IDF.

5.3.2 Feature Selection

The feature selection submodule selects the most important features using techniques such as mutual information and chi-square.

5.3.3 Feature Transformation

The feature transformation submodule includes techniques such as PCA and LDA.

5.4 Machine Learning Models

The machine learning models module includes the following submodules:

5.4.1 Model Selection

The model selection submodule selects the best model based on performance metrics such as precision, recall, and F1 score.

5.4.2 Hyperparameter Tuning

The hyperparameter tuning process is crucial for improving the performance of machine learning models. In our phishing domain detection system, we use various hyperparameters, such as learning rate, number of hidden layers, and dropout rate, in our neural network models.

To tune the hyperparameters, we use a combination of manual and automated techniques. First, we perform a grid search over a range of hyperparameters to identify the best combinations. We then use the best combinations from the grid search as a starting point for a more refined search using a random search algorithm. This approach allows us to explore a wider range of hyperparameters and identify the best ones more efficiently.

During the hyperparameter tuning process, we use cross-validation to evaluate the performance of the models on a held-out validation set. We also monitor the training and validation loss to ensure that the models are not overfitting to the training data.

Once we have identified the best hyperparameters, we train the final models using the full training dataset and evaluate their performance on a separate test set. We report the performance metrics for the final models in the results section.

5.4.3 Deployment for localhost Django

After completing the development and testing phases, the next step is to deploy the Phishing Domain Detection system on a localhost server using Django. Here are the steps to deploy the system:

1. Install necessary packages:
    - Python 3.9.13
    - Django
    - Scikit-learn
    - Numpy
    - Pandas
2. Clone the project repository from GitHub.
3. Open the command prompt and navigate to the project directory.
4. Run the following command to migrate the database:

    ```
    python manage.py migrate
    ```
5. Create a superuser to access the admin panel:

    ```
    python manage.py createsuperuser
    ```
6. Run the Django server using the following command:

    ```
    python manage.py runserver
    ```

7. Access the admin panel at **http://localhost:8000/admin/** and log in with the superuser credentials.

8. Add the necessary data to the database through the admin panel.

9. Access the Phishing Domain Detection system at **http://localhost:8000/** and test it with sample data.

10. If necessary, modify the source code and repeat the deployment process.

5.5 System Evaluation

To evaluate the performance of our phishing domain detection system, we use several metrics, including accuracy, precision, recall, and F1 score. We also analyze the confusion matrix to understand the types of errors the model makes and identify areas for improvement.

6.Technology stack

Front End	HTML/CSS/JS
Backend	Python Django
Database	SQLite
Visualization	Sea-born , Matplotib
Data version control	DVC
Source version control	GitHub

5. Proposed Solution

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that’s best fit for the above case.

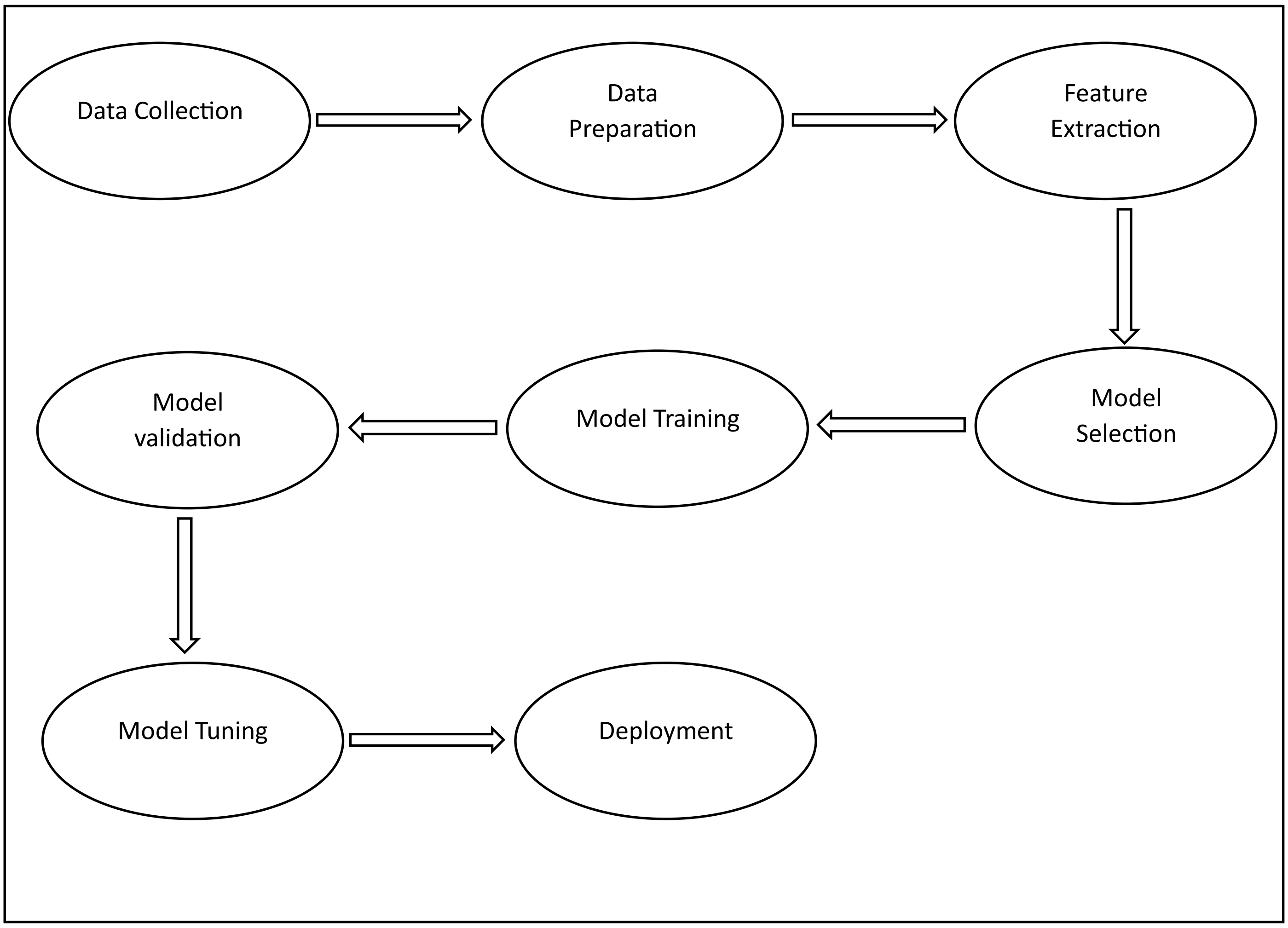
6. Proposed Architecture

The proposed architecture for phishing domain detection involves several key components and modules, each responsible for a specific task in the overall system. The architecture includes the following:

1. **Data Collection Module**: The data collection module is responsible for collecting data from various sources, such as web crawlers, public blacklists, and user reports. This data is then preprocessed to remove irrelevant information and stored in a database for further analysis.
2. **Feature Extraction Module**: The feature extraction module analyzes the collected data to extract relevant features that can be used to identify phishing domains. These features include domain age, IP address, SSL certificate, URL structure, content analysis, and other factors that contribute to the likelihood of a domain being a phishing site.
3. **Machine Learning Module:** The machine learning module uses the extracted features to train machine learning models to identify phishing domains accurately. The module includes several algorithms, such as Random Forest, Decision Trees, and Support Vector Machines (SVM), to classify domains as either phishing or legitimate.
4. **User Interface Module:** The user interface module provides a web-based interface for users to submit domain URLs for analysis. Users can also view the results of previous analyses and interact with the system to provide feedback on the accuracy of the predictions.
5. **Reporting Module:** The reporting module generates reports on the system's performance, including the accuracy of the predictions and the number of false positives and false negatives. The reports can be used to improve the system's accuracy and identify areas for further development.
6. **Alert Module:** The alert module is responsible for generating alerts when a phishing domain is detected. The alerts can be sent to users, administrators, or other systems to take immediate action, such as blocking the domain or warning users not to access it.

Overall, the proposed architecture provides a comprehensive solution for detecting phishing domains using machine learning and data analysis techniques. The system's modular design allows for easy scalability and flexibility to adapt to changing threats and requirements.

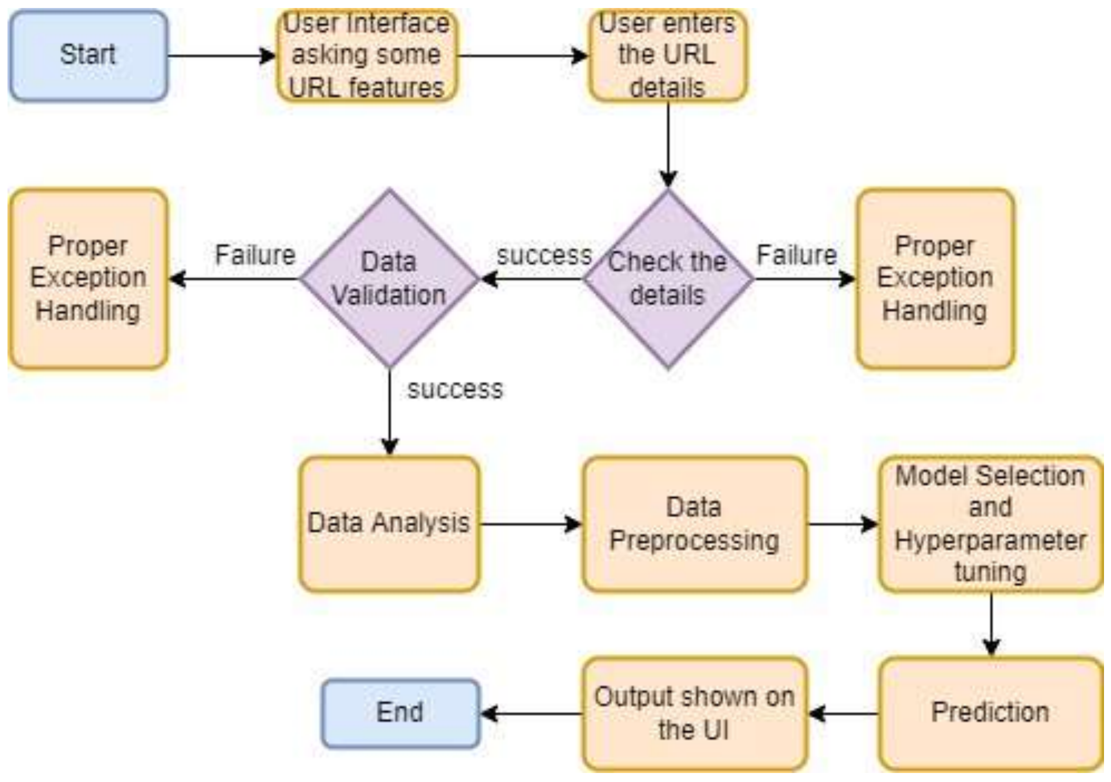
7. Model training/validation workflow



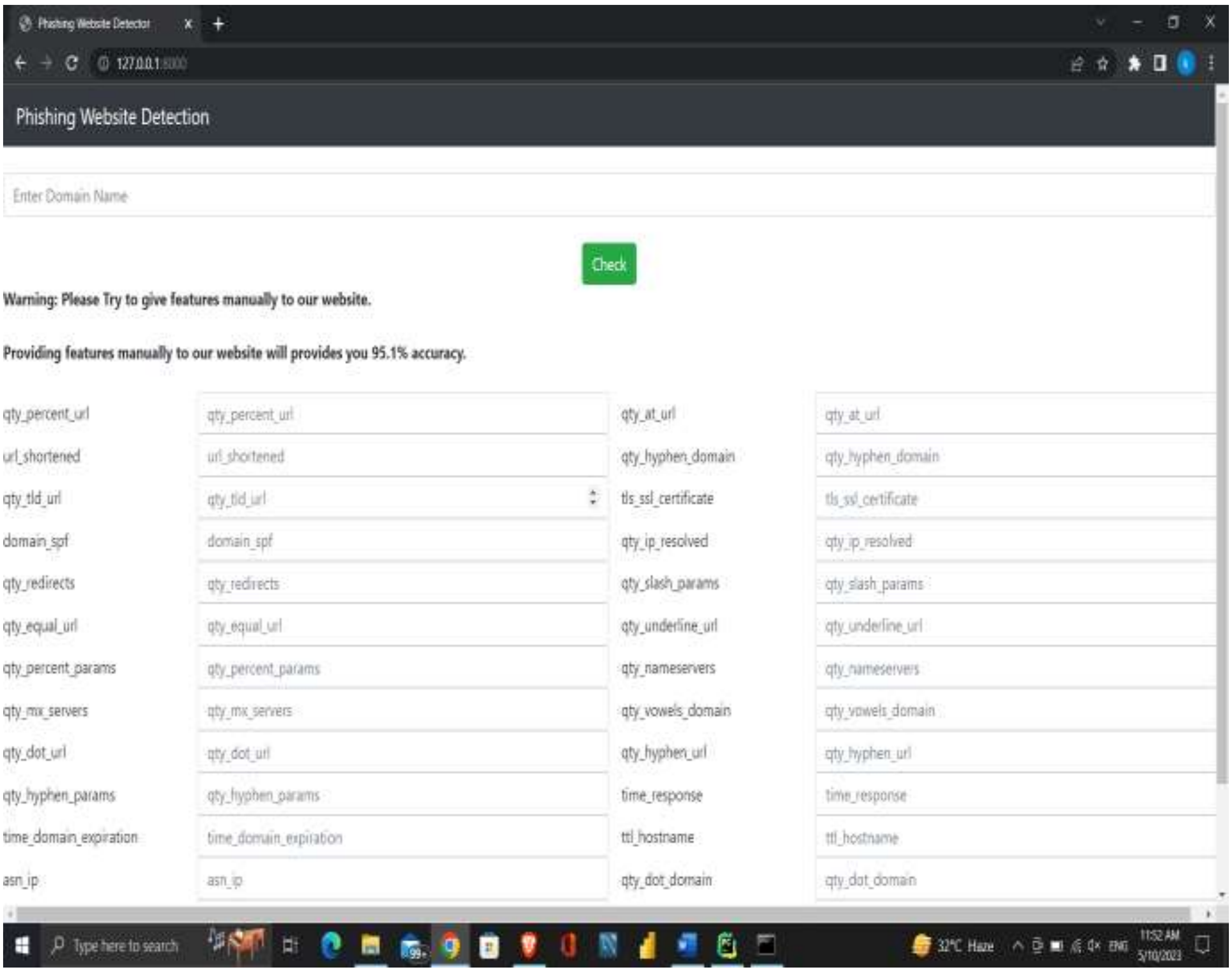
1. **Data Collection**: The first step in the model training process is to collect a large and diverse dataset of known phishing domains. The dataset should contain examples of different types of phishing attacks, including spear-phishing, clone phishing, and whaling attacks.
2. **Data Preparation**: Once the dataset has been collected, it needs to be cleaned and prepared for use in the model training process. This involves removing duplicates, verifying domain names, and filtering out irrelevant data.
3. **Feature Extraction**: Next, features need to be extracted from the dataset. Features are specific pieces of information that can help the model distinguish between legitimate and phishing domains. Common features used in phishing domain detection include the length of the domain name, the presence of certain keywords, and the use of certain characters or symbols.
4. **Model Selection**: After the features have been extracted, a suitable machine learning algorithm needs to be chosen to train the model. Popular algorithms used in phishing domain detection include Random Forest, Support Vector Machines (SVM), and Neural Networks.
5. **Training the Model**: Once the algorithm has been selected, the model needs to be trained using the prepared dataset. The model learns to differentiate between legitimate and phishing domains based on the extracted features.
6. **Model Validation**: After the model has been trained, it needs to be validated using a separate dataset of known phishing domains. This is done to ensure that the model is accurate and effective in detecting phishing domains.
7. **Model Tuning**: If the model is not accurate enough, it needs to be tuned by adjusting the algorithm parameters and feature selection. This process is iterative and continues until the model reaches an acceptable level of accuracy.
8. **Deployment**: Once the model has been trained and validated, it can be deployed in a production environment to detect phishing domains in real-time.

Overall, the model training and validation workflow is critical for ensuring that the model is accurate and effective in detecting phishing domains. It involves collecting and preparing data, extracting features, selecting a suitable machine learning algorithm, training the model, validating it, tuning it if necessary, and deploying it in a production environment.

8.User I/O workflow


Start
User Interface asking some URL features
User enters the URL details
Check the details
Failure
Proper Exception Handling
success
Data Validation
Failure
Proper Exception Handling
success
Data Analysis
Data Preprocessing
Model Selection and Hyperparameter tuning
Prediction
Output shown on the UI
End

11.Output



Phishing Website Detection

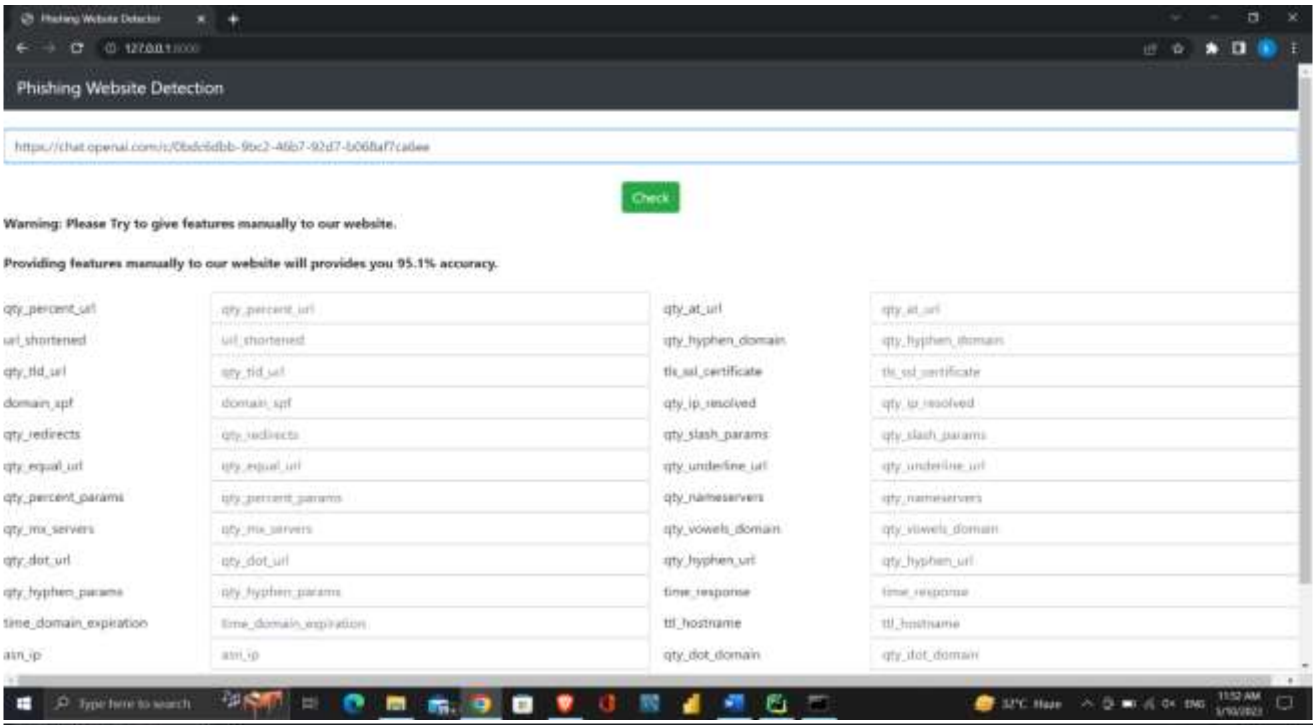
Enter Domain Name

Check

Warning: Please Try to give features manually to our website.

Providing features manually to our website will provides you 95.1% accuracy.

qty_percent_url	qty_percent_url	qty_at_url	qty_at_url
url_shortened	url_shortened	qty_hyphen_domain	qty_hyphen_domain
qty_tld_url	qty_tld_url	tls_ssl_certificate	tls_ssl_certificate
domain_spf	domain_spf	qty_ip_resolved	qty_ip_resolved
qty_redirects	qty_redirects	qty_slash_params	qty_slash_params
qty_equal_url	qty_equal_url	qty_underline_url	qty_underline_url
qty_percent_params	qty_percent_params	qty_nameservers	qty_nameservers
qty_mx_servers	qty_mx_servers	qty_vowels_domain	qty_vowels_domain
qty_dot_url	qty_dot_url	qty_hyphen_url	qty_hyphen_url
qty_hyphen_params	qty_hyphen_params	time_response	time_response
time_domain_expiration	time_domain_expiration	ttl_hostname	ttl_hostname
asn_ip	asn_ip	qty_dot_domain	qty_dot_domain



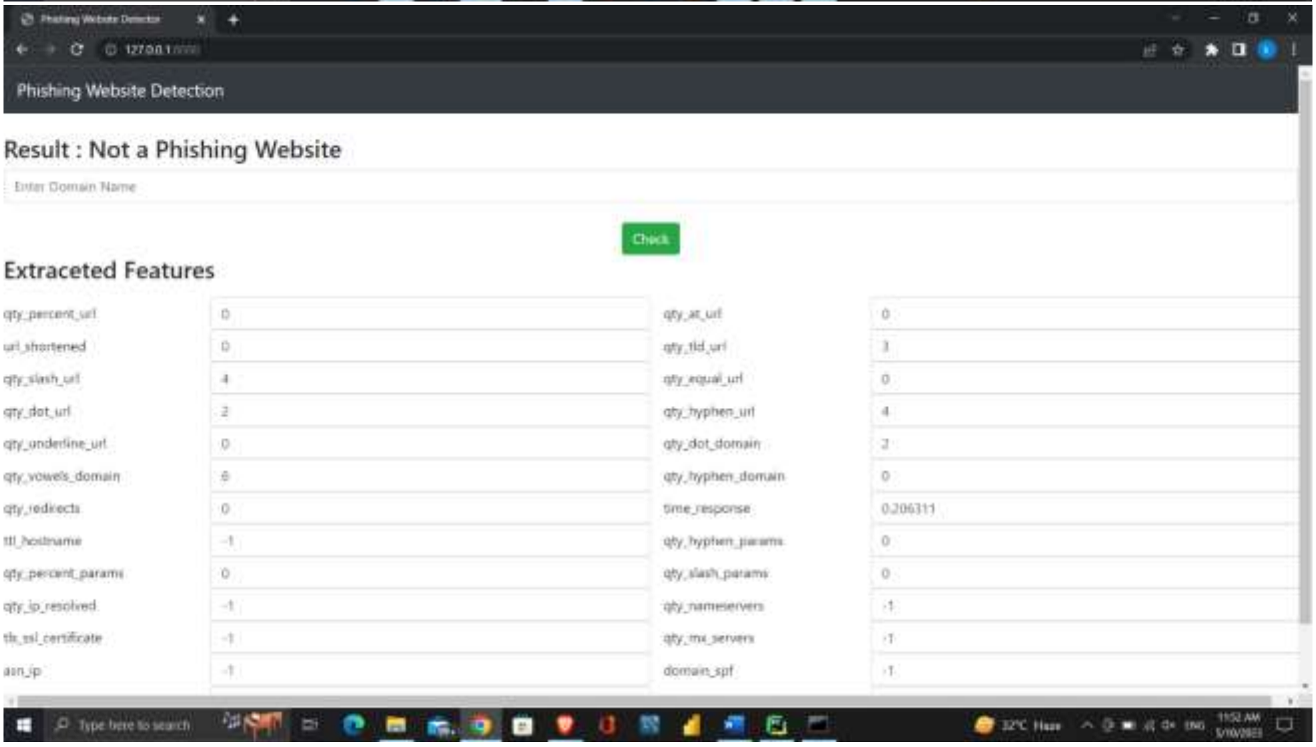
Phishing Website Detection

https://chat.openai.com/c/0bdc6dbb-9bc2-46b7-92d7-b068af7ca6ee

Warning: Please Try to give features manually to our website.

Providing features manually to our website will provides you 95.1% accuracy.

qty_percent_url	qty_percent_url	qty_at_url	qty_at_url
url_shortened	url_shortened	qty_hyphen_domain	qty_hyphen_domain
qty_tld_url	qty_tld_url	tls_ssl_certificate	tls_ssl_certificate
domain_spf	domain_spf	qty_ip_resolved	qty_ip_resolved
qty_redirects	qty_redirects	qty_slash_params	qty_slash_params
qty_equal_url	qty_equal_url	qty_underline_url	qty_underline_url
qty_percent_params	qty_percent_params	qty_nameservers	qty_nameservers
qty_mx_servers	qty_mx_servers	qty_vowels_domain	qty_vowels_domain
qty_dot_url	qty_dot_url	qty_hyphen_url	qty_hyphen_url
qty_hyphen_params	qty_hyphen_params	time_response	time_response
time_domain_expiration	time_domain_expiration	ttl_hostname	ttl_hostname
asn_ip	asn_ip	qty_dot_domain	qty_dot_domain



Phishing Website Detection

Result : Not a Phishing Website

Enter Domain Name

Extraceted Features

qty_percent_url	0	qty_at_url	0
url_shortened	0	qty_tld_url	3
qty_slash_url	4	qty_equal_url	0
qty_dot_url	2	qty_hyphen_url	4
qty_underline_url	0	qty_dot_domain	2
qty_vowels_domain	6	qty_hyphen_domain	0
qty_redirects	0	time_response	0.206311
ttl_hostname	-1	qty_hyphen_params	0
qty_percent_params	0	qty_slash_params	0
qty_ip_resolved	-1	qty_nameservers	-1
tls_ssl_certificate	-1	qty_mx_servers	-1
asn_ip	-1	domain_spf	-1

12.Key performance indicators (KPI)

- Key indicators displaying a summary of the anomaly detection in the URLs.
- Total number of slash in URL
- Total number of characters in URL
- Total number of dots in URL domain section
- Total number of dots in URL directory section
- Total number of hyphen in URL directory section
- Total number of filename characters
- Total number of underlines in directory section of URL
- ASN IP number of URL
- Enter no of days of time domain activation
- Enter no of days of time domain expiration
- Enter time to live

13. Conclusion

The Phishing Domain Detection System will detect whether the domains is real for fake and avoid user to become a victim of phishing based on various data used to train our algorithm, so we can identify the fake domain(URL) so prevent the loss of crucial data by avoiding phishing.

14. References

1) Datasets for phishing websites detection - ScienceDirect
2) Phishing Websites Dataset - Mendeley Data
3) https://getbootstrap.com/docs/4.3/getting-started/introduction/
4) https://www.youtube.com/watch?v=1BSwYlJUxK0&list=PLZoTAELRMXVOk1pRcOCaG5xtXxgMalpIe
5) https://www.youtube.com/watch?v=ioN1jcWxbv8&list=PLZoTAELRMXVPQyArDHyQVjQxjj_YmEuO9
6) https://www.youtube.com/watch?v=uMlU2JaiOd8&list=PLZoTAELRMXVPgjwJ8VyRoqmfNs2CJwhVH
7) Y. Zhang, L. Gao, and Y. Tang, "A Novel Phishing Website Detection Method Based on Convolutional Neural Network," in Proceedings of the 14th International Conference on Machine Learning and Cybernetics, pp. 1837-1842, 2015.