Bahria University,

Karachi Campus



COURSE: SOFTWARE CONSTRUCTION

TERM: FALL 2021, CLASS: BSE- 5 (A)

PROJECT NAME

PHISHING WEBSITE DETECTION

ENGR.RAMSHA MASOOD. Dr.Amina jameel

Submitted By.

Sami Ullah khan (02-131192-071)

Muhammad saad kamal (02-131192-075)

Signed Remarks: Score:

# ABSTRACT

A phishing technique was described in detail in a paper and presentation delivered to the 1987 International HP Users Group. Ever since the concept of phishing has been introduced it’s growing so fast. The problem is we do not have enough resources to detect this phishing website on a real time. Phishing attempts have grown 65% in the last year and it’s still growing so fast because the average cost of a phishing attack for mid-size companies $1.6 million. The technique for detection for phishing website has been proposed and project has been implemented in this called ‘PHISHING DETECTION THROUGH MACHINE LEARNING’. In this project I have emphasized on the detection through the URL also called ‘DECEPTIVE PHISHING’, because it is the most common way people visit the websites and links. The main aim for this project is to detect the malicious website and detect the location of the server. On top of that twenty different techniques have been used to analyze the URL type which decides if the URL is phishing or not. In the second part of the project another type of phishing has been introduced. It is called ‘SPEAR PHISHING’; in this technique I have used the email header to detect the email path and some security protocols to make sure that the email we get is safe or not. In the third section I have noted some common type of phishing through Wikipedia. It is very useful for the people who do not have any prior knowledge in this area. So it is mandatory for those people to have some basic idea of phishing and its type along with how to get rid of it using this website.

# ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without mentioning about the people who made it possible because "Success is the abstract of hard work & perseverance, but steadfast of all is encouraging guidance". So, I take this opportunity to acknowledge all those whose continuous guidance and encouragement served as a beacon light & crowned my effort, thus leading to successful completion of this work.

Specially to YouTube .

Table of content

1. [Introduction 6](#_TOC_250009)
2. [Literature Review 6](#_TOC_250008)
3. [Scope of the project 8](#_TOC_250007)
   1. [Existing system VS Proposed System 9](#_TOC_250006)
4. [Overview of phishing 10](#_TOC_250005)
5. [Practical use of phishing detection website 11](#_TOC_250004)
6. [System architecture 12](#_TOC_250003)
   1. Unified modeling language diagram 12
      1. Use case diagram 13
      2. Block machine diagram 14
      3. Activity diagram 15
7. [System requirements 16](#_TOC_250002)
8. [Project architecture 17](#_TOC_250001)
9. [Project source code 28](#_TOC_250000)
10. Results 50
11. Conclusion and Future work 59
12. References 60

# Introduction

In this modern world everyone is accessing the internet for different purpose. As the internet grew; there are some problems like stealing data, hacking and fake web pages came into being. To get rid of these phishing websites ‘phishing website detection website’ is very useful. It helps users to understand what phishing is and how to stay safe while surfing the internet, because people use their personal information like credentials and bank account details.

This project has three main parts. In the first part it has ‘URL detection’ method. In this part what it does is when we access to the webpage, we don’t know that the web page we are accessing is safe or not. So, in this first part when you copy the URL and search it tells you weather the URL is phishing or not. In the second part ‘Mail phishing’ it tells user to copy the email header and paste in to the given header to analyze it. Then it displays the result. In the last part called ‘Wiki’ it tells you about the basic types of phishing and gives the basic idea of what is phishing for the new user who doesn’t have any prior knowledge. So, it is very useful nowadays because we have to do most of the things on the internet, so it should be safe and this project achieve this goal.

## Literature Review

* 1. Anti-phishing solutions can be positioned at different levels of attack flow where most researchers are focusing on client-side solutions which turn to add more processing overhead at the client side and lead to losing the trust and satisfaction of the users. In this paper they introduced the anti- phishing method called ‘Wide scope and fast website phishing detection using URLs lexical features. Nowadays many organizations make centralized protection of spam filtering. This paper proposes a system which can be integrated into such process to increase the detection performance in a real time. The simulation results of the proposed system showed a phishing URLs detection accuracy with 93% and provided online process of a single URL in average time of 0.12 second. They have used the below formula to calculate the accuracy.

Accuracy: defined as the percentage of the correct classification over all attempts of classification. [1]

* 1. In the research paper ‘Using domain top similarity features in machine learning’ they proposed a web base similarity to add features in detection. they first experiment 200 web data in which 100 webs were phishing, the result in terms of f-measure was 0.9250 with 7.50% of error rate. The system uses 8 different features to detect the phishing web. They used the vector system to differentiate the good and bad web page and calculate the error rate. The result part includes that the result from the SVM and Random forest were 8.50% but the AdaBoost had 10.00% accuracy. The last result section tells that It is easy to implement and can achieved with 19.50% error rate and 0.8312 f-measure.

## Scope of The Project

Phishing is a considerable problem which differs from the other security threats such as intrusions and Malware which are based on the technical security holes of the network systems. The weakness point of any network system is its Users. Phishing attacks are targeting these users depending on the trikes of social engineering. . There are many possible ways to do phishing but unfortunately, we have some techniques available nowadays can stop spear and email phishing and therefore we need to build a system which can help us to stop phishing attacks. One real example is the Google mail phishing attack happened some years ago and many users lost their trust on Google. So, to avoid all these big threats, we need to think possible ways to avoid large scale attacks to protect users. The scope for the phishing in this project is to save maximum users on the web and the one who work on emails. The motto of the PHISHING DETECTION project is when user is about to enter any information such as personal info or bank information we recommend them to search that link or analyze the header of that email in the web detect site first and if it shows the result is legitimate then they are safe to give the information over the internet. The scopes of the project are as follows:

* + - To develop a system which can cover the large scope of safe surfing on the internet with using URL detection method; URL is the most common type user interact with possible viruses and attacks
    - To begin the awareness of current threat and possible future fraud methods for all users which can educate them about how to stay away from attacks and spam websites and not to become them bate
    - A good practical solution for big companies to keep their customer safe and win their trust and give satisfied services.
    - Mail header analyzing system introduced for big organizations and people as a good filter to avoid scam happened through emails and links which lead them to fake web page.
    - A very quick and user-friendly system to check and proceed with safety to save time, money, energy and peace of mind.

## Overview of phishing

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. In this process the victim receives the message that looks like it has been sent by the known person or organization. In this message it has some links or viruses which tricks user into divulging personal and bank information such as password, username and account IDs. Phishing is very common among attackers because it is easier for them to trick people to make them click on the malicious link which seems legitimate than trying to break

through the computer’s defense system. The malicious links within the body of the message designed to make it appear that they go to the fake organization using that fake company’s logo and other legitimate content.

## Practical use of phishing detection website

The main aim of phishing is to steal personal information and use it for some benefits. Now as a part of solution we need to detect this attack before the attacker make us fool. For example, if you get and email from the bank you have your account in saying that you must register account through the given link otherwise it will be closed. So, what most of the people do is they believe that that email is from the bank and they trust that their information will go to the bank, however the attacker spoofs the email and steal all the information. The process has five steps.

* + - **Planning**: attacker decides which business or organization to target and how to get email ids of the customers for that organization. They often use the mass-mailing technique.
    - **Setup**: Once they know which business to spoof and what their customers are, they find a way to deliver the message with the spoof links or documents with company’s logo to get trust from customers.
    - **Attack**: This is the most familiar step for everyone; in this step phisher sends a phony message that seems like from reputable source.
    - **Fraud**: The data or information phisher gathered; they use it to fraud like transfer money from other account or make illegal purchase.

To get rid of this fraud and spoofing emails or messages we need to develop a system like phishing detection website which helps people to stop getting into all these troubles and save their time, money and energy. We need to build a system like detection website which people can trust and check the URL link before entering their person data and credentials.

**TECHNOLOGY**

Python. Jupyter notebook.

Pandas

Numpy

Matplotlib

LogisticRegression

train\_test\_split confusion\_matrix

Flask

pickle

FUNCTIONALITIES

Phishing is an attack wherein the attacker exploits social engineering techniques to perform identity theft. Phishing traditionally functions by sending forged e-mail, mimicking an online bank, auction or payment sites, guiding users to a bogus web page which is carefully designed to look like the login to the genuine site

**Unified modeling language 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. UML is specifically constructed through two different domains they are:

* + 1. UML Analysis modeling focuses on the user model and structural model views of the system.
    2. UML design modeling focuses on the behavioral modeling, implementation modeling and environmental model views.

## 6.1.0 Use-case Diagram

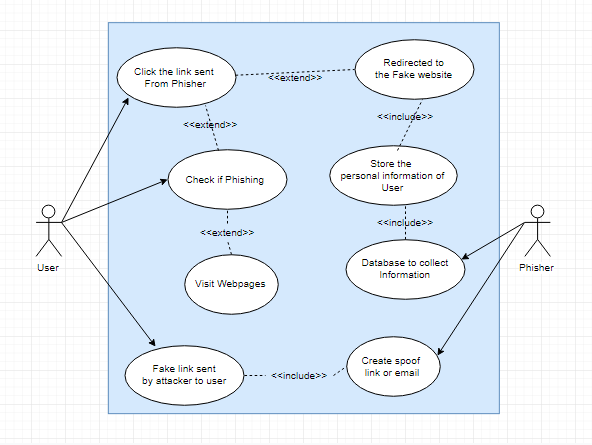


Figure 1. Use-Case Diagram

## 6.2.0 State-machine Diagram

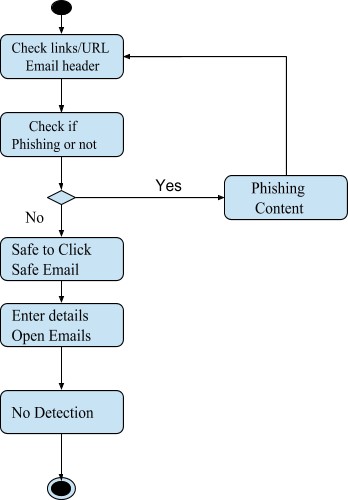


Figure 2. State-Machine Diagram

## 6.3.0 Activity Diagram

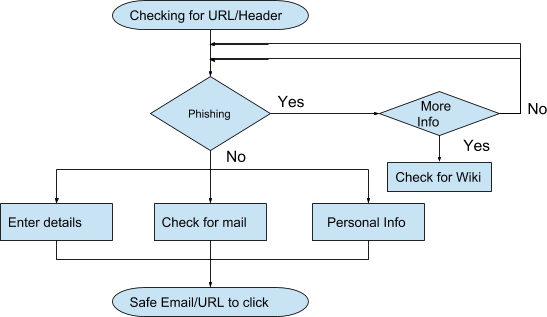


Figure 3. Activity Diagram

## System Requirements

The below mentioned minimum requirements are used for developing this application.

We can also use the updated or other requirements depending on the user.

## Software Requirements:

|  |  |
| --- | --- |
| Operating System | Windows 8, Linux, Ubuntu |
| Programming Language | Python, HTML, JavaScript, CSS |
| Web Server | Lenovo G470 |
| Editor | Sublime text or any editior |
| IDE | Spyder OR Python IDE, Eclipse |

**Hardware Requirements:**

|  |  |
| --- | --- |
| Processor | Intel i5 |
| Hard Disk | 500 GB |
| RAM | 4GB |

1. MODULE DISTRIBUTION

|  |  |  |
| --- | --- | --- |
| 65180 | Sami ullah khan | Algorithm implementation and back end development and data collection |
| 65183 | Saad kamal | Front end and compatibility of backend and front end. |

## Project source code:

**PREDICTION\_APP.py**

import uvicorn

from fastapi import FastAPI

import joblib,os

app = FastAPI()

#pkl

phish\_model = open('phishing.pkl','rb')

phish\_model\_ls = joblib.load(phish\_model)

# ML Aspect

@app.get('/predict/{feature}')

async def predict(features):

    X\_predict = []

    X\_predict.append(str(features))

    y\_Predict = phish\_model\_ls.predict(X\_predict)

    if y\_Predict == 'bad':

        result = "This is a Phishing Site"

    else:

        result = "This is not a Phishing Site"

    return (features, result)

if \_\_name\_\_ == '\_\_main\_\_':

    uvicorn.run(app,host="127.0.0.1",port=8000)

PHISHING SITE URL .IPYNB.

import pandas as pd # use for data manipulation and analysis

import numpy as np # use for multi-dimensional array and matrix

import seaborn as sns # use for high-level interface for drawing attractive and informative statistical graphics

import matplotlib.pyplot as plt # It provides an object-oriented API for embedding plots into applications

%matplotlib inline

# It sets the backend of matplotlib to the 'inline' backend:

import time # calculate time

from sklearn.linear\_model import LogisticRegression # algo use to predict good or bad

from sklearn.naive\_bayes import MultinomialNB # nlp algo use to predict good or bad

from sklearn.model\_selection import train\_test\_split # spliting the data between feature and target

from sklearn.metrics import classification\_report # gives whole report about metrics (e.g, recall,precision,f1\_score,c\_m)

from sklearn.metrics import confusion\_matrix # gives info about actual and predict

from nltk.tokenize import RegexpTokenizer # regexp tokenizers use to split words from text

from nltk.stem.snowball import SnowballStemmer # stemmes words

from sklearn.feature\_extraction.text import CountVectorizer # create sparse matrix of words using regexptokenizes

from sklearn.pipeline import make\_pipeline # use for combining all prerocessors techniuqes and algos

from PIL import Image # getting images in notebook

# from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator# creates words colud

from bs4 import BeautifulSoup # use for scraping the data from website

from selenium import webdriver # use for automation chrome

import networkx as nx # for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

import pickle# use to dump model

import warnings # ignores pink warnings

phishing\_data1 = pd.read\_csv('phishing\_urls.csv',usecols=['domain','label'],encoding='latin1', error\_bad\_lines=False)

phishing\_data1.columns = ['URL','Label']

phishing\_data2 = pd.read\_csv('phishing\_data.csv')

phishing\_data2.columns = ['URL','Label']

phishing\_data3 = pd.read\_csv('phishing\_data2.csv')

phishing\_data3.columns = ['URL','Label']

warnings.filterwarnings('ignore')

for l in range(len(phishing\_data1.Label)):

    if phishing\_data1.Label.loc[l] == '1.0':

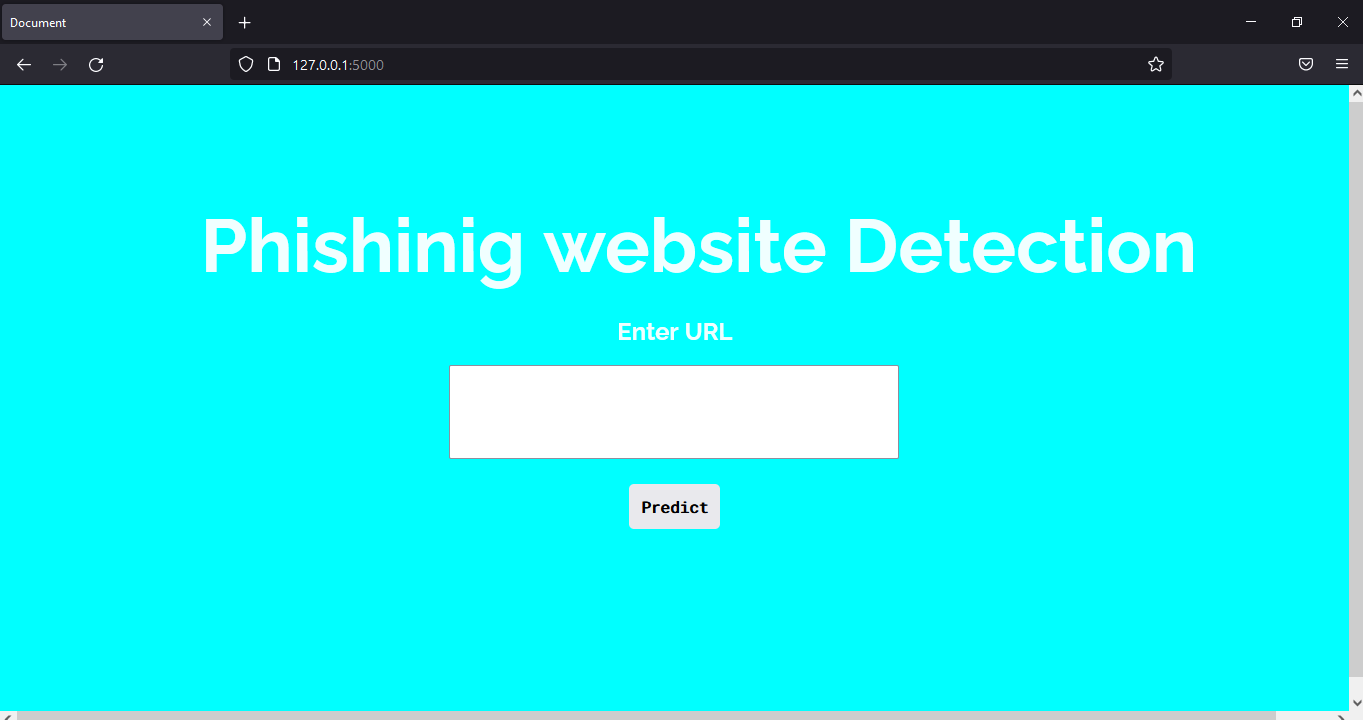
        phishing\_data1.Label.loc[l] = 'bad'

    else:

        phishing\_data1.Label.loc[l] = 'good'

## INTERFACES

In this project I have used many parts to build the whole system.



## Project Results:

## 

## 

## 

## Conclusion and future work

As a conclusion of the project ‘Phishing detection with machine learning’ is a new concept. There are twenty different methods have been used to carry out the result which is fast, quick and reliable for all the users. Also, the model used for training and testing gives the best accuracy of 97% for the random forest algorithm for machine learning. The methods used for the mail header analyzer is more advanced and gives the result according the path email header had visited, and the protocols needs to be followed. So, it is very safe method of detection of spoof OR phishing content on the internet.

The future work includes admin side. The current website is open source and we can improve it with some advance features. The first one is to make it access by cloud to everyone and when people try to access the web it will automatically ask for permission to track the search

and suspicious search will be blocked. On top of that the Google-chrome extension can be made to get the on-hand safety just like antivirus. The second method introduces the admin page. This is the page where a centralized person has the control over the phishing website and admin has a right to maintain or edit the content. last but not the least, if people have some doubts then they will have the facility of live chat over web page. Also, there are some tags shown in the current website which aware user about what is an ideal website has; for example, page rank, index number, host name etc. There could be one notification center which tells the real time news about what is happening around the globe. To conclude there will be some new and good method to prevent phishing threat in upcoming future and we will have advanced way to prevent.

## References:

* 1. Y. Daeef, R. B. Ahmad, Y. Yacob and N. Y. Phing, "Wide scope and fast websites phishing detection using URLs lexical features," 2016 3rd International Conference on Electronic Design (ICED), Phuket, 2016, pp. 410-415. doi: 10.1109/ICED.2016.7804679
  2. N. Sanglerdsinlapachai and A. Rungsawang, "Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection," 2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket, 2010, pp. 187-190. doi: 10.1109/WKDD.2010.108
  3. S. Parekh, D. Parikh, S. Kotak and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 949- 952. doi: 10.1109/ICICCT.2018.8473085
  4. G. Liu, B. Qiu and L. Wenyin, "Automatic Detection of Phishing Target from Phishing Webpage," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 4153-4156. doi: 10.1109/ICPR.2010.1010
  5. G. Liu, B. Qiu and L. Wenyin, "Automatic Detection of Phishing Target from Phishing Webpage," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 4153-4156.doi: 10.1109/ICPR.2010.1010
  6. J. Li and S. Wang, "PhishBox: An Approach for Phishing Validation and Detection," 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl

Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, 2017, pp. 557-564. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.101

1. <https://en.wikipedia.org/wiki/Phishing>
2. <https://www.sciencedirect.com/science/article/pii/S1071581918303628>
3. [https://archive.ics.uci.edu/ml/datasets/Phishing+Websites](https://archive.ics.uci.edu/ml/datasets/Phishing%2BWebsites)
4. <http://eprints.hud.ac.uk/24330/>
5. <https://documenter.getpostman.com/view/1034072/S17oxAFE>
6. <https://www.semrush.com/analytics/backlinks>
7. <https://toolbox.googleapps.com/apps/messageheader/>
8. <https://www.iptrackeronline.com/email-header-analysis.php>