"Detection of Phishing Websites Using Machine Learning"

1.Rohit Kundan Sonawane TE-B-59.
2.Sakshi Pandurang Mahajan. TE-A-55
3.Chetan Jayram Mahajan. TE-A-52
4.Vaishnavi Vilas Gadhe. TE-A-29

Department of Computer Engineering
Under the guidance of
Prof. A. T. Bhole Sir
Associate Professor

26 March 2022

Overview

Introduction

- Phishing costs Internet users billions of dollars per year. It refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting internet users.
- Phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords.
- This paper deals with methods for detecting phishing web sites by doing feature extraction of urls by Machine learning techniques and Natural Language Processing.

Objectives

- To explain what phishing websites are? And how they are major threat to peoples.
- To collect phishing websites database and perform processing on them
- After by applying various feature extraction techniques, fitting of the model with machine learning algorithm
- Improving accuracy of the model
- Deployment of model of web-page and make it ready mo use for end users.
- Users can enter their website on our website and check whether it is Phishing or not.

Problem Defination

- URLs sometimes known as “Web links” are the primary means by which users locate information in the Internet.
- Aim of the phishers is to acquire critical information like username, password and bank account details.
- Our aim is to derive classification models that detect phishing urls using machine learning and natural language processing. In Jupyter Environment

Hardware and Software Requirement

Hardware Requirement:-

- Intel Core i3 or above
- Processor: - 1.2GHz or above.
- RAM: - 2GB or above.
- Internal Storage: - 100GB or above.
- Internet Connectivity

Software Requirement :-

- Windows 7 or Higher
- Python 3.6.0 or Higher
- Visual Studio Code
- Flask

Tools Languages Used

Language Used:-

- HTML
- CSS
- Python

Software Used:-

- Jupyter Notebook
- Visual Studio Code
- Python 3.6

DataSet Used:-

- DataSet Of Phishing Website

Implementation Steps in Project

Implementation Steps :-

- Importing Some Useful Libraries.
- Data Preprocessing.
- Tokenizing the Strings.
- Stemming.
- Vectorization.
- Feature Extraction.
- Fitting the model in Support Vector Machine.
- Making the Pipeline.
- Loading the model with Pickle
- Predict the Output

Modules and Libraries used in Project

Modules used in Project :-

- Python Pandas.
- Python Numpy.
- Python Sklearn.
- Python Matplotlib.
- Python Nltk
- Python Pickle

Work Flow and Feature Extraction

Work Flow And Feature Extraction Steps :

- Firstly we have checked whether there is any null value present in the dataset or not and if so removed it.
- Use of RegexpTokenizer : with the help of tokenize.regexp() module, we are able to extract the tokens from strings by using regular expression with RegexpTokenizer() method.
- Use of PorterStemmer : stemmer is used to produce morphological variants of a root/base word. Stemming is desirable as it may reduce redundancy as most of the time the word stem and their derived words mean the same. And then after stemming joining of words is done.

Work Flow and Feature Extraction

- Feature Extraction using TfidfVectorizer: It's the main part of natural language processing . TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.
- Splitting the data: Train-test-split is used to split the data into training dataset and testing dataset.

Support Vector Machine Algorithm

S.V.M. = Support Vector Machine :-

- The S.V.M. performs classification by finding the hyper plane that maximizes the margin between two classes.
- The vectors that define the hyper plane are the support vectors. Below is the figure showing how it works

S.V.M. Concept

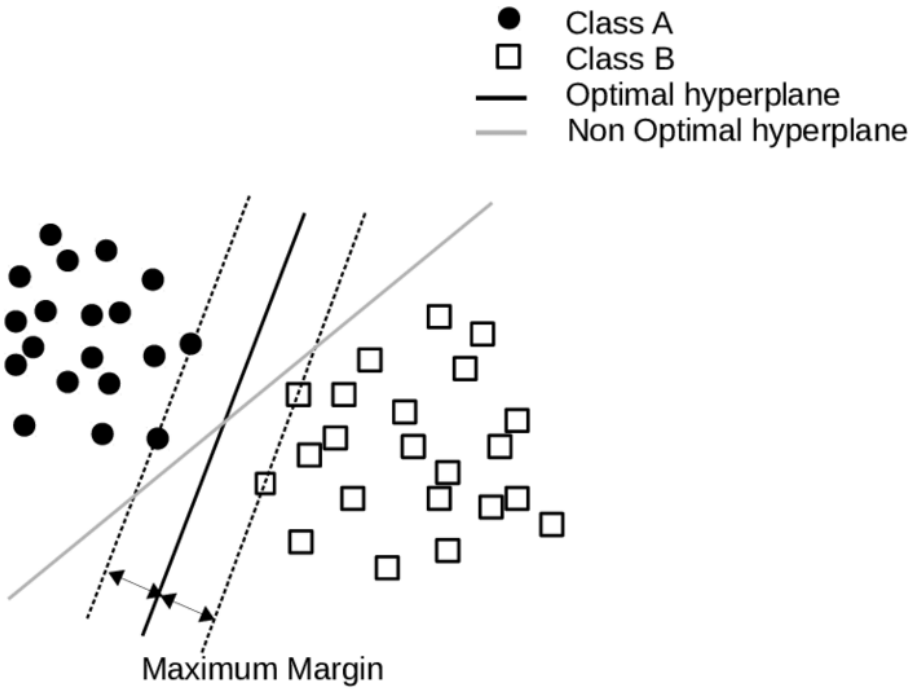

Figure: Support Vector Machine Algorithm

Data Flow Diagram

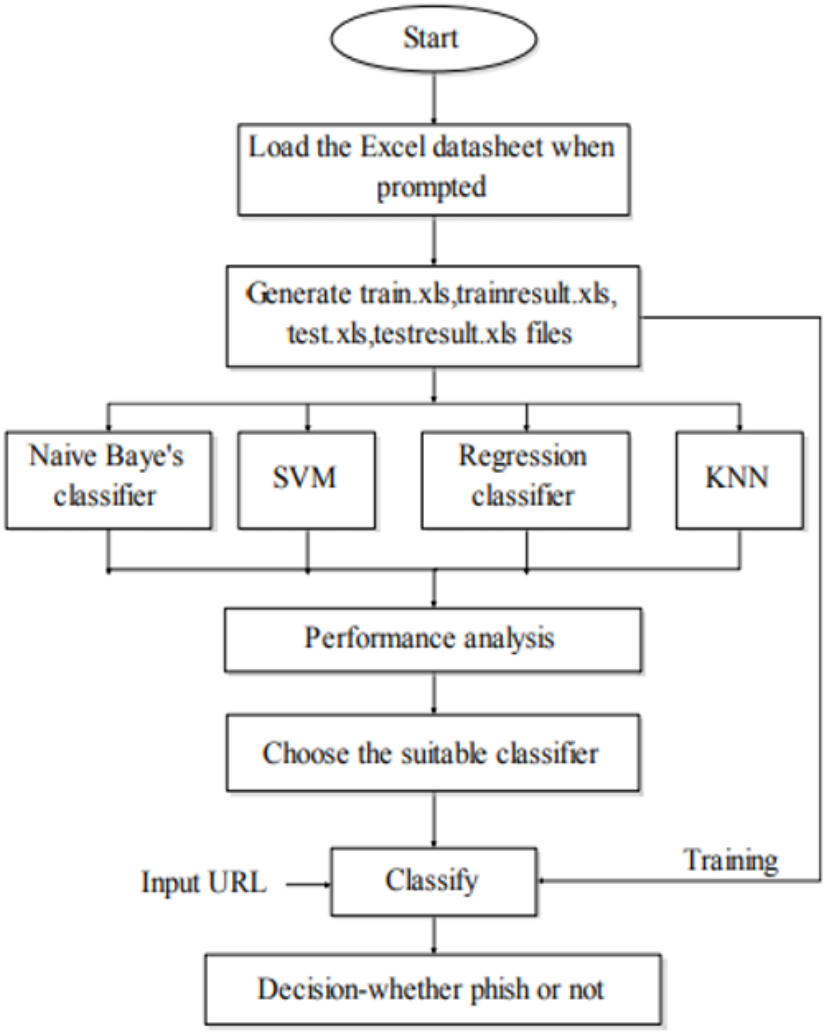

Figure: Data Flow Diagram

Use Case Diagram

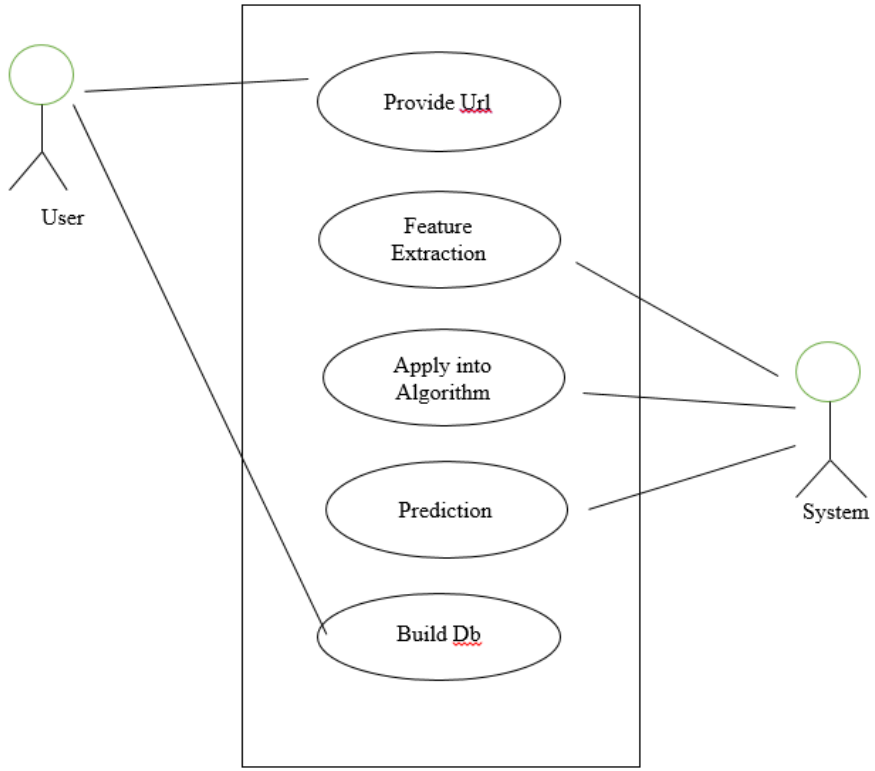

Figure: Use Case Diagram

Activity Diagram

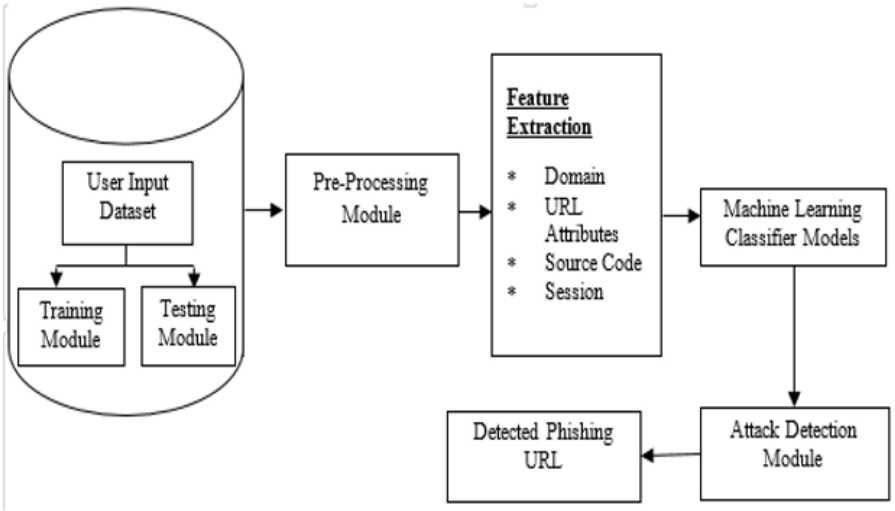

Figure: Activity Diagram

Component Diagram

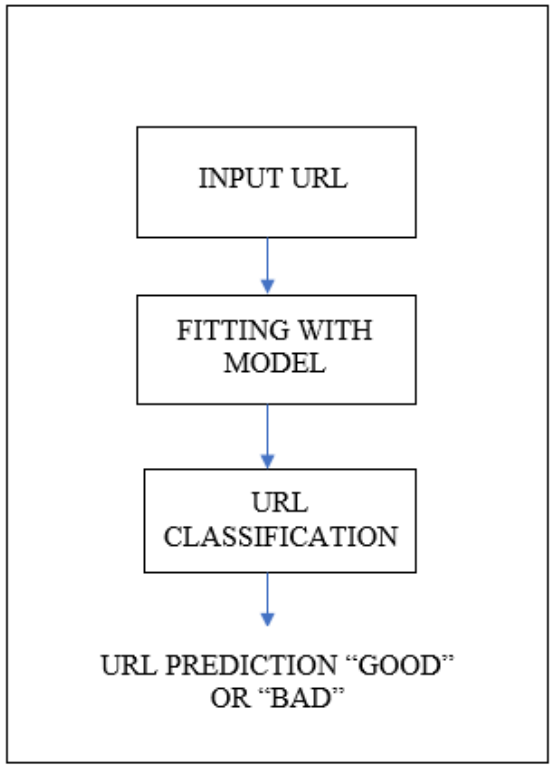

Figure: Component Diagram

Conclusion

Conclusion :-

- To summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have tested SVM machine learning algorithm on the 'Phishing Websites Dataset' and reviewed their results. We then built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using Support Vector Machine with and accuracy of 97.94 Percentage.

Result

Result :-

- Here we got model accuracy of 97 Percentage which is quite good. So Support Vector Machine is ideal to use.

Classifier	Accuracy Rate%	Error Rate%
SVM	Training-99% Testing-97%	3%

Figure: Result

Interface of Webpage

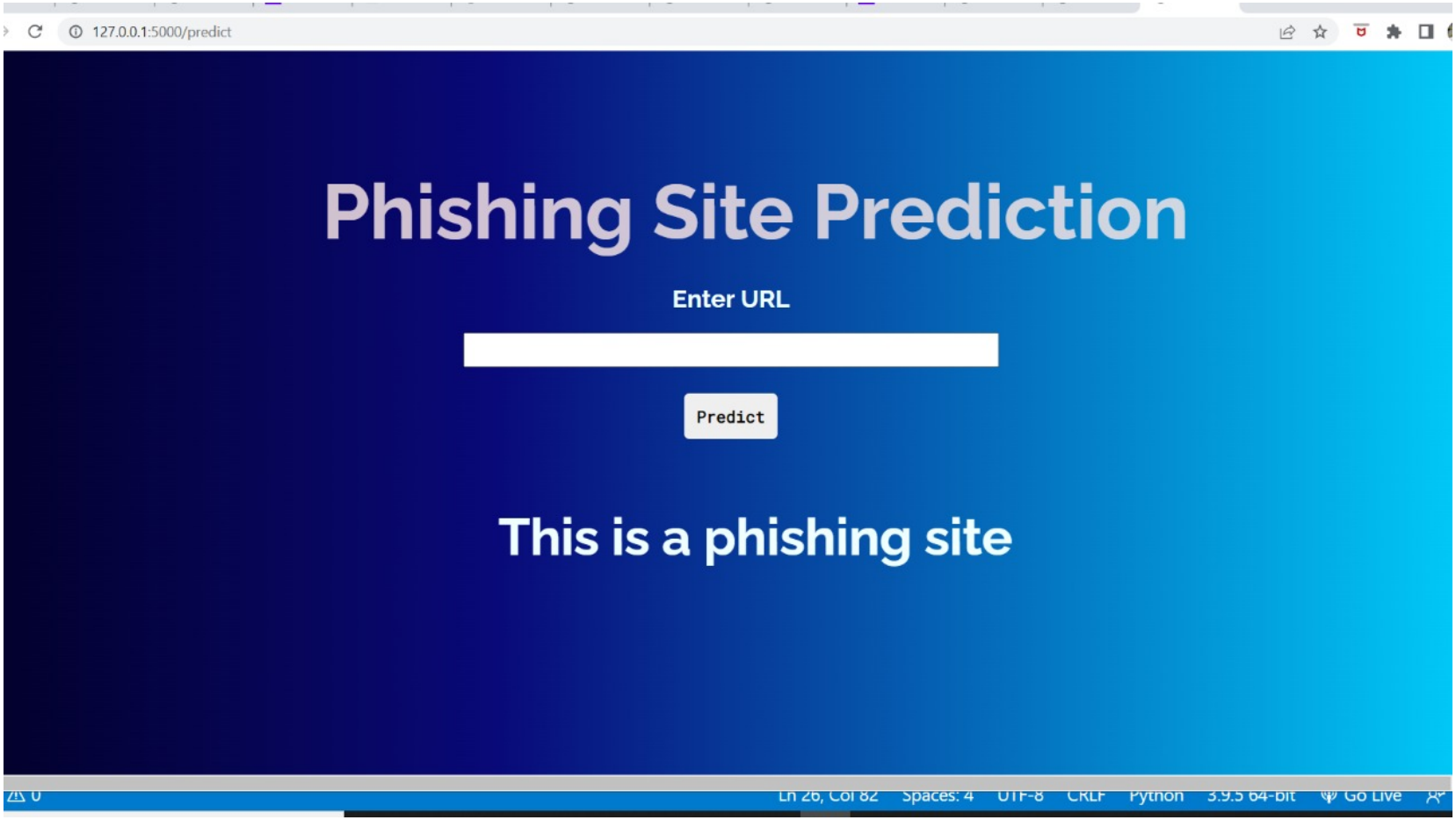

Figure: Interface of Webpage

References

References :-

- Detection of phishing websites using machine learning International Journal of Engineering Research Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV10IS050235 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 10 Issue 05, May-2021
- Joby James, Sandhya L, Ciza Thomas Detection of phishing websites using machine learning techniques. 2013 International Conference on Control Communication and Computing (ICCC)https://www.researchgate.net/publication/269032183
- Mustafa AYDIN, Nazife BAYKAL (2015) Feature extraction and classification phishing websites based on URL. IEEE

Thank you