**Phishing Website Detection**

A Minor Project Synopsis Report

Submitted in Partial Fulfilment of the Requirement for the Award of Degree of

**BACHELOR OF TECHNOLOGY**

**(COMPUTER SCIENCE AND ENGINEERING)**

To

****

**UTTARAKHAND TECHNICAL UNIVERSITY, DEHRADUN**

Submitted by

**Ankit Raj (180410101012)**

**Shantanu Paul (180410101044)**

****

**Shivalik College of Engineering**

**Dehradun, Uttarakhand**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Minor Project Synopsis**   |  |  |  | | --- | --- | --- | | **Faculty Use only:** | | | | **Selected / Rejected** | **Name of the Guide** | **Name of the Project coordinator** | |  |  |  |   ----------------------------------------------------------------------------------------------------------------------------------------   |  | | --- | | **1. Mini Project Topic / Title: Online Quiz Portal** | | Tentative: | | Finalized (Guide will write): |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **2. Details of Group Member:** | | | | | | | **Roll. No.** | **Course** | **Stream** | **Name of Student** | **Mobile Number** | **Email Id** | | 180410101012 | B.Tech | CSE 4th | Ankit Raj | 7352372491 | ankitrajrks.cse@sce.org.in | | 180410101044 | B.Tech | CSE 4th | Shantanu Paul | 7017346198 | shantanupaulscp.cse@sce.org.in | |
| **Project Overview/Background:**  Introduction**:** Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This project deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm  Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website.  Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks  account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing [1].  In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.  The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade  blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which  proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack.  Phishing attacks are categorized according to Phisher’s mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs . Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites. |

Need of Work / Reason for selection of this project:

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc.

Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

This study was motivated by the multiple millions of dollars that have been lost due to fraudsters operating fake versions of data collection websites and the need for a safer internet experience as we [progress](https://projectchampionz.com.ng/tag/progress/) in the internet and communication age.

Need of this work also include:-

* Developing a phishing detection system with new features.
* Creating a reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
* Studying previous work on the proposed topic and looking for ways to improve them.
* Optimizing the system.
* Implementing security [standards](https://projectchampionz.com.ng/tag/standards/) with the system.
* Creating the system which can also give suggestions to guest users.

Objective:

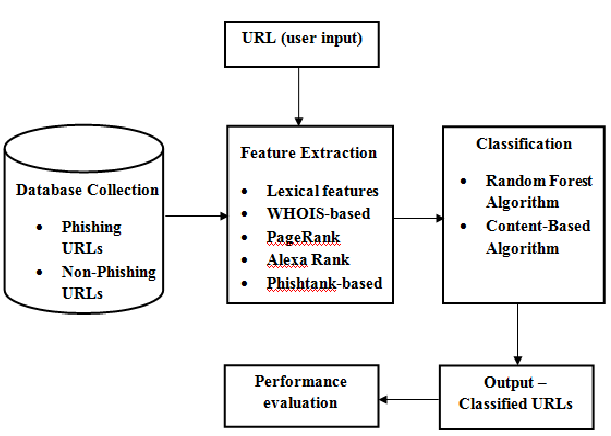
* There are number of users who purchase products online and make payment through e- banking.
* There are e- banking websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons.
* This type of e-banking websites is known as phishing website.
* In order to detect and predict e-banking phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm.
* We implemented classification algorithm and techniques to extract the phishing data sets criteria to classify their legitimacy.
* The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate.
* Once user makes transaction through online when he makes payment through e-banking website our system will use data mining algorithm to detect whether the e-banking website is phishing website or not.
* This application can be used by many E-commerce enterprises in order to make the whole transaction process secure.
* Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
* With the help of this system user can also purchase products online without any hesitation.

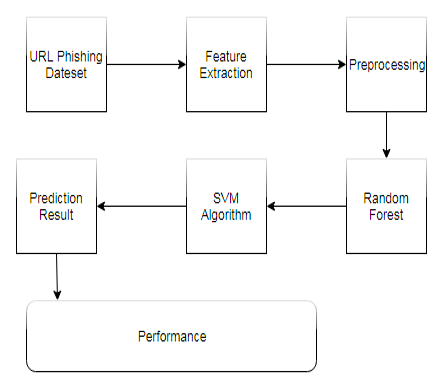
Problem Statement (explain what you want to implement in this project in

short sentence): Phishing attacks are categorized according to Phisher’s mechanism for trapping alleged users. Several forms of these attacks are keyloggers, DNS toxicity, Etc. The initiation processes in social engineering include online blogs, short message services (SMS), social media platforms that use web 2.0 services, such as Facebook and Twitter, file-sharing services for peers, Voice over IP (VoIP) systems where the attackers use caller spoofing IDs . Each form of phishing has a little difference in how the process is carried out in order to defraud the unsuspecting consumer. E-mail phishing attacks occur when an attacker sends an e-mail with a link to potential users to direct them to phishing websites.

* . In order to detect and predict e-banking phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm.
* We implemented classification algorithm and techniques to extract the phishing data sets criteria to classify their legitimacy.
* The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate.
* Creating a reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
* Creating the system which can also give suggestions to guest users.

Block diagram with explanation:



Other diagram for better understanding: 

Benefits to the surrounding/society:

Website Phishing costs internet users billions of dollars per year. Phishers steal personal information and financial account details such as usernames and passwords, leaving users vulnerable in the online space.

* This system can be used by many E-commerce or other websites in order to have good customer relationship.
* User can make online payment securely.
* Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
* With the help of this system user can also purchase products online without any hesitation.
* Creating a reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
* Creating the system which can also give suggestions to guest users.
* The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate.

**Methodology:**

Proposed Techniques or methods to be implemented:

1.Dataset Description:

We used the dataset provided by UCI Machine Learning repository⁴ collated by Mohammad et al³. The dataset has 11055 datapoints with 6157 legitimate URLs and 4898 phishing URLs. Each datapoint had 30 features subdivided into following three categories:

* URL and derived features
* Page’s source code-based features: Includes URLs embedded in the webpage and HTML and JavaScript based features.
* Domain based features
* Studying the way of extraction and relevance of features, we dropped 5 features out of 30, namely: Port Number, Abnormal URL, Pop-up Window, Google Index and Number of Links Pointing to a Page. Port Number was dropped due to feature drift. Rest were dropped due to unavailability of methods to extract them programmatically or absence of public APIs.

2.Visualising the Dataset:

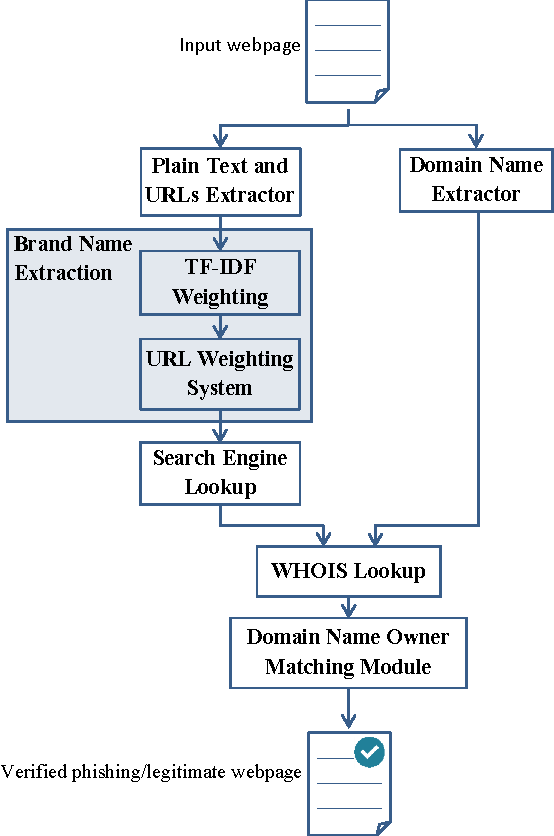
To see separability of the two classes, we plotted the t-SNE curve. The curve implied that though the classes are separable, they are not clustered together, and either transformation of the features or non-linear model is required to obtain good results.

3.Preprocessing the dataset:

We splitted the available data into training and testing data using 80:20 split. Post that, since we had only 7075 data points in the training data, we trained it using 5 fold cross validation. Hence, we achieved a train:val:test split of 64:16:20.

We one-hot encoded the features to avoid any biases due to numerical values.

Project Flowchart:



Advantages:

* This system can be used by many E-commerce or other websites in order to have good customer relationship.
* User can make online payment securely.
* Data mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
* With the help of this system user can also purchase products online without any hesitation.

### Limitations:

* . If Internet connection fails, this system won’t work.
* All websites related data will be stored in one place.

### Applications:

* This system can be used by many E-commerce or other websites in order to have good customer relationship.
  1. User can make online payment securely
* With the help of this system user can also purchase products online without any hesitation.
* A reporting platform for other users of the platform to report fake websites in order to build the knowledge base.
* The e-banking phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate.
* This system can also give suggestions to guest users.

### Project requirements:

### Equipments: A Computer with good internet connection and latest applications with proper functioning.

Facilities required (include software, hardware):

HARDWARE

* Windows 7,8,10 64 bit
* RAM 4GB

SOFTWARE

* Data Set
* Python 2.7
* Anaconda Navigator

Python’s standard library

* Pandas
* Numpy
* Sklearn
* tkMessageBox
* matplotlib

Processor: Intel i3or more

Ram: 2 GB or more

Hard disk: 40 GB hard disk recommended

**References:**

<https://nevonprojects.com/detecting-phishing-websites-using-machine-learning/>

<https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms>

<https://towardsdatascience.com/whataphish-detecting-phishing-websites-e5e1f14ef1a9>

<http://202.62.95.70:8080/jspui/bitstream/123456789/12636/1/1NH16CS054.pdf>

<https://www.ijert.org/detection-of-phishing-websites-using-machine-learning>

<https://www.activestate.com/blog/phishing-url-detection-with-python-and-ml/>

<http://norma.ncirl.ie/5100/1/nagasunderraopawarbaburaopawar.pdf>

<https://www.pantechsolutions.net/detecting-phishing-websites-using-machine-learning/>

<https://www.turkjphysiotherrehabil.org/pub/pdf/321/32-1-348.pdf>

<https://github.com/abhishekdid/detecting-phishing-websites>

<https://www.kashipara.com/project/python/3897/detection-of-url-based-phishing-websites-using-machine-learning-with-python>

Signature of Guide

Signature of Minor Project Co-ordinator :