**Classification and detecting Phishing website**

**Team Members:**

Latha Narayani Balaji

Mohammed Johny Shaik

Kushal Muktala

**Objectives:**

A phishing website is a social manipulation technique that imitates legitimate (URLs) and webpages. The purpose of this study is to use the dataset collected to train machine learning methods and deep artificial neural networks (Ann) to anticipate phishing websites. The required URL and website filters based on the content, attributes(features) are extracted from the dataset, which includes all phishing URLs of websites. Each model's performance is determined by comparison.

**Motivation:**

In 2020, the most common sort of cybercrime was phishing. - Federal Bureau of Investigation.

The best firewall or antivirus software can be implemented by any firm. However, most businesses do not have effective information security awareness initiatives for their employees. A single negligent employee can do a lot of damage to the organization.

**Significance:**

Phishing detection systems are essential in providing people to have a safe and secure online environment.

**Dataset:**

[https://archive.ics.uci.edu/ml/datasets/phishing+websites#](https://archive.ics.uci.edu/ml/datasets/phishing+websites)

**Classifiers Used:**

1. Random Forest Classifier
2. Decision tree
3. Logistic regression
4. KNN
5. XGBoost Classifier
6. ANN
7. Support Vector Machines
8. Multilayer perceptron

**Features:**

We are targeting close to 17 feature combing based on content and URL.

Table

Description automatically generated

**Background**:

Phishing is a typical cybercrime that involves sending a deceptive email or message to mislead recipients into visiting a fake website and then stealing sensitive information including credentials, ids, and credit card information for monetary benefit. An attacker first constructs a phishing website that closely resembles a legitimate website. On the one side, intruders forged the URL of the real website, particularly the DNS and overall network directories, using misspelled words, identical alphabet characters, and other tactics.

Building a large range of phishing sites is time-consuming for the phisher, with frequent modifications to URL or hostnames. It's plausible to assume that phishers use tools to create phishing sites. Although phishing websites built with tools have various URL or hostnames, the data of the phishing websites are presumed to be identical or similar. We observed several phishing websites that are structurally comparable to those built using recognized tools based on website structural features of our study of phishing sites recorded with Phish Tank.

**Design:**

Diagram, text

Description automatically generated

**Analysis:**

Chart, scatter chart

Description automatically generated

In T-SNE projection in component one we legitimate website URL. And in component two we have taken phishing website URLs. We use histogram plotting to see the data distribution.

A picture containing text, crossword puzzle

Description automatically generated

**Heatmap:**

A picture containing chart

Description automatically generated

Metrics for Evaluation:

1. Accuracy.
2. Recall
3. False Positive rate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithms | Train accuracy |  | Test  accuracy |  |  |
| Random Forest Classifier |  |  |  |  |  |
| Decision tree | 0.82 |  | 0.80 |  |  |
| Logistic regression |  |  |  |  |  |
| KNN |  |  |  |  |  |
| XGBoost Classifier |  |  |  |  |  |
| ANN |  |  |  |  |  |
| Support Vector Machines |  |  |  |  |  |
| Multilayer perceptron |  |  |  |  |  |

**Implementation:**

So far, we have acquired the required data set. We combined two data set for this project. We combined legitimate URL and phishing URL dataset. We are done with data exploration part. Now we are working on the classifier. decision tree performed well on this data set. Included the accuracy in the below table.

Implementation Status Report:

**Work completed**:

* + Creating the proposal.
  + Dataset selection --- we have combined two data set (phishing URL data set and legitimate URL dataset to create a dataset to fit our project requirement.
  + Data exploration – using python histogram we observed the data distribution.
  + Selecting the classifiers- we have selected eight classifiers for performance evaluation.
  + Python programming for the above specification.
  + Recording the accuracy.

**Work to be completed**:

We need to write code for rest of the seven classifiers. Based on the accuracy we have calculate Recall and FPR rate. Based on the score we need to evaluate the best performing model.

**Responsibility**:

Latha Narayani Balaji.

Programmer, Programming using python.

Creating the proposal.

Selecting the classifiers.

Mohammed Johny

Programmer, Programming using python.

Dataset selection.

Data exploration.

Kushal Muktala

Programmer, Programming using python.

Recording the accuracy.

**Contribution:**

|  |  |
| --- | --- |
| **Name** | **Percentage** |
| Latha Narayani Balaji. | **40%** |
| Mohammed Johny | **30%** |
| Kushal Muktala | **30%** |

**References:**

[1] <https://www.datatechnotes.com/2020/11/tsne-visualization-example-in-python.html>

[2] <https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1>

[3] Kulkarni Arun and Leonard L. Brown, Phishing Websites Detection using Machine Learning, 2019.

[4] Sadia Afroz and Rachel Greenstadt, PhishZoo: An Automated Web Phishing Detection Approach Based on Profiling and Fuzzy Matching, 2018.

[5]. Purvi Pujara and M. B. Chaudhari, Phishing Website Detection using Machine Learning: A Review, 2018.

[6] Aaron Blum, Brad Wardman, Thamar Solorio and Gary Warner, Lexical Feature Based Phishing URL Detection Using Online Learning, Alabama: Department of Computer and Information Sciences the University of Alabama at Birmingham, 2016.

[7] Rishikesh Mahajan, *Phishing Website Detection using Machine Learning Algorithms*, 2018.

**GitHub URL:**

<https://github.com/Lathabalaji/NLP_Project>