Project Proposal

PhishScan: Investigating Methods for Detecting Phishing Attacks

Group 10
Ian Oxley, Jack Hsieh, Naveen Kumar Manokaran, Siddharth Ranjan, Willem Grier

I. Introduction

1. Objective

Our primary objective is to identify whether a Uniform Resource Locator (URL) is a phishing link or a legitimate website based on information such as the domain, page ranking, active duration. We will be using different training processes and methods to classify a link as a phishing scam or legitimate. We will compare the process and the model accuracy between those methods and evaluate the trade offs.

2. Motivation

Phishing scam is some of the simplest yet effective scam method and it is causing more trouble around the world than it ever has. While the world was suffering from covid-19, these pesky scammers also made people suffer financially. “Phishing incidents rose by a staggering 220% compared to the yearly average during the height of global pandemic fears” (David Warburton, 2020)[1].

3. Expected Outcome

The first model we will try is Decision Tree and Random Forest. This includes feature preprocessing, feature selection, and feature extraction. We will also utilize techniques such as bootstrapping and boosting. The other method is using deep learning techniques such as (GNN) Graph Neural Network, (RNN) Recurrent Neural Network or a mix of both (GRNN). GNN is done by embedding various features into our graph as nodes and constructing a neural network to do the training.

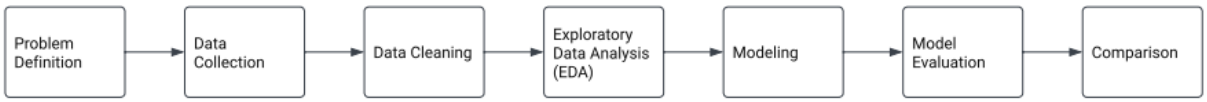
We expect to see a difference in accuracy between these methods. Cross validation will be used in both processes, giving us an idea of how well each performs. We will also compare the process of using each method. For example, how easy it is to tune the models, how fast they run, and how much work needs to be put into preprocessing the data. We will also compare the accuracy, precision, recall, F1-score of these models in classifying the phishing and legitimate websites. Ultimately, this research will contribute an understanding of which classification methods are most useful for the task of identifying phishing attempts.

4. Dataset

We will be using the “Phishing websites Data” dataset, which is hosted on Kaggle. Our dataset consists of 71677 data points with 11 features, including a text feature and 10 numerical features. The dataset will require some cleaning up before we can do the training or feed into our graph as embedding. The link to the dataset: https://www.kaggle.com/datasets/aman9d/phishing-data/data

II. Execution Plan

1.Steps to follow



Problem Definition:
This step involves defining the scope, objectives and constraints of our project.

Data Collection:
Gathering the required data for the machine learning and deep learning model.

Data Cleaning:
This step involves removing missing data,removing null values and transforming data.

Exploratory Data Analysis(EDA):
This step involves statistical and visual techniques to summarize the main characteristics of the data, identify patterns, detect anomalies, and explore relationships between variables.

Modeling:
Choosing the correct machine learning and deep learning model and training them.

Model Evaluation:
Assessing the performance of the trained models using appropriate evaluation metrics.

Comparison:
In this step we compare the performance of our deep learning and machine learning model based on various evaluation metrics and analyze their strengths and weaknesses in the phishing detection.

2.Workload Distribution

In our next meeting, we will first assign each individual member a method to implement. They will be responsible for completing all of the steps denoted in Steps to follow for that specific model. We have created a Gantt chart to provide an estimate of the timing for each of the tasks. We expect certain models to be more challenging than others to implement and are prepared in case one takes longer than expected.

3.Expected challenges and how to handle

We are expecting to reach challenges with version control and code collaboration throughout this project. Github and Google Collab will provide us an effective means to handle this challenge and collaborate effectively.

In Addition, each group member will individually develop and evaluate a model. As a result, each team member will need an understanding of each other's program so we can reproduce the results and evaluate whether they effectively utilized their model. We will handle this challenge of providing documentation so our peers can utilize our code and understand its mechanisms.

III. Evaluation Plan

1.Outcome Evaluation

The success of this project will be evaluated according to:

- Completion of implementation of each classification method
- Thorough and insightful comparison of classification methods

2.Performance Evaluation and Peer Review

The performance of individual members will be evaluated according to:

- Participation in group discussion
- Weekly communication of progress via Whatsapp
- Proper code maintenance and sharing

IV. Timeline

Jack (A)	Willem (B)	Ian (C)	Siddharth (D)	Naveen (E)	As a Team (ALL)

Tasks

1. Finalize selection of classification techniques: *Each team member will be assigned a method and will be the principal member responsible for implementing that method.*
2. Data Collection and Cleaning: *Clean data and distribute identical datasets to all members.*
3. EDA, Preprocessing, and Feature Extraction: *Discover patterns, spot anomalies, and test assumptions about input features.*
4. Implementation Stage 1: *Configure development environments and begin implementation.*
5. Milestone: *Report all findings utilizing the template provided for course evaluation*
6. Implementation Stage 2: *Complete implementation. Determine how to compare performance.*
7. Performance Evaluation: *Collect performance metrics in standardized format for each method.*
8. Synthesis and Report: *Compare methods and make conclusions.*
9. Final Report: *Collate findings in a concise report for course evaluation.*

Task ID	Est. Duration (Weeks)	2/19	2/26	3/4	3/11	3/18	3/25	4/1	4/8	4/15	4/22
1	1	ALL		SPRING BREAK							
2	1	ALL									
3	2	A-E	A-E								
4	2		A-E		A-E						
5	1				A-E						
6	3					A-E	A-E	A-E			
7	2							A-E	A-E		
8	1									ALL	
9	1										ALL

References

[1] David Warburton (Nov 11, 2020). 2020 Phishing and Fraud Report. F5 labs. https://www.f5.com/labs/articles/threat-intelligence/2020-phishing-and-fraud-report

[2] https://ieeexplore.ieee.org/document/9214132