CS235 Fall'22 Project Proposal: Detection of Phishing Websites

HRITVIK GUPTA #1, NetID: hgupt010
YASH AGGARWAL #2, NetID: yagga004
SHUBHAM SHARMA #3, NetID: sshar180
NITYASH GAUTAM #4, NetID: ngaut006
SIDDHANT POOJARY#5, NetID: spooj003

Additional Key Words and Phrases: data, mining, phishing websites, computer security, supervised learning, machine learning

ACM Reference Format:
Hritvik Gupta #1, Yash Aggarwal #2, Shubham Sharma #3, Nityash gautam #4, and Siddhant Poojary#5. 2022. CS235 Fall'22 Project Proposal: Detection of Phishing Websites. 1, 1 (October 2022), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A phishing website is a website that appears similar to a genuine/original website in terms of appearance and overall characteristics. Attackers use these phishing websites to get user data and deploy malware on users' machines. The objective of our project is to overcome these issues by using specific machine learning algorithms [Support Vector Machine, Perceptron, Decision tree, Naive Bayes' Algorithm, and K-Means Clustering] to predict the authenticity of a website by labeling them as benign or phishing using certain data features [Length of URL, Length of query parameters, Number of '@', '.', '/', '_' characters, etc.]. Further, we will compare the different models and then cross-validate the models with the best possible evaluation metrics.

1.1 Project Type

Software Type

2 PROBLEM DEFINITION

Phishing websites disguise as legitimate websites in terms of the domain name, URL, DNS, and other features of a legitimate website. This raises the chances of heavy malicious attacks on multiple machines of one or more users at a given time. Therefore, detecting phishing websites becomes essential to eliminate user interaction with these websites and, in turn, eliminate the chances of such malicious attacks. In this project, we will aim to classify if a website is phishing or benign based using ML algorithms.

Authors' addresses: Hritvik Gupta #1NetID: hgupt010; Yash Aggarwal #2NetID: yagga004; Shubham Sharma #3NetID: sshar180; Nityash gautam #4NetID: ngaut006; Siddhant Poojary#5NetID: spooj003.

Manuscript submitted to ACM

3 DATASET DESCRIPTION

The dataset is obtained from the University of New Brunswick's website: https://www.unb.ca/cic/datasets/url-2016.html, which contains separate excel files of benign, phishing, spam, malware, and other types of websites. Our project will only use legitimate and spam files as labeled data. Assuming that the number of legitimate websites is much greater than the number of spam, we will take only part of the entire dataset in comparable proportions to avoid bias. We will then combine the files, shuffle the rows, and split the data for testing and training. Also, we will create an excel file with the final output so as not to do this preprocessing every time.

4 PROPOSED APPROACH

The first and foremost task is pre-processing the data and extracting relevant features from the URLs. We will first separate the website's domain, URLs, IP, etc., of the phishing and legitimate websites. After getting the labeled data, we vectorize the feature using one-hot encoding and visualize the shape and size of the dataset. After that, we apply window segmentation to feed the dataset into the machine learning algorithm. In other words, we will divide the data into 1000, 2000, and 5000 samples and then compare the results to see if the machine learning algorithms do not overfit and get the best possible results. After getting the final results from the set of 5 machine learning algorithms applied, the final task is to compare and visualize each machine learning algorithm's performance upon different metrics. Finally, we finalized the best possible machine learning algorithm.

5 EVALUATION PLAN

Given the nature of the problem to find out the best possible algorithm which can function with high accuracy, the Confusion Matrix approach will be treated as the baseline for evaluating the performance metrics of all the algorithms. Using the baseline approach, algorithms will be evaluated against five parameters: Area Under the Curve, F1 Score, Accuracy, Precision, and Recall.

6 PROJECT TEAM & PROJECTED LABOR DIVISION

Team Members: HRITVIK GUPTA, YASH AGGARWAL, SHUBHAM SHARMA, NITYASH GAUTAM, SIDDHANT POOJARY.

Labour Division: Each team member will work on a different algorithm throughout the quarter, initially using an 'off-the-shelf' approach and then an approach coded from scratch. Further, all the team members will work together to write the reports, proofread them, and check for uniformity, accuracy and plagiarism. This work also includes but is not limited to finding corresponding research papers to look for inspiration, cite them as required, and work on presentation.

Algorithm Division: HRITVIK GUPTA [Support vector machine], YASH AGGARWAL [Perceptron], SHUBHAM SHARMA [Decision tree], NITYASH GAUTAM [Naive Bayes' Algorithm], SIDDHANT POOJARY [K-Means clustering].