Detecting Phishing Websites using Machine Learning

A

Mid Term Report

submitted in partial fulfillment of the
requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE

By :

Name	Roll No	Branch
Kushagri Pandey	R103215036	CSE-BAO
Garima Dixit	R144215007	CSE-MI
Amulya Yadav	R143215003	CSE-HI

Under the guidance of

Dr. Piyush Chauhan
Assistant Professor

UPES
UNIVERSITY WITH A PURPOSE

School of Computer Science
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
Bidholi, Via Prem Nagar, Dehradun, Uttarakhand

2018-19

CANDIDATES DECLARATION

We hereby certify that the project work entitled **Detecting Phishing Websites using Machine Learning** in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science And Engineering and submitted at School of Computer Science, University of Petroleum And Energy Studies, Dehradun, is an authentic record of our work carried out during a period from **January, 2019** to **May, 2019** under the supervision of **Dr. Piyush Chauhan, Assistant Professor**.

The matter presented in this project has not been submitted by us for the award of any other degree of this or any other University.

Name	**Kushagri Pandey**	**Garima Dixit**	**Amulya Yadav**
Roll No.	**R103215036**	**R144215007**	**R143215003**
Branch	**CSE-BAO**	**CSE-MI**	**CSE-HI**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

(Date: 13 March 2019)

Dr. Piyush Chauhan
Project Guide

Dr. T.P. Singh

Head
Department of Informatics
School of Computer Science
University of Petroleum And Energy Studies
Dehradun - 248 001 (Uttarakhand)

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Dr. Piyush Chauhan**, for all advice, encouragement and constant support he has given us through out our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank to our Head of the Department, **Dr. T.P. Singh**, for his great support in doing our **Detecting Phishing Websites using Machine Learning** at **SoCS**.

We are also grateful to **Dr. Manish Prateek Professor and Director SoCS** and **Dr. Kamal Bansal Dean CoES**, UPES for giving us the necessary facilities to carry out our project work successfully.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Kushagri Pandey**	**Garima Dixit**	**Amulya Yadav**
Roll No.	**R103215036**	**R144215007**	**R143215003**
Branch	**CSE-BAO**	**CSE-MI**	**CSE-HI**

ABSTRACT

There are number of users who purchase products online and make payment through various websites. There are multiple websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons. This type of websites is known as phishing website. In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on using classification Data mining algorithm. We implement classification algorithm and techniques to extract the phishing data sets criteria to classify their legitimacy. The 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 online when he makes payment through the website our system will use data mining algorithm to detect whether the 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. Admin can add phishing website url or fake website url into system where system could access and scan the phishing website and by using algorithm, it will add new suspicious keywords to database. System uses machine learning technique to add new keywords into database.

Keywords: Data mining, Machine Learning, Phishing.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

List of Figures

1 Introduction

Phishing [1] is a cyber attack that uses disguised email as a weapon. The goal is to trick the email recipient into believing that the message is something they want or need — a request from their bank, for instance, or a note from someone in their company — and to click a link or download an attachment.

What really distinguishes phishing is the form the message takes: the attackers masquerade as a trusted entity of some kind, often a real or plausibly real person, or a company the victim might do business with. It's one of the oldest types of cyberattacks, dating back to the 1990s, and it's still one of the most widespread and pernicious, with phishing messages and techniques becoming increasingly sophisticated.

Types of phishing

Hand over sensitive information. These messages [2] aim to trick the user into revealing important data — often a username and password that the attacker can use to breach a system or account. The classic version of this scam involves sending out an email tailored to look like a message from a major bank; by spamming out the message to millions of people, the attackers ensure that at least some of the recipients will be customers of that bank. The victim clicks on a link in the message and is taken to a malicious site designed to resemble the bank's webpage, and then hopefully enters their username and password. The attacker can now access the victim's account.

Download malware. Like a lot of spam, these types of phishing emails aim to get the victim to infect their own computer with malware. Often the messages are "soft targeted" — they might be sent to an HR staffer with an attachment that purports to be a job seeker's resume, for instance. These attachments are often .zip files, or Microsoft Office documents with malicious embedded code. The most common form of malicious code is ransomware — in 2017 it was estimated that 93 percent of phishing emails contained ransomware attachments. Phishing detection has been implemented by modeling the previously detected phishing websites. Each web-site (phishing and non-phishing) is compromised of a set of items and properties. These items and properties can be described as variables. In a specific-website, these variables can be assigned avalue based on the content of that web-site. Thus, the problem can be formed as a detecting of future phishing websites based on modeling the previously detected websites using data mining processes. Classification, as one of the most important category of the data mining processes, and one utilized with the phishing detection, is the process of predicting the output class (family/category) of a given input with unknown class. To use classification in phishing detection, the modeled website should be described based on predetermined set of properties/features.

2 Literature Review

SVM-The objective [3] of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N—the number of features) that distinctly classifies the data points.To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.Hyperplanes [4] are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.Support vectors

are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM. USE OF SVM- SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined.Well SVM it capable of doing both classification and regression. In this post I'll focus on using SVM for classification. In particular I'll be focusing on non-linear SVM, or SVM using a non-linear kernel. Non-linear SVM means that the boundary that the algorithm calculates doesn't have to be a straight line. The benefit is that you can capture much more complex relationships between your datapoints without having to perform difficult transformations on your own. The downside is that the training time is much longer as it's much more computationally intensive.

RANDOM FOREST CLASSIFIER-Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, you don't have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does). USE- Random Forest is also considered as a very handy and easy to use algorithm, because it's default hyperparameters often produce a good prediction result. The number of hyperparameters is also not that high and they are straightforward to understand. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this, that measures a features importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1.

3 Problem Statement

With the rise in the number of phishing websites it has become increasingly insecure for users to share their personal details with a website e.g. bank details, account details, PAN details etc.

4 Objective

To detect phishing website using the URL and classify it on the basis of its features.

5 Pert Chart

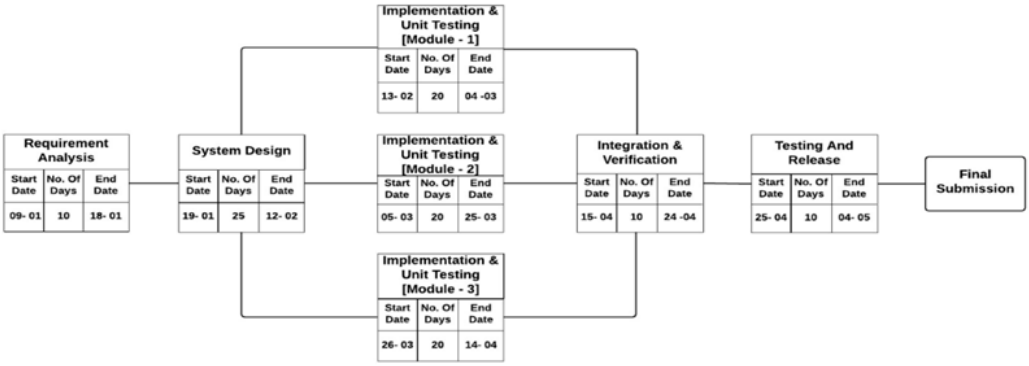

Figure 1: Pert Chart

6 Flow Chart

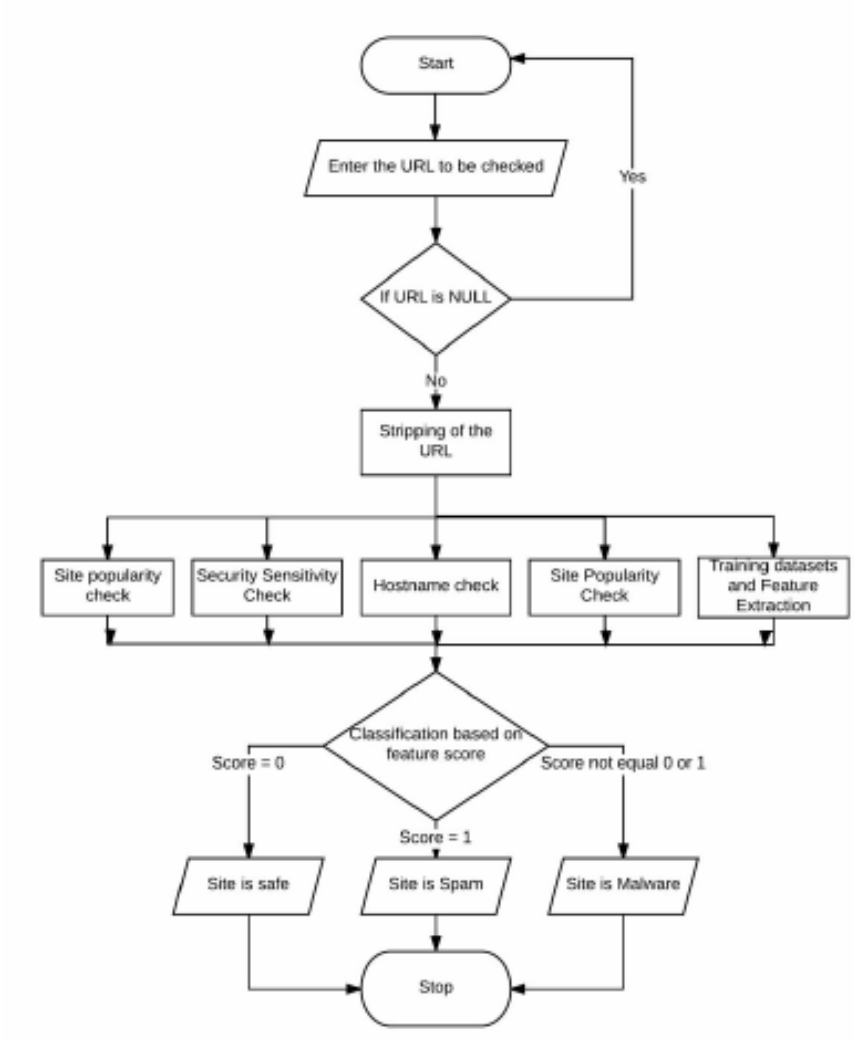

Figure 2: Flow Chart

7 Design Methodology

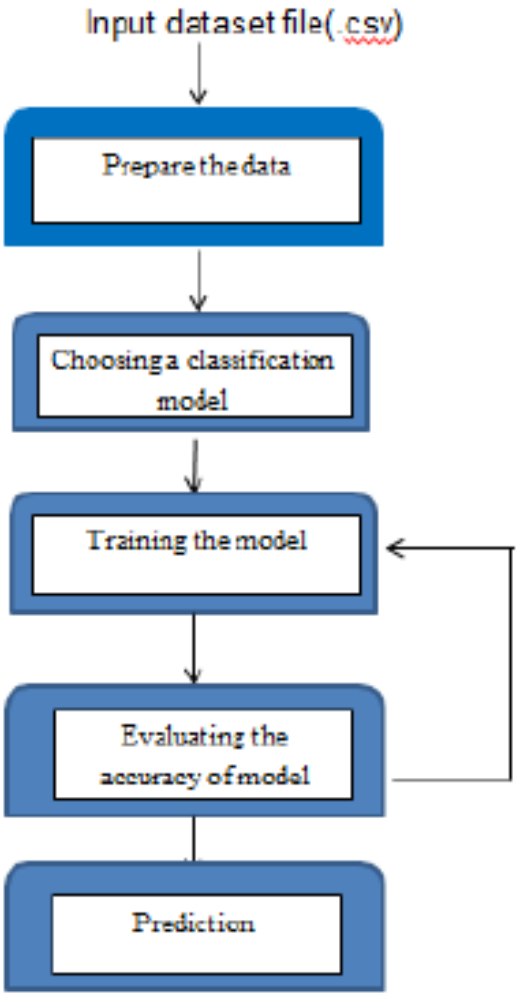

Figure 3: Methodology

References

[1] J. James and C. Thomas, "Detection of Phishing URLs Using Machine Learning Techniques," no. February, 2019.

[2] E. P. Pujara, "Phishing Website Detection using Machine Learning : A Review," no. February, 2019.

[3] H. Sampat, M. Saharkar, A. Pandey, and H. Lopes, "Detection of Phishing Website Using Machine Learning," pp. 2527–2531, 2018.

[4] I. Vayansky and S. Kumar, "Phishing – challenges and solutions," vol. 3723, no. April, 2018.