A Comprehensive Feature Selection And Classification Methods For Phishing Detection

Pravin Dattatray Pawar

(MSc Data Science)

Thesis Proposal

December 2019

Table of Contents

Abstract 2

1. Background 3

2. Problem Statement 5

3. Aim and Objective 6

4. Significance of the Study 6

5. Scope of the Study 6

6. Research Methodology 6

7. Required Resources 6

8. Research Plan 6

References 6

# Abstract

ToDo

Asd

# Background

Phishing is criminal activity carried out using social engineering technique. It’s kind of cyber-attack in which attackers/phishers try to steal sensitive information of internet user (Chiew et al., 2019). Attacker target user via online or offline methods such as phishing website, email, SMS, fake calls, etc. Most commonly used phishing method is phishing website, as an exponential growth in internet. Many peoples are now relying on internet services for personal as well as business purpose, hence internet fraud is biggest threat to internet users. Through phishing website attacker try to get users personal information like date of birth, userid, password or financial information like bank details, credit card details and related passwords. The core intension to steal such information is, to get some financial benefit by using this information or selling such sensitive data to external word (Rao et al., 2020).

Anti-Phishing Working Group [APWG] tracks number of phishing website present across the globe. As per APWG Q3 report, 266,387 number of phishing website detected in third quarter of 2019 which is 46% raise in Q3 2019 as compare to Q2 2019 with 182,465 phishing websites and approx. 100% raise as compare to Q4 2018 with 138,328 phishing website (Anti-Phishing Working Group, 2019). It shows significant increase in phishing website. OpSec security found that 30.80% phishing attack targets SaaS [Software as a Service] and webmail sector followed by 19.80% on Payments and 19.40% on Financial Industry(Anti-Phishing Working Group, 2020).

Phishing attacks can be of two types, Malware phishing and Deceptive phishing (Sumathi & Sujatha, 2019).

* 1. **Malware Phishing**

Malware means malicious software which is harmful to computer system/user, intention to develop such software is to damaging or gaining access of user’s system it can be personal or office laptop, mobile device. Trojan, worms, spyware, ransomware are some examples of malware. Phisher attach harm full software or script to email or website made to look useful and prompt user to download or install.

* 1. **Deceptive Phishing**

In deceptive phishing attack, phisher try to impersonate a legitimate entity such as company, financial institute, etc. and try to steal sensitive information. Such attacks use following methods to phish target.

* Website: Phisher identify target website based on popularity & exposed vulnerability.

Because of this it become very difficult for the user to identify phishing website by just eyeballing website URL (Uniform Resource Locators) or content which looks like legitimate website. When user faces phishing attack, user actually want to visit some legitimate website and end up landing on some different site which looks like legitimate site but it was not.

* Email: Once fake website is ready, Phisher run email campaign to circulate URL. These emails contain some sort of offers of top branded sites or some sort of urgency asking user to act on it immediately. It copies contents like text, logo or style from legitimate website so that it appears to be a genuine email.
* Social Media / SMS: Such website URL also circulated via SMS or other chatting application as well as via social media pages like facebook, twitter
* Phone Call: In phishing call, phisher pretend to be from some organization or some financial institute and try to acquire personal information over call like date of birth, address, credit card details, etc.

Damage of such attack can be temporary or permanent loss of any online account like social media, bank, some organization specific account or any financial loss. It can be a financial as well as reputational loss for organization. According to Federal Bureau of Investigation (FBI) report 2018, losses associated with business email compromises [BEC] had cost organization $5.3B from Oct 2013 to Dec 2016 where as in financial year 2017 more than $675M losses occurred because of BEC.

Motivation behind this study to safeguard internet users from internet frauds occurred via phishing website, as such malicious URL’s are present all over the internet which can impact many internet users. This study proposed improved solution to detect phishing URL using different features of website URL as well as website content with better accuracy.

# Problem Statement

Phishing attack become very critical when it hits banking and financial industries as it can lead to financial losses of normal users as well as financial institute or banks it can also damage reputation of such banks which eventually impact the overall business.

Various studies conducted on phishing detection problem. Many different methods are used to detect phishing URL such as listing approach, content based and URL based approach using machine learning algorithms applied on different data sets.

* 1. **List-based approach**

Most commonly used list-based approach is Blacklist URL.

ToDo

# Aim and Objective

The main aim of this research is exploring domain of web security specifically phishing detection and propose machine learning based model to detect such website using content and URL based features. The goal of this research is, safeguard internet user activity by providing real time phishing detection.

The objective of the research based on proposed aim are:

* To investigate existing phishing detection technique and there performance.
* To determine important feature using different feature extraction technique.
* To develop classification-based prediction model to detect phishing website
* To evaluate and compare performance of predictive models and proposed best model

# Significance of the Study

# Scope of the Study

# Research Methodology

# Required Resources

# Research Plan

# References