INFORMATICS INSTITUTE OF TECHNOLOGY

In collaboration with the

University of Westminster, UK

A Machine Learning Approach to Detect Image Based Web Phishing Attempts

A dissertation by

Sandun Abeysooriya

Supervised by

Ms. Jayani Harischandra

Co-Supervised by

Mr. Guhanathan Poravi

Submitted in partial fulfillment of the requirements for the BEng (Hons) Software Engineering Degree Department of Computing

November 2020

Abstract

Phishing is one of the oldest type of cybercrimes which tricks people to obtain their sensitive information such as usernames, passwords, and credit card information. Time to time the attackers have changed their methods for phishing to avoid being detected by the security protocols. If a phishing website is hosted under a legitimate domain or if the website content contains only images, it would be able to avoid being detected by the security system algorithms since there's only a slight chance of identifying any suspicious elements from such sites to make a prediction.

In this proposed machine learning approach to detect image-based web phishing attempts, the main aim was to develop a system that would be able to detect phishing websites that uses image-based attempts to avoid being detected. This image-based phishing detection system has taken a novel approach by using combined sets of features to identify any suspicious elements from the site such as URL, domain, content, and images. A brand new dataset has been compiled with previously addressed element data in order to train a high accurate model with the use of best fitting ml algorithm. In this case random forest algorithm has been selected for the model classification after comparing with several other algorithms. Compiled dataset was fed up with enough phishing and legitimate URL data from various source with unique elements to maintain a higher accuracy.

A finalized prototype application has been developed based on the proposed solution for the end users. The development was carried out in a Python environment since it goes well with data science related projects. Before finalizing the project, the implemented system was tested properly according to a certain criteria and evaluated with the guidance from domain and technical experts in order to confirm that the taken approach was a success.

Keywords: Phishing Detection, Machine Learning, Random Forest, Python.