DETECTION AND CATEGORIZATION OF MALICIOUS URLS WITH A DEEP LEARNING APPROACH

Don Manula Ransika Udugahapattuwa

A dissertation submitted in partial fulfillment of the requirement for
Master of Science in Big Data Analytics

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
Robert Gordon University, Aberdeen

2022

Abstract

In this 21st century, the world is being digitalized each and every day. The Covid pandemic made the digitization even faster through the use internet. Uniform Resource Locators abbreviated as URLs are publicly accessed by anyone who will be navigating through the internet. Therefore, URLs are a great tool for cyber threat actors (people who means harm through cyber space) to utilize in order to attack anyone who tries to access a website.

The proposed system will utilize deep-learning-based binary and multiclass machine learning engines to identify if a URL is malicious or benign. If the URL is malicious, the multiclass classifier will categorize it under one of four available cyber threat categories.

The system has been trained well and has acquired over 90% accuracy in multiple deep learning algorithms namely Multilayer Perceptron, Keras-Tensorflow based model and FastAI based model. The evaluation process has taken feedback from academic personnel as well as industrial experts while conducting self-evaluations in both quantitative and qualitative measures in order to identify the strengths and project improvements

Key words: URL categorization, deep learning, Malicious URL detection, Neural Networks