INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Malicious URL detection based on machine learning.

A Project Proposal by

Mr. Udayasanthiran Ahtshayan

Supervised by

Mr. Sharmilan Somasundaram

Submitted in partial fulfilment of the requirements for the BEng (Hons) in Software Engineering degree at the University of Westminster.

November 2023

ABSTRACT

The rise in internet usage has coincided with an increase in cyber dangers, notably the prevalent issue of rogue URLs. These URLs are frequently used in phishing scams, malware distribution, and other types of criminality. Because of the quickly developing nature of these threats and the difficulty to scale efficiently to address the vast amount of URLs created everyday, current approaches for identifying malicious URLs, such as blacklisting or rule-based systems, have proven ineffective. As a result, there is an urgent need for a more effective, precise, and scalable method of detecting and neutralizing the dangers posed by bad URLs.

In answer to this issue, the author has developed a sophisticated machine learning model based on Logistic Regression and TfidfVectorizer. To categorize URLs as benign or dangerous, Logistic Regression, a machine learning approach generally used for binary classification issues, was applied. TfidfVectorizer, a feature extraction method that turns text data into numerical vectors, was utilized, on the other hand, to convert the URLs into a format acceptable for the Logistic Regression model. This approach provides a score to each token in the URL, depending on its frequency in the URL and rarity in the total dataset. The model was trained on a huge dataset of URLs that had been categorized as benign or dangerous.

The model's performance was evaluated using essential data science metrics for binary classification tasks, such as accuracy, precision, recall, and the F1 score. Testing was carried out on a different dataset from the training set. The model demonstrated good accuracy, indicating its ability to accurately categorize URLs. Precision was also high, indicating a low proportion of false positives, and the model's ability to catch the bulk of dangerous URLs was supported by a high recall score. The F1 score, which is a harmonic mean of accuracy and recall, attested to the model's solid performance even further. This novel way to detecting fraudulent URLs marks a big leap in the world of cybersecurity.