TU Clausthal

Abschlussvortrag Masterarbeit Harini Sree Boinpally

„Phishing URL detection using Machine Learning in Cyberspace“

Phishing is a method that tricks the users into revealing sensitive information and is a common method for cyberattacks. According to statistics, the number of phishing attacks has increased by over 150% annually since 2019. The advancement in AI is helping cybersecurity researchers classify phishing URLs with the machine and deep learning algorithms. This study focuses on the classification of phishing URLs using machine learning algorithms. URL features play a vital role in building machine learning models to attain the efficient classification of Phishing URLs. We analyzed the URL features (F1, F2) of two datasets: UCI2016 and Mendeley2020. Datasets were built by extracting F1 and F2 from URLs found on PhishTank and Tranco and categorized them as old URLs (OU) or new URLs (NU) based on the year they were collected. The datasets built are referred to as OUF1, NUF1, OUF2, NUF2. Various machine learning models are built on these datasets to analyze the performance of each model with respect to each dataset. The models' accuracies dropped when trained on NUF1, with the RSVM model achieving only 58% accuracy. The reasoning for the drop in the accuracies is found to be that the feature set (F1) consists of javascript and HTML based URL features that are not ex-tractable as they have become obsolete. Additionally, the feature based on external servers such as the rank of the page has posed an access restriction while extracting. These reasons lead to the insufficient extraction of URL features in some instances thereby attaining low accuracies. An exploration of the Mendeley2020 dataset was made by extracting URL features and training various models. The models achieved good accuracy with the new URL dataset at 97.7% and the old URL dataset at 95.4%. Feature engineering techniques were applied to improve the models' performance, resulting in a 1.4% increase in accuracy on old URLs. Our analysis of Mendeley2020 has revealed that certain characters such as ",_,:,?" are not compliant with domain name system standards, and are counted as domain-based features. Therefore, it is crucial to select relevant and appropriate URL features to build machine learning models that can ensure effective classification of phishing URLs.

Betreuer der Arbeit:	Prof. Dr. Mohammad Ghafari, Prof. Dr. Benjamin Leiding
Datum:	Montag, 19. Juni 2023, 17:00 Uhr
Ort:	Center for Digital Technologies Besprechungsraum 2.04 Wallstraße 6 Goslar Webkonferenz: https://webconf.tu-clausthal.de/b/car-s3a-9pu-bno Code: 845949