Final Year Project (CS4E2) Overview

(By - Arnav Malhotra, 17317424)

Motivation:

Malicious URLs/websites are quite common and a serious threat to cybersecurity hosting unsolicited content (spam, exploits, phishing etc.). These lure a lot of unsuspecting users who become victims of various kinds of scams resulting in loss of billions of dollars every year. To prevent this, blacklists are usually used but since they are not exhaustive, they lack the ability to detect latest malicious URLs. However, some machine learning techniques have given promising results. One such technique is Feature based approach but feature extraction is a tedious process and requires domain expertise. Hence, the objective of this project is to build a classifier which can detect malicious URLs using a **Featureless Deep Learning** approach.

Feature-based VS Featureless Approach:

First step of Feature-based approach requires Feature Analysis which can be done by analysing 4 sets of features of the URL- Blacklist, Lexical, Host-based, Content-based. Then, those features are passed onto the machine learning classifier (such as Logistic Regression, k-Nearest Neighbours, Decision trees) for classification.

On the other hand, the Featureless approach uses Word2Vec for pre-processing of the URL. This has two hidden layers which are trained to “embed” characters the are close-by in n-dimensional space. For classification, it can use any kind of neural network (such as a DNN, CNN or RNN).

The drawback of using Featureless approach is that it requires much more data and training time as compared to Feature-based approach.

Design of Experiment:

The dataset being used is balanced which contains 194798 URLs in total. The size of the dataset might result in overfitting as comparatively larger datasets are recommended when using the Deep Learning classifiers.

Architecture of Feature-based approach-

1) Feature extraction
2) Logistic Regression / k-Nearest Neighbours / Decision Trees

Architecture of Featureless approach-

1) Word2Vec
2) Simple LSTM / 1D Convolution + LSTM / 1D Convolution + Fully Connected Layers

Results:

A basic model of feature extraction was implemented along with a bunch of classifiers to check the efficiency of the Feature-based approach. Classifiers are mentioned in decreasing order of performance; K-Nearest Neighbours, Decision Trees, Logistic Regression.

Planned Results:

In case of Featureless approach, 1D Convolution + Fully Connected Layers architecture is expected to perform best out of the 3. Since the Feature extraction model implemented is very basic, performance of Word2Vec is expected to be better.