Detecting Malicious URLs Using Lexical Analysis

Here are features for lexical analysis

Table

Description automatically generated

Experiment Summary

1. data collection:

Around 114,400 URLs were collected initially containing benign and malicious

URLs in four categories: Spam, Malware, Phishing and Defacement.

Four single class datasets by mixing benign and malicious URLs and one multi-class dataset by combining all four malicious URLs and benign URLs.

2. feature selection

CFSSubsetEval and Infogain as feature selection algorithms.

CfsSubsetEval evaluates the worth of a subset of features by considering the

individual predictive ability of each feature along with the degree of redundancy.

Infogain searches the space of feature subsets by greedy hill-climbing strategy

augmented with a backtracking facility.

3. Classification:

\* K-Nearest Neighbours algorithm (KNN)

\* C4.5 and RandomForest

ARCHITECTURE as follow:

Diagram

Description automatically generated

4. Analysis and result

We observe that tree based classifiers, with Random Forest yields highest

accuracy among the classifiers tested.

A close up of text on a white background

Description automatically generated