Research Proposal

Subject:
Big-data URL classification by NLP features

Team members:
Sari Saif 324832914
Simcha Teich 323104562

Paper:
Machine learning based phishing detection from URLs
Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri

Abstract:
The article refers to the world of "cat and mouse" in which the attacker observes how he is discovered, and accordingly perfects his methods. In our case - URL phishing.
URL phishing focuses on exploiting human weaknesses - so their language must be taken into account. When extracting the features, one should refer not only to basic data such as "amount of points in the URL" or its size, because the attacker can easily play with them. But also to take gibberish words into account, and words that only resemble real words, and words that are brand words, and a word that consists of several words. Through this analysis, it will be difficult for the attackers to bypass the identification mechanism because it is dynamic and changes per language.

After extracting the features (which is the main work) the article presents comparisons between 7 models and chooses the best of them.
The models are: Naive base, Random forest, knn (k=3), Adaboost, k*, SMO And Decision Tree.
There are 40 NLP features, and 1000+ regular ones. After the CfsSubsetEval filter algorithm, 102 filters remain.

The results show that random forest algorithm combined with NLP features only, gives the best result (accuracy) at a rate of 97.98%

In our work we will try to find additional NLP features that will increase the percentage of success.

Related work
Our related work is divided into two types: filtering by lists (black list and white list) and machine learning.
In related works containing machine learning, two previous articles by the researchers themselves stand out (in the current article they improve upon them).

https://ieeexplore.ieee.org/document/8093406
https://link.springer.com/chapter/10.1007/978-3-319-76348-4_59

Malicious URL Classification Using Extracted Features, Feature Selection Algorithm, and Machine Learning Techniques (https://scholar.google.com/scholar?hl=iw&as_sdt=0%2C5&q=Malicious+URL+Classification+Using+Extracted+Features%2C+Feature+Selection+Algorithm%2C+and+Machine+Learning+Techniques&btnG=)

Dataset:
We found the dataset of the researchers themselves, it contains about 35000 legitimate URLs and also about 35000 malicious URLs.
Each record comes with 2 columns - the URL itself, and a label (malicious or not). We already do the feature extraction ourselves.

https://github.com/ebubekirbbr/pdd/tree/master/input

To get even better results we can train the model with additional datasets that can be easily found on kaggle.com.