Phishing URL Detection

Ilan Sirisky, Nir Meir

February 2023

Abstract:
Phishing is a considerable problem that differs from the other security threats such as intrusions and Malware which are based on the technical security holes of the network systems. The weakness point of any network system is its Users. Phishing URLs mainly target individuals and/or organizations through social engineering attacks by exploiting the humans' weaknesses in information security awareness. These URLs lure online users to access fake websites, and harvest their confidential information, such as debit/credit card numbers and other sensitive information.

1 Introduction

With the steady acceleration in information technology, we are no longer immune to being victims of cybercrime. The use of the Internet has become essential in the modern era and an integral part of technological development, which leads to discoveries and reduction of time, effort, and costs.

This work focuses on a URL phishing attack that directly depends on social engineering and specifically targets individuals by deluding them with fake websites in which the victim falls prey to these hackers , give his sensitive information such as e-mail account and other sensitive information related to credit card details and confidential information that may affect the reputation of the individual or institution.

There have been many studies recently attempting to come up with a suitable solution for detecting phishing URLs. These solutions can be grouped into commonly four classifications: predefined list, signature-based, content-based, and machine learning.

In this work, we will focus on the machine learning classification technique. The machine learning approach mainly depends on the ability to learn the characteristics of the websites that are listed under the phishing category, then implementing the prediction ability to distinguish the legitimate websites from the faked ones in the means of the different machine learning techniques (Prediction, Classification, etc.)

2 Related Work

Nowadays, many anti-phishing techniques are proposed, but still, there is a challenge to get high accuracy detection with a low ratio of false-positive detection. In this section, a review of related work techniques and their features is presented.

A study in 2016 [1], used a real-time detection system approach using URL features only, a dataset of 46,5461 URLs was used with three classifiers (J48, SVM, and Logistic Regression) the highest accuracy was 93% which was gained by J48 classifier.

In 2017 authors of [2], implemented a middleware system to detect phishing websites. Multiple algorithms, including Random Forest, SVM, and K-Nearest Neighbor (KNN); a dataset of 11055 URLs were collected, the highest accuracy 96% was obtained using RF algorithm.

A model in 2018 was proposed by authors in [3] using a URL identification strategy utilizing the Random Forest algorithm. A dataset was gathered from PISHTANK; only 8 out of 30 features were used for analysis. Finally, an accuracy of 95% was achieved by this model. Where authors of [4] used a new design called Extreme Learning Machine (ELM) based on the RF algorithm using 30 URL features, ELM detecting accuracy was 95.34%.

From another perspective, authors in [5] proposed a technique through content analysis and URL features extraction. Artificial Neural Network, Decision Tree, Support Vector Machine, K-Nearest Neighbor and Naive Bayes algorithms were used in this approach. The highest accuracy (96.01%) was obtained using Artificial Neural Network algorithm.

And lately in 2021, authors of [6] used a dataset of 1056937 labeled URLs which was processed to generate 22 features that were reduced further to a smaller set. Random Forest, Gradient Boosting, Neural Network and Support Vector Machine (SVM) classifiers were all evaluated, and results show the superiority of SVMs, which achieved the highest accuracy in detecting the analyzed URLs with a rate of 99.89%.

Table 1 Summerization of previous works and their techniques

Author	Used algorithms	Accuracy
[1]	J48, SVM and LR	93% using J48
[2]	RF, SVM and kNN	96% using RF
[3]	RF	95%
[4]	RF	95.34%
[5]	SVM and NB	90% using SVM
[6]	RF, GB, NN and SVM	99.89% using SVM

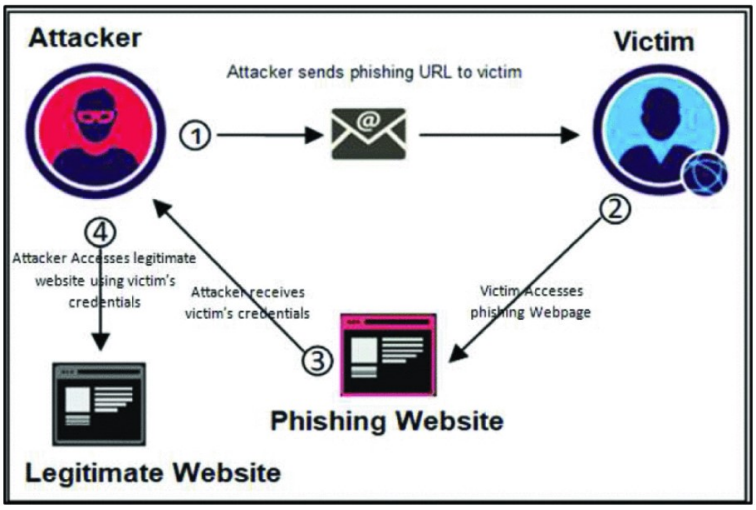

Figure 1: Website Phishing Life-cycle

3 Data set Description

In this work we had two different datasets:

The first dataset we found at the Kaggle website which contained 160k legitimate URLs and 100k phishing URLs. But because we didn't want to work with imbalanced data because it can cause over fitting and simmilar problems. So we added another 60k verified phishing URLs from an open-source service called PhishTank. This service provide a set of phishing URLs in multiple formats like csv, json etc. That gets updated hourly by users. So overall we have 320k URLs in this dataset.

The second one we used is a dataset that has a total of 675k URLs. The special thing about this data set that 50% of dataset was generated by Domain Generator Algorithms (DGAs). The 337.5k URLs are equally distributed among 25 DGA families, thus we have 13.5k domains for each family. Among such DGAs, 13 are time dependent algorithms and 12 are time independent ones. The rest of the 337.5k URLs are legitimate URLs that are verified by Alexa 1 Million sites.

In each of the datasets we split the data in a 80:20 ratio, 80% used for training and the remaining 20% is used for testing.

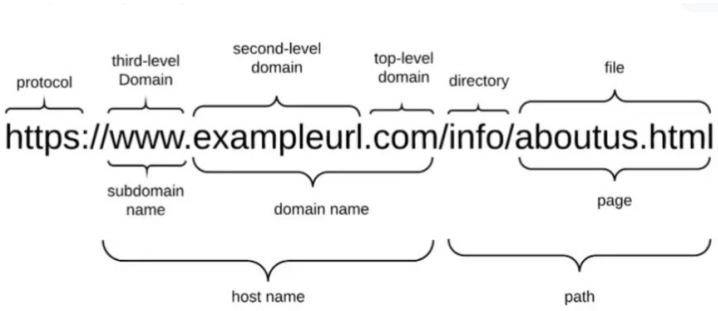

Figure 2: URL breakdown

4 Results

Like we stated above in the dataset description section, we had 2 different datasets. So for each data set we trained a number of classification models, with different features for each dataset.

In the first dataset, we extracted 19 features overall and ended up using 17 of them. We categorized them into 3 groups: 1, Length based features. 2, Count based features. 3, Binary features.

The features in the first group are: Length of URL, Length of Host-name, Length Of Path, Length Of First Directory, Length Of Top Level Domain.

The features in the second group are: Counts of '@', '?', '%', '.', '=', 'http', 'https', 'www'. As well as number of digits, letters and re-directions.

The third group contained: if the URL is using and IP address and if the URL is using a shortening service.

We trained 3 models on this dataset and in all of them we obtained an accuracy and f1-score of 100%. The models are: Decision Tree, Random Forest and a MLP algorithm using deep learning.

In the second dataset, we extracted 13 features overall and ended up using 10 of them. Here as well we can categorize them into groups:

1. Length based features: Domain Name Length, Subdomain Length Mean.

2. Binray based features: Has a Valid Top Level Domain, Contains TLD as Subdomain, Contains Digits.

3. Ratio based features: Vowel Ratio, Digit Ratio, Ratio of Repeated Characters in a subdomain, Ratio of consecutive digits, Ratio of Consecutive Consonants, Entropy of subdomain. And an additional feature: Number of Subdomains.

On this dataset we trained 4 classification models:

1. Decision Tree: we obtained an accuracy of 75% and f1-score of 79%.

2. Random Forest: we obtained a similar accuracy of 76% and f1-score of 79%.

3. Gradient Boosting Classifier: we obtained the highest accuracy of 83% and f1-score of 84%.

4. MLP algorithm: we got an accuracy of 76%.

5 Summary

Our study found that achieving 100% accuracy in detecting random phishing URLs was relatively easy based on our work and previous studies. However, the constant creation of new phishing domains presents a persistent challenge. To address this, we collected additional data generated by DGAs and trained a new model to work in conjunction with the existing models, expanding the range of detectable phishing domains. Our results suggest that using an ensemble approach with the best models from both datasets can increase the likelihood of accurately detecting phishing domains.

6 References

1. Wide scope and fast websites phishing detection using URLs lexical features
2. Malicious web content detection using machine leaning
3. A New Method for Detection of Phishing Websites: URL Detection
4. Phishing web sites features classification based on extreme learning machine
5. Malicious Web Page Detection: A Machine Learning Approach
6. URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis