ISTANBUL KÜLTÜR UNIVERSITY

Department of Computer Engineering

2020-2021 Fall

CLASSIFICATION OF PHISHING WEBSITES BASED ON MACHINE LEARNING TECHNIQUES

1600003764 Safa ORHAN, 1600003762 Eyüp USTA

Project Advisor: Assis. Prof. Dr. Öznur ŞENGEL

Project Description

Phishing aims to obtain users' credentials by cheating them via presenting fake web pages which appearing legitimate sites. For example, you receive mail from a bank or a company to your e-mail address. You trust it to be a well-known organization and click the URL in the incoming mail and your information is stolen. Phishing detection mechanisms are grouped into heuristics based, visual similarity based, list based, and Machine Learning (ML) based technical categories. This study is based on ML and Deep Learning (DL) techniques. In this project, a dataset will be used to compare ML and DL methods while detecting phishing websites. Also, precision, accuracy, F1-score, and recall will be evaluated to check DL and ML algorithms.

The purpose of this project is to classify phishing sites using ML and DL techniques, to facilitate the detection of such sites. To develop various models in the classification process. To reach the best result by testing these developed models on the data set. There are many problems in phishing such as people, security vulnerabilities, and attackers constantly updating themselves. As a result of our work, we will propose a model against attackers trying to exploit people's weaknesses using seemingly safe URLs. Thus, it will be more effective against fake e-banking, e-commerce style URLs coming via e-mail.

The objective of the project is to find and develop the model that gives the best result, that is, the closest to 100% value. It is our success criterion that the model we will obtain as a result of the studies has the closest value to 100% in comparison methods. It is impossible to obtain 100% value in such a study.

Design Summary

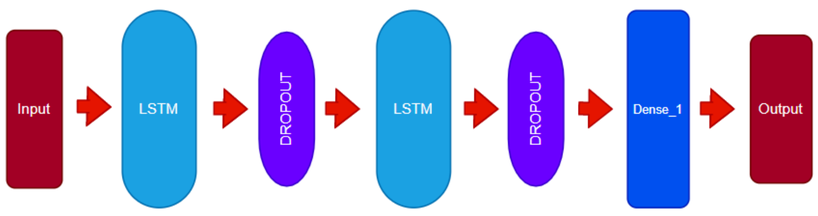

Figure 1. RNN Model

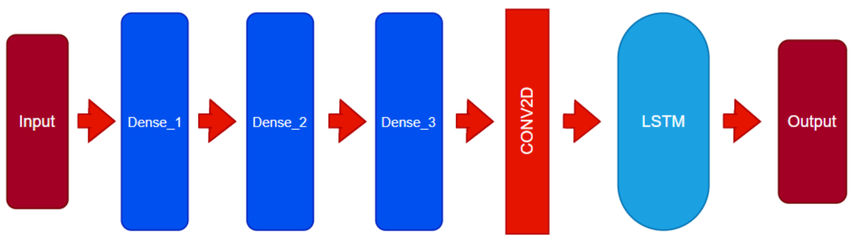

Figure 2. Combined Model

Implementation Summary

Random Forest Model:
Max features: auto
Class weight: balanced subsample
Warm start: true
Criterion: gini
Accuracy: 99.09%

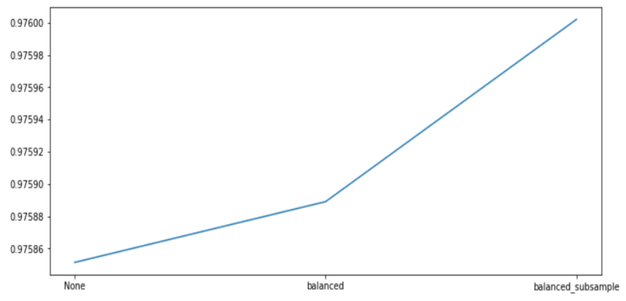

Figure 3. Average accuracy values according to class weights

RNN Model:
Optimizers: Adamax
Activation Functions: Sigmoid
Loss Functions: Binary cross entropy
Accuracy: 98.31%

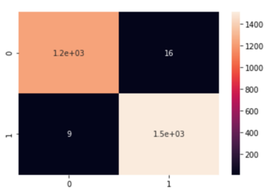

Figure 4. RF's Confusion Matrix

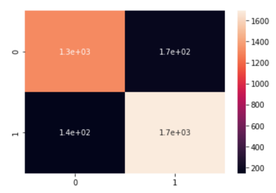

Figure 5. RNN Model's Confusion Matrix

Combined Model:
Combination: CNN – LSTM
Optimizers: Adamax
Activation Functions: Softplus
Loss Functions: Binary cross entropy
Accuracy: 96.26%

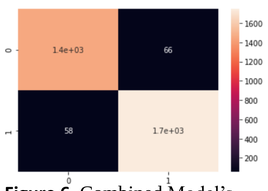

Figure 6. Combined Model's Confusion Matrix

Results and Discussion

DL algorithms have good accuracy values, but other metrics such as f-measure, precision and recall did not give good results. For this reason, we did not recommend DL models. Nevertheless, the DL algorithm that worked the best was RNN. Although ML algorithms give good results in all comparison methods, we recommended it because Random Forest gave the best result. Since it is easy to create trees in the Random Forest, we concluded that the models that give the best results are in the Random Forest. Of course, we think that the reason why this is so likely to be due to the structure of our data set. As a result, the model we propose will be successful against phishing attacks.

Table 1. F1 Score, Precision, Recall values of ML and DL techniques used.

Model	F1 Score	Precision	Recall
Random Forest	0.9917	0.9895	0.9940
RNN Model	0.9732	0.9769	0.9694
Combined Model	0.9656	0.9635	0.9678

Table 2. Model features, epoch values, test sizes, and accuracy values

Model	Feature	Accuracy	Test Size	Epoch
RNN Model	Non-categorical	98.31%	30%	150
Combined Model	Non-categorical	96.26%	30%	150

Table 3. Comparison of accuracy value and cross-validation accuracy values according to Random Forest

Model	Accuracy	Cross-Validation Accuracy
Random Forest	99.09%	97.24%

Table 4. Parameters, training and test accuracy values, epoch values, and test sizes used in models.

Model	Parameters	Train Accuracy	Test Accuracy	Epoch	Test Size
RNN Model	Sigmoid/Binary Cross Entropy/Adamax	98.31%	97.20%	150	30%
Combined Model	Softplus/Binary Cross Entropy/Adamax	98.54%	96.26%	150	30%

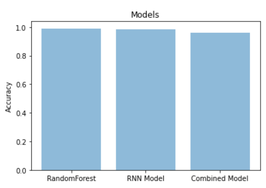

Figure 7. Models' best accuracy value

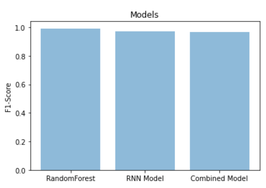

Figure 8. Models' best F1-score value

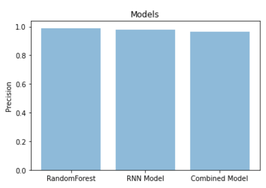

Figure 9. Models' best precision value

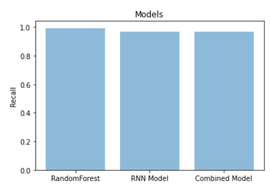

Figure 10. Models' best recall value