A Machine Learning Approach to Phishing Detection

Kendra Maggiore
Texas State University
kam355@txstate.edu

Jon Pugh
Texas State University
JPugh90@gmail.com

James Knepper
Texas State University
jbk30@txstate.edu

Abstract

With the evolution of email as one of the primary means of communication, certain problems arise with exploitation such as Phishing. This project will deal with the application of machine learning algorithms to detect phishing attacks. The algorithms we will use are logistical regression and support vector machine. These algorithms will be trained and tested on large data sets, and the results of such applications will be compared for accuracy and efficiency.

1. Introduction

Email is a widely used and effective means of electronic communication, and it is a commonly accepted method of contact between entities regardless of the specific field. Since email is such a primary correspondence in the modern world, there are those who wish to exploit email service and its users. A common exploitation technique of email services is known as Phishing. Phishing is a major threat to today's email communication. In a Phishing attempt, a malicious entity will send an email to a user that appears to be legitimate. The purpose of the email is to solicit the user to provide personal or sensitive information, often by containing a URL to an unsecure or malicious website.

2. Existing Solutions

[3] is an example of a phishing email detection machine learning algorithm that utilizes Rapid Miner. We plan to use a similar feature set but we will implement our algorithm with Python to provide greater flexibility in development. It also uses a dataset that was created in 2007. We will use a dataset available from [2] that was created in 2018. If time allows we will implement TF-IDF as discussed in [1] as an additional feature.

TFIDF	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	96.5	0.832	0.837	0.835
KNN	97.6	0.921	0.837	0.877
Logistic Regression	96.8	0.986	0.704	0.821
Naive Bayes	94.7	0.77	0.694	0.733
Random Forest	97.1	1.0	0.719	0.837
AdaBoost	97.7	0.927	0.842	0.882
SVM	98.7	0.978	0.898	0.936

Table 1. Results from [1]

Table 1 shows the findings of [1]. It can be seen that SVM yeilded the highest accuracy for their TF-IDF-based algorithm. As a result, we will focus on implementing SVM in our algorithm. Additionally we design and implement a logistic regression algorithm for comparision.

3. Preliminary Plan

- Coordinate with team members through Github: https://github.com/machine-learning-spring-2019/Phishing-Email-Detection
- Identify best feature set using [3] as a baseline.
- Convert latest .mbox from [2] to a .csv containing feature set.
- Design and implement Logistic Regression algorithm.
- Design and implement SVM algorithm.

3.1. Tenative Plans

The following will be added if time allows:

- Add TF-IDF to feature set.
- Design and implement additional algorithms studied in [1] and compare results.
- Use a URL's PhishTank classification [4] as a feature.

References

[1] S. K. Harikrishnan NB, Vinayakumar R. A machine learning approach towards phishing email detection. Technical report, Center for Computational Engineering and Networking(CEN), 2018.

[2] J. Nazario. https://monkey.org/ jose/phishing, 2018.

[3] D. Ocampo. https://github.com/diegoocampoh/machinelearningphishing, 2017.

[4] PhishTank. https://www.phishtank.com/developer_info.php, 2019.