Anna Nardelli

Dr. Yu

CS 620

13 December 2023

Final Report

The goal of the project is to be able to tell whether an email is intended to be used for cyber crimes. Malicious spam emails are often the gateway to larger cybersecurity events which can be very costly and invasive in a corporate environment. There are many different types of spam emails, which makes it more difficult to predict whether an email has honest intentions or not. Some examples of spam emails are antivirus emails, which alert the receiver that there is a virus on their computer when in reality clicking the email’s link will lead to a virus on their computer. There is also spoofing, where a threat actor fashions their email to look like it’s coming from a reputable source such as a well-known company. There are also sweepstakes and lottery emails in which the receiver is told that they have won a prize and they must click a link to claim it. Additionally, there are money scams in which a threat actor takes on the persona of a person in need and asks the receiver to send them money. To a person who is untrained in the dangers of spam emails, these messages could appear to be legitimate and they could fall into the trap and inadvertently cause serious harm. This is why it’s important that email service providers try to protect their customers from spam using detection practices.

Machine learning can be very effective in the detection of spam emails. Machine learning algorithms are very effective at detecting patterns that would not be easy for a human to notice. For my final project, I used a dataset containing emails to train an algorithm to detect whether they are spam or not. The dataset includes word frequencies for each email and tags it as either spam or not spam. I would also like to take the opportunity to test a few different algorithms for their effectiveness. I tested logistic regression, K-Nearest-Neighbor, and Perceptron to determine which algorithm is the best at predicting spam email instances.

The results of the project determine that the logistic regression is the most accurate algorithm for analyzing spam emails. The K-Nearest-Neighbors algorithm produced a lower accuracy rate than the logistic regression, and the model was very overfitted. The perceptron model also had a lower accuracy rate than the logistic regression, but it was not overfitted. See the chart on the next page for the specific data. Though the logistic regression algorithm may be slightly overfitted, the accuracy is still very high on the testing set and they are within 2 percent of each other, so it it not such a sever overfit.

The logistic regression defined in this project is an effective model for analyzing emails for spam content. The project’s goal was to create an algorithm with at least an 80 percent accuracy and this was achieved. In the future, I would like to modify this project to be able to use datasets that haven’t already been tokenized and combined into word frequencies. I would also like to be able to input my own emails into the project to test for spam.

| Algorithm | Training Accuracy | Testing Accuracy |
| --- | --- | --- |
| K-NN (4 Neighbors) | 0.930 | 0.868 |
| Logistic Regression (C=1) | 0.999 | 0.972 |
| Perceptron | 0.893 | 0.903 |

References

cisco.com/c/en/us/products/security/email-security/what-is-spam.html#~types-of-spam

<https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv/data>