Lucas Hyatt
Olivia Pannell
CIS 433

Detecting Spam with Machine Learning

The Problem:

Society is facing an onslaught of malicious emails that has reached epidemic proportions. Nearly everybody with an email account is exposed to spam. Identifying dangerous content within social platforms should be considered one of the most important security problems facing the internet. Our team is proposing the creation of a spam detection mechanism which can be used to identify malicious emails through machine learning. Combining topics of security with machine learning is a primary goal of this project, and we find both fields to be extremely fascinating and important. We are most interested in creating an accurate model, and with the availability of Spam Data Sets we feel that we have enough to not only train but validate the accuracy of our model. This challenge is what has attracted our team to this topic. Current solutions for spam detection are relatively accurate, but they suppress important data regarding the number and type of unsolicited emails being received by an account. We are proposing an algorithm which detects spam and summarizes spam activity to the user.

Approach:

Our approach to this security issue will center around the use of a machine learning model. Using the Python programming language and potentially some built-in libraries, we plan to develop, train, and test a learned model. There are data sets available online which (hopefully) provide us with plenty of data to effectively train the model. We plan to withhold at least 20% of the data we use in order to test the model (data which the model has never seen).

Plan:

Obviously, the primary goal will be to deliver a trained model capable of detecting spam. Initially, our team will research the necessary approach for creating such a model, as neither of us have experience with machine learning. After we complete this, we will begin training. The final phase to our project will be testing. We are approaching this project with Agile development in mind, emphasizing working software over documentation. This development method will allow us greater ability to respond to change. In terms of deliverables, we are planning to produce a model with a high degree of accuracy, code used for training the model, training and test data, instructions for using the model, a power point demonstrating the model's capabilities, a final report, and a poster which depicts the highlights of our project.

Week 2, 3	Week 4, 5, 6	Week 7, 8	Week 9, 10
Research how to build the model.	Find data and train the model.	Testing + Project Refining	Deliver and Present Work