Phishing for Spam

Part 1: Project Proposal

Sean O’Sullivan

2024-06-08

## 1 Description of data and data source

The data set(s) that I’d liked to examine are a handful of annotated data sets pertaining to fraudulent (spam) emails. They come from a handful of sources, te largest of the bunch is from the Conference on Email and Anti-Spam. That data set and a pair of others found on [Kaggle](https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset/data?select=CEAS_08.csv) contain text information from the subjects and bodies of the emails as well as the senders, receivers, timestamps, and whether or not URLs are contained within.

## 2 Questions/Hypotheses to be addressed

I’d like to explore the features that are prevalent in the identification of spam emails. Utilizing text analysis in addition to the other email attributes supplied my end goal would be to model a classifier that has a high rate of efficacy in identifying spam emails for automated detection as used by some popular email providers and large internal IT departments across the corporate landscape.

I expect that a number of keywords common in spam emails will be relevant, but it’s also likely that potential sender domains, whether the sender and receiver domain are the same, and what day of week or time of day the emails are sent are also highly correlated with intentionally spammy behavior.

While there may be some models better suited for this than others I intend to attempt a handful of classifiers and parameter hyper-tuning in order to compare and determine the best overall predictor of spam emails.