Email Spam Detection

ECS 171 Machine Learning
Group 5 Project Report

Group Members
Joe Zhu, Zhenshuo Xu, Omar Taha, Amar Singh, Lucas Punz

University of California, Davis
Dec 1st, 2023

Introduction and Background Information

0.1 Background Information

As university students we receive a large amount of spam emails. These emails clutter our inboxes and make it extremely difficult to access the information that we need. This constant barrage of spam not only disrupts us from our daily workflow, but also slows our organizational efforts preventing students from staying focused on academic and personal tasks.

Email spam detectors have been around for a while now, the first coming in the 1990s where two computer scientists created a database of ip addresses that they found often sent spam emails. Since then the technology has evolved and we now have Machine Learning related spam detectors that are much more efficient at blocking unwanted content.

0.2 Uses

An accurate spam detection service has many benefits for the consumer. Some of which include:

- **Time and Productivity Savings:** Spam detectors help users save time by automatically filtering out irrelevant and potentially harmful emails. Users don't have to manually sift through numerous spam messages, allowing them to focus on important communications.
- **Enhanced Security:** Spam detectors play a crucial role in maintaining cybersecurity. They can identify and block phishing emails, malicious attachments, and links that may contain malware. This helps protect users from potential security threats and keeps sensitive information safe.
- **Improved Email Organization:** By separating spam from legitimate emails, spam detectors contribute to a more organized inbox. Users can easily locate important messages without the clutter of unwanted or suspicious content.
- **Reduced Risk of Scams:** Spam detectors are effective in identifying and blocking various types of scams, including fraudulent schemes and phishing attempts. This protects users from falling victim to scams that could compromise their personal information or financial assets.

1 Literature Review

The realm of email spam detection has seen several contributions from machine learning applications. So much so that machine learning is now used by every major email company as their main way to detect spam. Gmail, Yahoo Mail, Outlook, and others all use their own forms of Machine Learning to keep their users' inboxes relatively clean.

Tejinder Singh and his colleagues investigated some of the most common algorithms used in spam detection in his paper, "Study of Machine Learning and Deep Learning Algorithms for the Detection of Email Spam based on Python Implementation." In this paper it discusses some of the different models used by these large companies to detect spam, KNN, Naives Bayes, Deep CNN, etc. The paper then goes on to discuss the experiments that they conducted before finally settling on Deep CNN as its most accurate model for email detection. The Deep CNN received an accuracy score of almost 99 percent.

Similarly there has been a large amount of research done on the use of these algorithms in detecting spam. Most of the homemade experiments provide a margin of error of around 7 percent.

2 Dataset Description and Exploratory Data Analysis of Dataset

This dataset is a collection of 5172 emails that are labeled as either spam or important. There are 3002 total columns. The first column holds the email name, the last column is a 1 for spam and a 0 for not spam. The remaining 3000 columns are the 3000 most common words in all the emails. To prepare the dataset for the algorithms we dropped some columns with low variance.

This dataset is a csv file containing 5172 rows, one row for each email in the dataset. There are a total of 3002 columns. The first column is the email name, set with numbers in order to protect the privacy of the people in this dataset. The final column is a label of 1 for spam and 0 for not spam. The remaining 3000 columns are the 3000 most common words in all of the emails, excluding the non-alphabetical characters and words. Through some data exploration we were able to determine that there are a total of 3672 non spam emails and 1500 spam emails.

We have chosen this dataset for our project because it provides a comprehensive list of important values that can be used to determine email spam. It also provides a good split of spam and non spam emails that allows us to have access to enough data for both cases. Finally, this dataset is fairly popular, as several others have used it to complete similar projects online. We have set a target for a margin of error of around 7 percent as this is congruent with some of the numbers we have seen online.

3 Proposed methodology

This problem is a classification problem. We have decided to run several different ML algorithms on the data and compare the results. The models we are going to try are Logistic regression, Naives Bayes classifier, desicion tree and random forest, Artificial Neural Network (ANN), and Support Vector Machine. We are choosing those model based on their characteristic of outputing classification results.

In order to save computational power and remove unnecessary information we decided to drop the columns with low variance as they do not help with the training. We then checked if the dataset was linear by calculating the Pearson correlation score for all of the attributes. The highest score was 0.27 which concluded that the data was nonlinear.