CS753 Project Proposal
Improvement of Bayes in the Detection of Spam Email
Abhinav Gupta, Jordan Ramsdell, Timothy Ward, and Rachel Cates

I. Motivation

Email spam is still an everyday problem for many users, and spammers continue to find ways to circumvent spam filters. Spam not only interferes with the productivity of users, but it also often includes links to harmful sites. Methods for improving the detection of spam are still actively being researched, and many of them are related to common approaches in data science and information retrieval. We wish to apply techniques we have learned in information retrieval to the task of detecting spam.

II. Task

Given an email, the task is to label the email as either as spam or non-spam ("ham"). A solution to this task must accurately identify an email as spam, while minimizing the occurrence of false-positives. As such, this is a binary classification problem.

III. Data Set

In order to answer this question, the TREC spam corpus will be used as a data set. This data set consists of approximately 10,000 emails that are labeled "spam" or "ham".

IV. Evaluation

To measure the effectiveness of spam detection methods, we will be implementing naive Bayes as a baseline. Given our corpus of emails, we will divide them into a training and test sets. We will train naive Bayes on the training set of emails and measure the accuracy of this method in predicting spam emails using the test set. We will be using the F1 evaluation metric as a measure of the effectiveness of naive Bayes. Finally, we will train and evaluate our methods in

the same way as naive Bayes, and use the F1 score of naive Bayes as a baseline for comparison. Our expectation is that we will be able to come up with a method that exceeds this baseline.

V. Methods

We will be exploring methods that can be used to detect spam emails. One group of methods we wish to explore involve indexing the training set of emails using Lucene and treating the problem of classifying emails as a passage retrieval problem. We will do so by creating features that measure the similarity of an unlabeled email to labeled emails in our indexed corpus, and label the email according to what it is closest to.

We will also explore discriminant analysis methods for determining whether or not an email is spam. For example, we can use the traditional vector space model to represent emails as documents in a vector space, and then learn a separating hyperplane between spam and ham documents. Examples of these methods are the classical linear / quadratic discriminant analysis methods, and generalized discriminant analysis using kernel-based methods.