1. Dataset

We use the 2007 TREC SPAM corpus to explore classification methods. You can find the corpus at: https://plg.uwaterloo.ca/ gvcormac/treccorpus07/

2. Baseline Method

We use naive Bayes as our baseline for classification, where the task is to label emails in the 2007 TREC SPAM corpus as either "spam" or "ham".

3. Evaluation

We parse the emails in the 2007 TREC SPAM corpus and create a randomized 50/50 train/test split. We train the classifiers on text extracted from emails in the train set, and then have them classify emails in the test set. To evaluate performance, we use the F1 measure (we consider the "spam" label as a positive, and the "ham" label as a negative) with respect to correctly called labels in the test set of emails.

Note: the content of emails has been tokenized using the EnglishAnalyzer, and stopwords were removed.

4. Results

We obtained an F1-measure of 0.94 with the naive Bayes classifier. This is surprisingly high.