Results:

The spam filter performed well on emails similar to the spam filter’s knowledge base, and showed promising potential on unfamiliar emails. The Enron corpus is a large database of thousands of emails gathered from the employees of the Enron Corporation and classified as either spam or ham. After training on several thousand emails from the Enron corpus (27705), we attempted to classify a small subset of Enron emails (separate from the training base, of course). The results were relatively successfully, with…

However, whenever the spam filter looked at emails not similar to the knowledge base, the results are less optimal. The Ling corpus is a classified set of emails from an academic department, with considerably different styles and types of emails from the Enron corpus. When attempting to evaluate the Ling corpus with the Enron knowledge base, the spam filter correctly classifies $97.96\%$ of spam emails, but only correctly classifies $79.33\%$ of ham emails.

However, the spam filter shows quick adaptation to new emails. After only reviewing a few hundred emails from the Ling corpus, the performance increased… And after reviewing a few thousand emails of the Ling corpus (2602), the ham classification of the Ling emails (disjoint from the knowledge base) increased to $100\%$, with no degradation of the spam performance.