HW6-Text Mining

Richard Barad, Dave Drennan, Jarred Randall

2023-12-14

Contents

Introduction & Methods

This report employs several text mining techniques to analyze the Enron Email Dataset - emails obtained by the Federal Energy Regulatory Commission to investigate the financial scandal of the Enron Corporation in the early 2000s. This data set is downloadable from the data science website Kaggle.

While the original data set includes over 500,000 emails, we take a sample of 100 emails and combine them into a corpus—a large and structured set of texts. The process begins by cleaning the corpus, which includes the removal of extraneous characters, numbers, and punctuation, followed by the application of natural language processing techniques such as stemming. Stemming removes common word suffixes and endings (e.g. "-ing" or "-ed"), which can help highlight more unique words in the text analysis and avoid crowding them out with different variations of the same common word or words.

We also remove stop words—commonly used words such as "the", "is", and "and", which offer little value in our analysis due to their high frequency, but low informational content. The cleaned corpus is then visualized through a word cloud, which is a graphical representation of text data where the size of each word indicates its frequency or importance in the dataset.

Subsequently, we conduct a sentiment analysis to evaluate the emotional tone behind the words used in the Enron emails. Sentiment analysis is a method used to identify affective states and subjective information by assigning sentiment scores to the terms within the text. This score can range from negative to positive and is often derived from predefined lists of words in sentiment lexicons—dictionaries where words are mapped to sentiment categories. The sentiment analysis in this study is executed using one or more such lexicons.

By combining these methodologies, the report aims to uncover not just the frequency of word usage but also the underlying sentiments, potentially offering insights into the corporate culture of Enron during its final years.

Setup

Data

Corpus

Corpus Text Cleaning

Prior to analyzing the text from the emails, we first clean the text. Our cleaning steps include removing special characters, numbers, punctuation, whitespace, and word stems. We also remove any stop words - examples of stop words include words like “a”, “the”, “is” and “are”. We remove abbreviations and e-mail metadata which is present in the email header (e.g: “cc”, “bcc”,”date”,”subject”, .etc). Lastly, we remove all words which are not present in the Scrabble dictionary to focus on common English words.

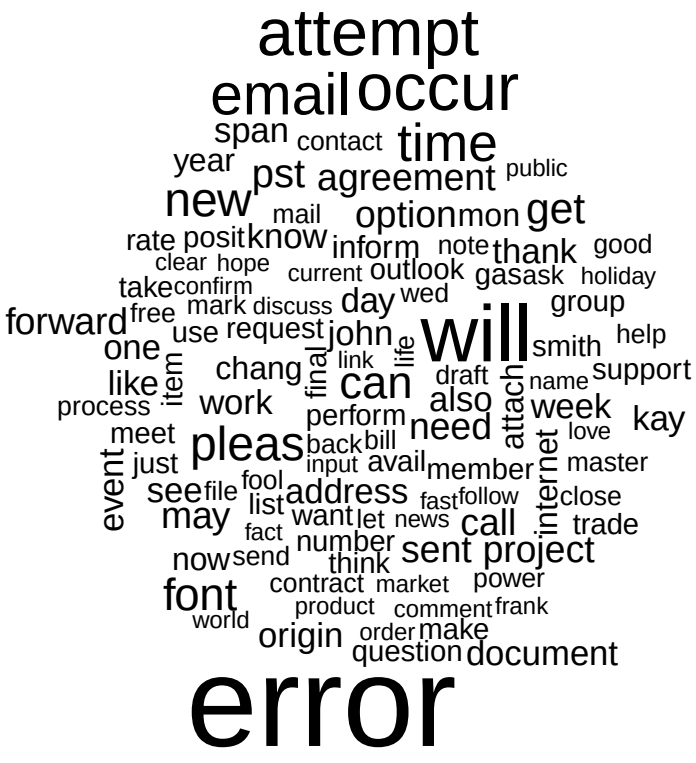
An additional step is included of creating a term matrix, which aggregates term counts across emails. This will be used to create a word cloud.

The output below shows an example of what part of the text in the first email looks like after all cleaning steps have been completed.

```
##  may      hold option agreement   kay    inbox   first option need  can  prior    new option agreeme
```

Frequency Word Cloud

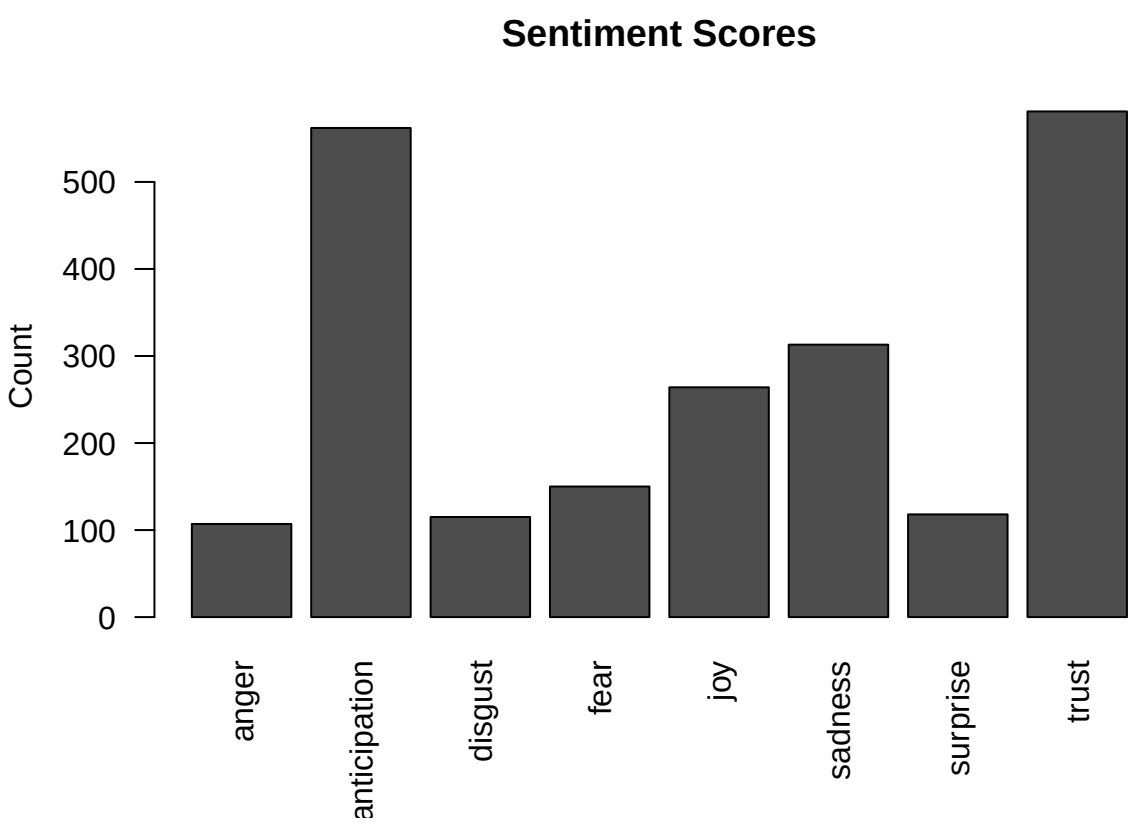
The word cloud helps us determine which words occurred most frequently across the sample of 100 emails. Only words which occur more than 15 times are included in the word cloud. Some of the most frequently used words include the words error, attempt, email, please, and time.



Sentiment Analysis

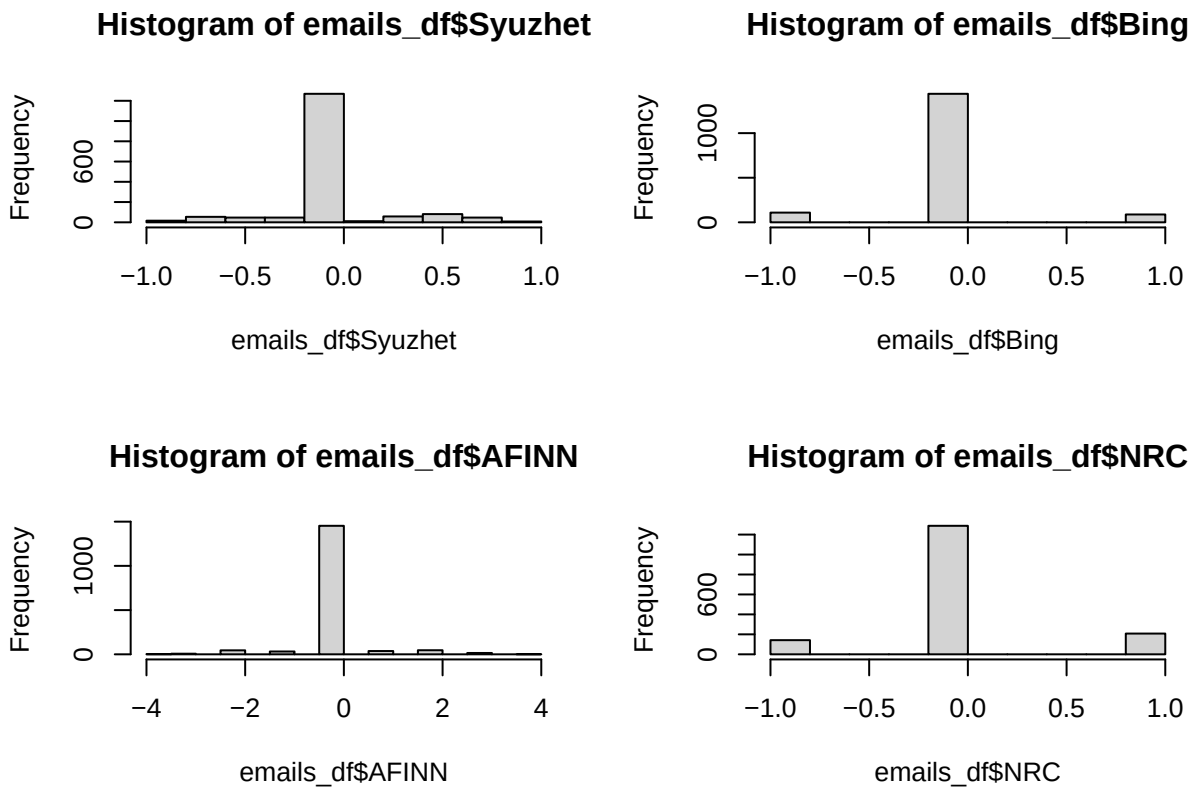
Plotting NRC Lexicon Sentiment Analysis

The graph below shows the number of words present in e-mails based on the sentiment assigned to the word using the NRC lexicon. The most common sentiments associated with words used in e-mails are anticipation and trust. The least common sentiments are anger, disgust, feat, and surprise.



Comparing Plots of Different Lexicons

The histograms below show how the words from the e-mails are classified using different lexicons. The Bing lexicons classify words as either positive (1), negative (-1) or neutral (0). The Syuzhet lexicon assigns sentiment of words on a -1 to 1 scale, with 1 indicating a very positive sentiment and -1 indicating a very negative sentiment. The AFINN lexicon assigns scores ranging from -5 (most negative) to 5 (most positive). Based on these results we can observe that the majority of words in e-mails are associated with neutral sentiments. The NRC lexicon used previously is also shown again as a comparison.



Positive, Neutral, and Negative Sentiment Counts by Lexicon

The table below shows the number of words associated with negative, positive, and neutral sentiments according to the four sentiment dictionaries. The table reinforces our previous conclusion that words with neutral sentiments are the most commonly used in our sample emails. The Syuzhet, AFINN, and NRC lexicons indicate that more positive words are present than negative words. However, the Bing lexicon indicates that more words with negative sentiments are present. We can also observe that the Syuzhet and NRC appear to classify more words to a sentiment that is not neutral when compared to the other two lexicons.

The second table below shows the same results, but presents the sentiments as the percent of the total words instead of as a raw count.

	Syuzhet	Bing	AFINN	NRC
-1	169	109	86	142
0	1264	1441	1453	1288
1	205	88	99	208

	Syuzhet	Bing	AFINN	NRC
-1	0.1031746	0.0665446	0.0525031	0.0866911
0	0.7716728	0.8797314	0.8870574	0.7863248
1	0.1251526	0.0537241	0.0604396	0.1269841

Discussion

These findings indicate that the sample of fifty random emails do not suggest a strong positive or strong negative sentiment regarding words used in Enron emails. The majority of words are neutral. Additionally, the number of positive words exceeds the number of negative words according to three out of four lexicons. In order to gain further insights, it would be useful to repeat this analysis on a larger sample of emails to see if the conclusions still hold true when using a larger sample size.

It could also be useful to repeat the sentiment analysis at the department level, and compare the sentiment of emails within the different departments of Enron to determine if some departments have a greater tendency to use words with negative sentiments. Likewise, adding a time element to compare words from different months or years could also show additional insights into the downfall of Enron.