Sentiment Analysis

Unsupervised Machine Learning on Enron Emails

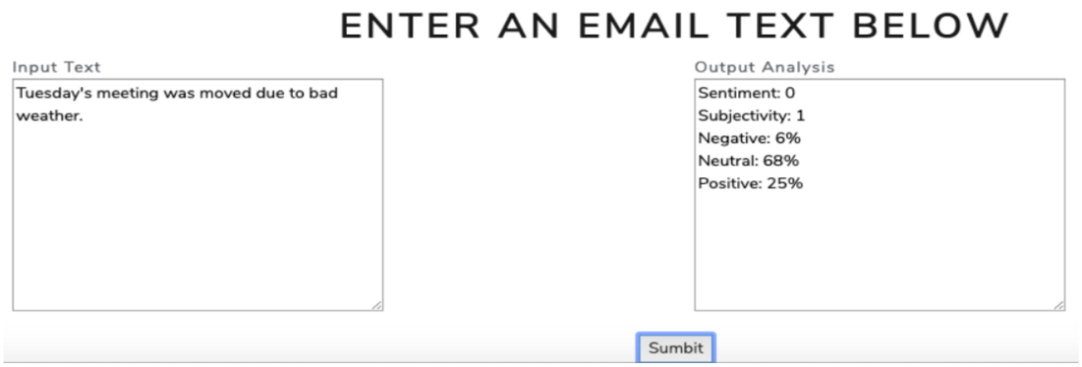
Introduction

The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse. Unfortunately, email datasets come as unstructured dataset in the form of text files or, whenever they contain any markup structure, the actual data might not be well formed. In that case, the data could be human-readable but hardly cleaned for further analysis. Therefore, the analysis should include additional mining steps and many integrity checks, in order to minimize any possible inconsistencies. Sentiment analysis is used to systematically identify, quantify, and study affective states and subjective information.

Methodology

The first step is the ETL (Extract, Transform, Load). The extraction part is about extracting data from emails in a structured form, that is we need to extract emails body and headers from the raw data. To do this we create an ‘Email Parser’ object in Pandas that implements the extraction and parsing logic. Email headers ('Message-ID', 'Date', 'From', 'To', 'Subject', 'Mime-Version', 'Content-Type', 'Content-Transfer-Encoding', 'X-From', 'X-To', 'X-cc', 'X-bcc', 'X-Folder', 'X-Origin', 'X-FileName') and body are separated, then we utilize TextBlob sentiment score for the inputted text and use the Naïve Bayes Classifier on the Email contents while splitting into train and test to predict either positive, negative or neutral (3 classes). Lexicon generates the sentiment score on the inputted text.

Results



Tuesday's meeting was moved due to bad weather.

Sentiment: 0
Subjectivity: 1
Negative: 6%
Neutral: 68%
Positive: 25%

Conclusions

Although the sentiment analysis does provide some useful insights into user input text on our front-end page, it is still unclear as to whether they can be used as investigation tools to detect fraud. However, there are some ways in which the analysis can be improved upon. For instance, analysis can be conducted by focusing on former employees such as Kenneth Lay and Jeffrey Skilling who were prosecuted for fraud-related crimes. If you would like to explore the data more, please click here.