Financial News Article Accuracy Mining

Gene Zaleski - Due 4/29/2021

Text Mining - Dr. Breitzman

Problem Statement

Trading stocks is a very volatile endeavor. People gain and lose billions of dollars every year, and there are countless industries devoted to the pursuit of making money on the stock market.

One of these Industries is the publishing of financial news articles. These articles can vary in accuracy and intent, but they should generally reflect reality for the sake of accuracy.

My goal of this project is to perform sentiment analysis on financial news articles downloaded from the internet, and compare them with the stock price around the time of publication to test both the accuracy of the sentiment classifier and the published article.

Approach (Overview)

My plan for this project was to download articles related to S&P 500 stocks and sort them by date, then plot the results versus the actual stock price that occurred to see if an article of positive sentiment reflected a climb in price, or if an article of negative sentiment reflected a drop.

I used NLTK's sanitization tools to sanitize my data.

I used NLTK's sentiment analysis pre-trained tools because I had no training set, and the Vader classifier had been extensively trained on social media datasets by the NLTK team, which I felt applied to internet articles.

Finally, I would use Yahoo Finance to compare my generated results with the actual reported stock prices.

Approach

For this project, I used Google News to mine links related to “[Stock Abbreviation] stock” for 10 pages of links of each S&P 500 stock. I used the year 2018 because I felt it was a solid year of good information I could possibly compare to 2019 if I wanted to (2019 to 2020 would be a bad comparison because covid altered many predictions). Once every link was downloaded, this netted me roughly 20,000 total articles .

Approach (continued)

Once all the articles were downloaded, I was able to sanitize my data and perform Sentiment Analysis.

To sanitize my data, I removed stop words and stemmed all downloaded articles.

Next, each article could be passed to the NLTK Vader Sentiment Intensity Analyzer. From this, I summed the total positive and negative sentiment score from each individual stock to create an overall sentiment, while also writing an individual document sentiment for a doc published on a certain date.

These results could next be compared to the reality stock prices.

Approach (continued)

With my written results, I could then download the relevant stock prices from 2018 and compare them with my calculated sentiments.

For each S&P 500 stock, I would download the stock prices from that year and plot my results against them.

Sorting my results by there Julian Date converted by calendar date, I put my results into chronological order and was able to plot an articles sentiment at the date it was published (color coded) and review the accuracy of my results.

Results

Plots for each stock are saved as PNGs in my compressed submission folder.