Problem Statement Formation

Can news headlines be used to predict stock market movements for more than 60% of the time?

Context

Scholars and researchers are looking into factors that can drive the stock values of a company. These factors can be classified as internal factors and external factors. Internal factors are derived inside the company, such as company earnings or company policies, whereas external factors are derived outside the company, such natural disasters or tensions between two countries. A news agency, under normal stances, produces news headlines that can capture both factors and therefore it is rational to examine the relationship between news headline sentiments to stock market movements.

Criteria for Success

Able to match news headlines sentiments to the index values for up to 60% of the time.

Scope of Solution Space

The analysis will be focused on the impact of Thomson Reuters new headlines to S&P 500 Index (ticker: GSPC).

Constraints

- This project only uses the news headlines from Thomson Reuters for sentiment analysis. The same modelling process and analysis techniques are theoretically applicable to news headlines from other new agencies.
- The model performance in this project is measured against a single target, a stock market index. Models used and the modelling process need to be reviewed if multiple targets, for example a group of selected stocks, are used to measure the model performance.

Stakeholders

Researchers in financial industry

Data Sources

- Historical index values – Yahoo Finance
- News headlines
    - Pham, L. (2020, July). Financial News Headlines Data, Version 1. Retrieved [07/29/2020] from https://www.kaggle.com/notlucasp/financial-news-headlines.
    - Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [07/28/2020] from https://www.kaggle.com/aaron7sun/stocknews.

Proposed Analysis

The following steps will be performed in this project:

1. Data wrangling - data cleaning and text preparation
    a. Remove stop words/numbers/punctuation
    b. Lower case
    c. Stemming

    d. Speech tagging
2. Modeling and analysis
    a. Use machine learning technique to construct a classifier to identify texts that expresses sentiment.
    b. Find an evaluation metrics to measure model performance.
    c. Map stock index values to the sentiment values and perform analysis.
3. Documentation

Deliverables

A report and a slide deck with details in

- Identified problem
- Key findings
- Modeling results and analysis
- Summary and recommendations