Stock Market Movement Prediction

Byungsu Jung
Department of Mathematics
bjung@uw.edu

Seoyoung Park
Department of ACMS(Data Sci&Stat)
syp1017@uw.edu

Abstract

The aim of our study is to roughly predict stock market price movement based on top 25 News headlines for each day. Considering this task involves NLP, we will implemented several NLP feature extraction techniques such as N-gram, TF-IDF and others. For each of feature extraction techniques, we will implement different machine learning models to evaluate which feature extraction showed highest prediction. After successfully identifying best feature extraction technique and machine learning model, we will further provide report on "Why selected feature extraction technique and machine learning model showed best performance", and we will propose future improvements,

1 Research Question

- Is it possible to roughly predict stock market price movement solely based on NLP?
- What is the best performing NLP feature extraction technique?
- After feature extraction, identify the word or words that has most impact on prediction.
- What is the best correlating machine learning model to the selected NLP feature extraction technique?
- If possible, calculate the optimized hyper-parameters that produces highest prediction.

2 Data set

We will use dataset from the **Kaggle**. This dataset consist of three different CSV files with each containing different type of data. The following is URL to our dataset.

`https://www.kaggle.com/aaron7sun/stocknews#DJIA_table.csv`

News data: Crawled historical news headlines from Reddit WorldNews Channel (`https://www.reddit.com/r/worldnews?hl`), which are ranked by Reddit users' votes. Only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)

Stock data: The Dow Jones Industrial Average, or simply the Dow, is a stock market index that indicates the value of 30 large, publicly owned companies based in the United States, and how they have traded in the stock market during various periods of time. (Range: 2008-08-08 to 2016-07-01)

- **RedditNews.csv**: The first column is the "date", and the second column is the "news headlines".

- **DJIA table.csv**: Gained from Yahoo
  - "Date": YYYY-MM-DD format
  - "Open": Opening weighted average stock value in USD
  - "High": All day high in USD
  - "Low": All day low in USD
  - "Close": Closing weighted average stock value in USD
  - "Volume": Number of trades
  - "Adj Close": Adjusted closing prices - adjusted for both dividends and splits - in USD

- **Combined News DJIA.csv**: The first column is "Date", the second is "Label", and the others are news headlines("Top 1" to "Top 25"). "Label" indicates whether the DJIA Adj Close value rose or not - Binary classification. If it rose or stayed the same, it's 1, and if it decreased, it's 0.

3 Libraries

We are planning to use following libraries for our project. The list below can be modified as we conduct the project.

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- sklearn
- scipy
- keras