

sentiment analysis impact on Stock Market prediction

Capstone Project – Literature Review



[Date]

Ryerson University

[Company address]

Table of Content

[Introduction 3](#_Toc42621243)

[Literature Review 4](#_Toc42621244)

[Dataset 6](#_Toc42621245)

[Approach 13](#_Toc42621246)

# Introduction

Predicting stock market movement -before a trading day starts- has always been a very interesting subject. In this project, I found a dataset with two channels: News Reddit data and historical financial dataset of DJIA that’s used as an indication of stock market performance and movement. DJIA stands for Dow Jones Industrial Average. It is a price-weighted stock market index that measures 30 large companies listed on NYSE (New York Stock Exchange). It covers all types of industries except Transportation and Utilities. So, if in a day the trade market started when DJIA is higher than the day before, we say “the market is up today”!

Reddit News data was captured from Reddit World News website. A place for major news from around the world, excluding US-internal news. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics like news, science, movies, video games, music, books, fitness, food, and image-sharing. Submissions with more up-votes appear towards the top of their subreddit and, if they receive enough up-votes, ultimately on the site's front page. Our dataset has only the top 25 headlines are considered for a single date.

It is important to mention that the world is changing rapidly and the traditional ways in analyzing financial data, plotting charts, building MS Excel sheets, and applying advanced formulas are not enough. There is huge amount of information scattered and dispersed around us in other forms like texts, videos, pictures, ... etc. that has impact on stock market performance as well. *This project is studying the impact of changes in world news on predicting stock market performance. More questions would rise such as: are stock stats like average prices, moving averages, median, and mode enough to predict the market? What about measuring stock market performance accuracy, stability, and efficiency?*

The project will start with detailed initial analysis visualizing stock market time-series data, generate graphs to observe how prices change over time. Perform bivariate and multivariate analysis, do more visualizations and investigate variables correlation using R language. Then explore the data and check for balanced or imbalanced dataset using Weka. Clean Reddit News text to prepare for further analysis. Sentimental Analysis is the approach to sense the market performance; whether the world news is negative or positive using R or Python libraries. This shall help predicting DJIA stock prices using a binary classification prediction model on training and testing data sets. The project will use data from 2008-08-08 to 2014-12-31 as Training Set, and Test Set is then the following two years data (from 2015-01-02 to 2016-07-01). This is roughly an (70% - 30%) split. The project will use AUC (Area Under the Curve) and Confusion Metrix for model evaluation. To measure the market efficiency and forecast stock movements, add natural learning processing like 10-k form that can be trained on historical DJIA market prices.

# Literature Review

*Thesis is about studying the impact of changes in world news on predicting stock market performance. Are stock stats such as average prices, median, and mode enough to predict the market? What about measuring stock market performance accuracy, stability, and efficiency?*

Many analyses have been conducted on DJIA performance and stock market valuation. All my reading and researches done were on the web. I started with the definition of DJIA, the companies consist of, how the stock index is priced, … etc. Also researched about Reddit World Global News, sentimental analyses, time-series analysis, and time-series modeling considering Wikipedia, Investopedia, InvestorFriend.com, and other business news on Linkedin, Yahoo Finance to libraries of universities and learning institutions. The most interesting business article I read was Dow Jones Industrial Average Valuation Analysis (Investorsfriend, 2017) that draws on Warren Buffett’s teachings. The analysis measures the fair value of DJIA and how its overpriced by 6% on April 2017considering other metrices not used by Dow Jones way of price-weighted method. So up until 2017, there was no sentiment analysis consideration in Warren’s Buffet method of teaching? It is a question I would like to gladly research on my leisure although I found it out of scope for this project.

A complete Tutorial on Time Series Modeling in R is what I started with (Srivastava, 2016). I found the step by step style of this tutorial very beneficial. How to do time series analysis explained in a framework way. Visualize the time series, Stationaries the time series, Plot ACF / PACF charts and find optimal parameters, Build the ARMA or ARIMA model, and then make Predictions. After that, I browsed a little booklet Using R for Time Series Analysis (Coghlan, 2010). I found Coghlan followed the same framework steps Srivastava used and I shall use those resources in the next step of this project (coding). Another research resource that’s more recently written on Towards Data Science medium by (Koehrsen, 2018) An Introduction to Time Series Analysis in Python. What’s interesting in this research is using the Facebook Prophet package that’s available for R and Python released on 2017 for data scientists globally to enjoy. “Prophet is designed for analyzing time series with daily observations that display patterns on different time scales”. The research was about a GM and Tesla stock prices prediction comparison. It would have been so similar to my project if there was no sentimental analysis involved in my project. Detailed analysis was made and concluded that there’s no day to day stock prediction model can be relied on, but on the long term there can. I actually believe it make sense. Another research on the same website is End-to-End Project on Time Series Analysis and Forecasting with Python (Li, 2018). The researcher used retail sales time series data to predict future sale values based on previously observed values. Overall, the forecast aligned with sales real values showing upward trend capturing seasonality toward the end of the year (i.e., the current year).

I had the pleasure of reading a master’s thesis design document on the web titled NLP and Sentiment Driven Automated Trading (Davda & Mittal, 2008). The author and the developer argued and proved that trading using technical approaches shall lead to losses and using sentimental analysis will lead to efficient market performance. They used yahoo Finance Headlines data with market prices from yahoo finance, built their own sentiment analyzer, normalized words and scored from -10 to +10 (bad to happy). Another document published for a conference for IEEE I found on Ryerson Library website, Using Sentimental Analysis in prediction of stock market investment (Khatri & Srivastava, 2016). In this paper, the sentiment analysis was performed on Twitter data, comments were categorized to four categories: happy, up, down, reject. The polarity index with market data is supplied to an artificial neural network to predict the model. The study concluded that it’s better to invest in a company with high sentimental score than a high closing price. With the same conclusion and more sophisticated analysis, a research I found when I was browsing Ryerson Library titled Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction (Seifollahi & Shajari, 2018). It is a valuable resource, emphasizes on sentiment analysis process and the importance of identifying words sense and avoid or even eliminate disambiguation. The researchers suggesting a system for FOREX to predict USD/EUR currency market prediction movement through social media news headlines word sentiment analysis. The main contribution was the introduction of novel approaches, similarity thresholds, verb normalization, and optimization measures to decrease execution times. Their system proposing outer perform (to the best of their knowledge) proposed market prediction accuracy from 83.33% to 91.67%.

All articles, thesis, document, and researches I read and studies share the same conclusion, sentimental analyses can give more insights on stock market prediction with analyzing market ratio (P/E price per earnings), earnings per share, ..etc of internal factors and external as well (like GDP).

# Dataset

The dataset is eight years’ worth of data (from 2008-06-08 to 2016-07-01) of multivariate time series stock data and text data with daily news corresponding to trading days from Kaggle. There are three files in this dataset. You can check the details of the dataset from the following link: <https://github.com/mayadakazem/Capstone-Project>

1. Stock Data File: It is historical DJIA stock prices with 7 columns and 1,989 rows downloaded directly from yahoo Finance. Each row represents a trading day at the NYSE. Stock prices of open, high, low, close values are in USD so there is **no** Currency column.
2. News Data File: It is historical Reddit News headlines ranked by reddit users' votes, and only the top 25 headlines are considered for a single date; i.e., for each date, there are 25 rows of headlines sorted by how “hot” they are as Top1. This dataset has two columns and 73,608 rows.
3. Combined Data File: This file has 27 columns of data the provider created from the above two files. For efficient use of data and fast sentiment data analysis, she included the Top 25 headlines as columns to have one row for each business day. Also, processed Stock data with News data to include one output attribute called Label that indicated when DJIA close price was up or the same, then Label is (1) and when DJIA close price was down, then Label is zero.

In all files, data is sorted descending by Date. More details regarding the dataset files are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| DJIA\_table.csv (Stock Data) | Attribute Description | Attribute Type | Comments |
| Date | Stock traded date for each business day the stock was traded | Date with format  yyyy-mm-dd | 2016-07-01  (In U.S.A. July first is a working day) |
| Open | price of the stock at market open | Number (float) | Example: 17924.24  (This is the first traded price of DJIA on 2016-07-01) |
| High | Highest price reached in the day | Number (float) | Example: 18002.38  (This is the highest value traded on 2016-07-01) |
| Low | Lowest stock price reached in the day | Number (float) | Example: 17916.91  (This is the lowest value traded on 2016-07-01) |
| Close | price of the stock at market close | Number (float) | Example: 17949.36  (This is the market closing value on 2016-07-01) |
| Volume | Number of shares traded | Number (Integer) | Example: 82160000  (Total Volume traded on NYSE on 2016-07-01) |
| Adj Close | price of DJIA at market close | Number (float) | This is considered market value of DJIA on 2016-07-01 |

|  |  |  |  |
| --- | --- | --- | --- |
| RedditNews.csv (News Data) | Attribute Description | Attribute Type | Comments |
| Date | Reddit News Date for each calendar day for the period from 2008-06-08 to 2016-07-01 | Date with format  yyyy-mm-dd | Example: 2016-07-01  Same date is repeated 25 times |
| News | Reddit Global World News Headlines | Text | Rows are sorted from highest headlines as voted by Reddit users (Top 1 to Top 25) for each date |

|  |  |  |  |
| --- | --- | --- | --- |
| CombinedNewsDJIA.csv | Attribute Description | Attribute Type | Comments |
| Date | Stock traded date for each business day the stock was traded | Date with format  yyyy-mm-dd | 2016-07-01  (In U.S.A. July first is a working day) |
| Label | Output Variable that has two values; either 0 or 1 | Numeric / Binary | "1" when DJIA Adj Close value rose or stayed as the same;  "0" when DJIA Adj Close value decreased. |
| Top 1 | Reddit Global World News headline | Text | Highest headline as voted by Reddit users  Example:  “A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours later. Trinidad Alvarez Lira had waited years for proof that she had been born in 1898.” |
| Top 2 | Reddit Global World News headline | Text | Second highest headline as voted by Reddit users |
| Top 3 | Reddit Global World News headline | Text | Third highest headline as voted by Reddit users |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| Top 25 | Reddit Global World News headline | Text | 25th highest headline as voted by Reddit users |

Here’s some Descriptive Analysis to the dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | Data Type | Minimum | Maximum | Mean | Std. Deviation |
| Date | Date | 2008-06-08 | 2016-07-01 | 2012-06-19 | Eight years data - dates are from 2008-06-08 to 2016-07-01 |
| Open | Quantitative / Numeric | 6,547.01 | 18,315.06 | 13,459.12 | 3,143.28 |
| High | Quantitative / Numeric | 6,709.61 | 18,351.36 | 13,541.30 | 3,136.27 |
| Low | Quantitative / Numeric | 6,469.95 | 18,272.56 | 13,372.93 | 3,150.42 |
| Close | Quantitative / Numeric | 6,547.05 | 18,312.39 | 13,463.03 | 3,144.01 |
| Volume | Quantitative / Numeric | 8,410,000.00 | 674,920,000.00 | 162,810,970.34 | 93,923,426.68 |
| Adj Close | Quantitative / Numeric | 6,547.05 | 18,312.39 | 13,463.03 | 3,144.01 |
| News, Top1, Top2, … , Top 25 | Qualitative / Text |  |  |  |  |
| Label | Qualitative / Nominal (0,1) |  |  |  |  |

# Approach

The flow chart below explains on high level the project approach to predict DJIA market movement using eight years (from 08-08-2008 to 07-01-2016) combined data of DJIA stock prices and Reddit headlines news data.

START

Stop

Data Model and Prediction

Sentiment and Weight Determination

Combined Data

Stock Data

Reddit News Data

Text Data Disambiguation / Cleaning

Stop Words Removal & Tokenization

Descriptive Data Analysis

Data Exploration