**Machine Learning**

**Predicting stock price change from Financial News**

**Group 21**

Mandy Gu

Willa Yu

Mayank Mani

Tejasvini Karunakarbabu

Contents

[Executive summary: 3](#_Toc35207757)

[Introduction 4](#_Toc35207758)

[a) Defining the sentiments: 6](#_Toc35207759)

[Long Short-Term Memory model: 7](#_Toc35207760)

[Doc 2 Vec Model: 7](#_Toc35207761)

[b) Aggregate sentiments to a daily level 8](#_Toc35207762)

[c) Compare the sentiment with the following day’s stock price 8](#_Toc35207763)

[Business Value provided 9](#_Toc35207764)

[Conclusion and Recommendation 10](#_Toc35207765)

[Reference: 10](#_Toc35207766)

# Executive summary:

News analytics is a growing field with several applications apart from stock price prediction such as automated detection of insider trading, circuit breakers to halt trading algorithms when news arrives, and stock screening to find interesting assets.

In the news data web scraping we obtain information of headlines, the origin of the news, published time, article and the companies mentioned in the article. These are vital inputs to execute Natural Language Processing (NLP) algorithms. The news data is the cause and stock prices are the effect in our modeling, therefore we gather information such as stock price points (opening, closing, high and low) of the corresponding company.

We sought to co-relate the news headlines with the growth rate of stock prices and use them to predict the change in stock prices for the next day.

Our process was divided into three stages. We manually mapped the news as positive, negative and neutral using domain knowledge to train the model. Then we use text analytics modeling techniques such as the bag of words and classification model to attach final sentiment. Post this, we established the relationship between news and sentiment.

We found that the aggregated sentiment of stock for a day is positively correlated with the opening price of the stock for the next day. We see that if the aggregate sentiment is mostly positive we can expect the stock to grow the next day and vice versa. We also find the effect of aggregate sentiment is highest on the opening stock. We conclude that the opening stock price increases by 0.103% for every 0.1 increase in positive aggregated sentiment. This calls for the early call of action the next day.

Our unique value proposition is to help our clients manage the stock market volatility by dynamic portfolio management on a daily time period so as to maximize their return on investment.

# Introduction

As a fin-tech company, we would like to provide our clients with insights on the impact of news on stocks. By focusing on ten key companies such as Facebook, Apple, and Tesla, we sought to predict if the stock price will increase or decrease. We can hedge risks for our clients by advising them to take appropriate measures based on our recommendations. From historic data collected through financial news and daily stock prices, we analyzed the relationship between sentiments of news, from headlines and articles, on stock prices.

Business strategy and Market

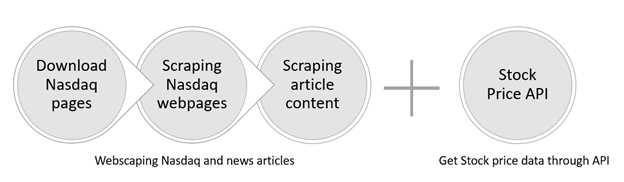
The entire world’s stock exchanges have a capitalization of $85 trillion USD, trending up from $25 trillion in 2009 a 320% increase.[1] Stock prices, trends, and growth are of keen interest for investors. The volatile nature of the stocks is a major hindrance to prediction. Estimated 93% of Mutual funds underperform in the market [1].

One major factor influencing stock is news and news analytics has gained traction since the advent of natural language processing. There are multiple analytical companies in this market such as RavenPack, Lexalytics, and InfoTrie who solely look at the text analytics of financial news to provide predictive analytics.

Our goal is to convert the vast amount of unstructured financial news into machine-readable sentiments and predict the impact on the stock market. Specifically, we are focusing on establishing the known relationship between news and stock prices by reading the headlines and article content of the news and looking at the impact of the sentiment on the stock prices.

Data characteristics

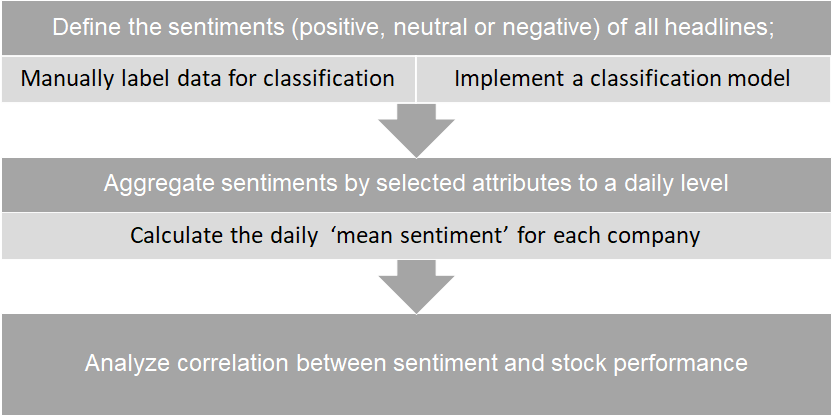
We used a two-pronged approach in data collection. Firstly, we stored the webpages from historical archives of Nasdaq for extracting the news and article details and Secondly, we used an application programming interface (Alpha Vintage API) to obtain stock price details for the companies.



*Data collection sources and methodology*

Analysis and Modelling

With five hundred headlines and stock prices for each company in our company list, our methodology is to summarized below.



## a) Defining the sentiments:

Two methods can be applied to analyze the sentiment of a headline:

1. Set a specific rule of detecting if the headline conveys positive information based on sentiment dictionary (for instance, if a negative word such as *loss* appeared before ‘*increased’*, it will be -1 \* 1 = -1. The sentiment for the headline will be determined by the final score after accounting for all the positive or negative words).
2. Build a supervised sentiment analysis model by manually labeling a subset of headlines and implement a classification model to identify the sentiment automatically.

We finally chose the second method although it took us much more effort to label data. The reason is that without a well-structured sentiment dictionary specific to the finance domain (pertaining to financial terms), even if the model has a high accuracy rate upon training data, it would still underperform in out of sample accuracy rate since it fails to learn the information outside the training data. As a result, we stuck to the second method of building a supervised model, where we labeled ‘2852 headlines’ in total.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Labeling** | **Positive** | **Neutral** | **Negative** | **Total** |
| Quantity | 721 | 1759 | 372 | 2852 |

*Table 1: The number of three kinds of headlines we labelled*

After we labeled the data, we preprocessed the headlines by removing punctuations and stop words, stemming, transferring all words to lowercase and oversampling. Several methods of embedding and building a model were carried out to do prediction and the model having the most predictive power was chosen, which was embedding the headlines by doc2vec (a model used to produce word embeddings) and implementing Multinomial Logistic Regression to do classification. The accuracy rate reached 82.81%.

### Long Short-Term Memory model:

The first method we applied was to build a word2vec model based on both headlines and their corresponding articles and input headlines to the Long Short-Term Memory (LSTM, a special case of Recurrent Neural Network that can memorize long-term dependencies within a document) by embedding it with weights from the word2vec model. The reason why we tried to incorporate the LSTM model was that it could take the position of words into consideration and memorized them. However, the accuracy rate was only around 63.25%.

Drawbacks that led to the failure of the model :

* The limited training data which prevented the LSTM model from learning
* Since we did not aggregate weights for all words in headlines, the dimension of the headline was too high to deal with.

To remedy the model, we needed to scrape at least hundreds of Megabytes of data from the website or to incorporate pre-trained models into our case, which is not specific to the Finance domain. Both solutions were not ideal enough, thus we turned to another method.

### Doc 2 Vec Model:

Another model we utilized was that instead of incorporating word2vec model, build a doc2vec model that can directly turn a headline to one set of weights and apply Multinomial Logistic Regression (MNLR) to predict the final sentiment. For doc2vec model, there are also two methods to carry out the model: Distributed Bag of words (DBOW) and Distributed Memory (DM), both of which are document embedding approaches. The accuracy rates of these two methods were 75.48% and 75.41%, respectively.

Although the model was much better than the previous one, it was still not ideal for us. Thus, we generated a new model that combined these two methods together and got an accuracy rate of 81.9%.   
Apart from MNLR, we also explored other methods such as ‘XGBoost’ which is an optimized distributed gradient boosting model, where we got a 76.96% accuracy rate. As a result, embedding headlines with doc2vec model and incorporating weights into an MNLR model won over the other methods in this case.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models ->** | **Word2vec & LSTM** | **Doc2vec & xgboost** | **Doc2vec DBOW & MNLR** | **Doc2vec DM & MNLR** | **Doc2vec combined & MNLR** |
| Accuracy rate | 63.3% | 77.49% | 76.23% | 76.25% | 82.81% |

*Table 2: A comparison of accuracy rates carried out by different methods*

## b) Aggregate sentiments to a daily level

After applying these models, we chose Doc2vec combined & MNLR with the highest accuracy rate to generate sentiments for all the 5000 news and then accounted for exceptions:

* Since companies usually have more than one headline every day, we need to first aggregate the news-level sentiments to daily sentiment for each company. The metric we use for everyday sentiment aggregation is ***‘mean sentiment’***.
* Additionally, the news released on weekends will only have an effect on the stock price of next Monday, so we combined these headlines together with the news published on the previous Friday.

## c) Compare the sentiment with the following day’s stock price

In order to analyze the relationship between the sentiment and the following day’s stock prices, we built a simple linear regression to investigate the correlation between the growth rate of all the companies’ daily prices and the average sentiment as the independent variable. We investigated into different times of prices and found out that daily open price had the most significant correlations with aggregated sentiments.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Daily Price Elements** | **Open price** | **Close price** | **High price** | **Low price** |
| Coef | 0.0103 | 0.0061 | 0.0070 | 0.0094 |
| P-value | 0.001 | 0.055 | 0.012 | 0.002 |

*Table: Correlation of aggregated sentiments with daily stock price elements of all the stocks*

# Business Value provided

Due to the complexities in quantifying and analyzing the news articles, people often neglect its effect on stock prices, thereby attributing the change to either the microeconomic and macroeconomic factors. For short term investors, relying on such ‘traditional’ factors is not an option as the effects of such changes take place over a longer time period. A short term investor will likely refer to news articles for suggestions to invest in certain stocks. But they suffer from three major drawbacks:

1. The news articles can itself be biased based on the author’s views
2. Due to the financial terminologies, sometimes the articles can be difficult to make sense
3. Such articles mostly focus on big companies and may not be representative

Hence, we tried to present our findings to address the above scenarios:

1. By aggregating thousands of news headlines taken over several weeks, we took care of the biases and they would average out
2. The user can look only at the predicted stock price changes, while the model takes care of all the financial terminologies and complexities
3. Our model can easily be expanded to cover as many companies we want and to include as many data points as required and can predict the stock price for all the companies

# Conclusion and Recommendation

To summarize, we were able to state confidently that the stock prices are affected by the news. The aggregated summary for a day is most positively correlated with the opening price of a particular stock for the next day. We conclude that the opening stock price increases by 0.103% for every 0.1 increase in positive aggregated sentiment.

We recommend that:

1. We see a direct correlation between sentiments and growth rate movement. Therefore, we can dynamically update our client’s investment portfolio based on the current mean sentiment for the portfolio
2. Moreover, the open price of stocks was affected the most by aggregated sentiments and hence, exchange stocks in early day time

In conclusion, through our final model we can use the sentiments to predict the performance of stocks in real-time and make changes to portfolios of customers to gain maximum advantage and growth.

# Reference:

1) <https://www.liberatedstocktrader.com/stock-market-statistics/>