An Approach To Textual Analysis For Stocks

Interactive Qualifying Project

Theodoros Konstantopoulos

Professor Dimitrios Koutmos

Worcester Polytechnic Institute

Table of Contents

List of Figures

1. Introduction

Numerous scientific attempts have been made to predict the stock market but there has not been one method that accurately predicts stock price movement. Most literature on stock market prediction relies on associating certain keywords with price movement but this approach has shown weak ability to forecast the direction of share prices.

News articles can enrich the knowledge of investors and affect their trading activities while public sentiments cause emotional fluctuations and intervene in their decision making. It is possible that investors may value information in different ways but this paper assumes that the information is valued based on the data produced by the sentiment analysis that is performed. This might be inaccurate in a lot of cases. However, since there is no proven way to predict the market, those inaccuracies could turn out to be suggestions that are profitable.

To test the hypothesis, a group of websites was chosen based on the Alexa site list for the investing sub-category under the business category. A custom crawler was written for every site that is listed on the Alexa website. Weights are generated based on the information for total sites linking a website. The weights are needed in order to compute a weighted average for the sentiment analysis results. The information to compute the weights for every website is retrieved from the same source.

The program updates this list every time it refreshes by pulling the table from the same source again to reflect any potential changes in website popularity. Then, a sentiment analysis is performed by an analyzer that is trained using social media and news data.

This paper is arranged as follows. Section 2 provides a summary of literature regarding stock market prediction. Section 3 discusses the background of the project; the terms and tools used throughout the paper. Section 4 describes the system design of the project. Section 5 discusses the data collected and the correlation between the sentiment and stock market performance. Section 6 concludes the paper and discusses the implications of the project.

2. Literature Review

The Efficient Market Hypothesis (EMH) theory is usually referenced as a basis of all the approaches that attempt to make stock market forecasts. According to the EMH, there is no advantage in doing forecasts because the market reflects the latest news, making it impossible to outperform it.

Two philosophies have been developed based on the assumptions of the EMH. Those are the fundamentalists and the technicians. According to the fundamentalists, the price of a security can be determined from the overall economy, the particular industry's sector, and from the company itself. On the other hand, technicians examine historical and time-series data. They believe that opportunities can be found by finding patterns in historical volume of an asset. However, technical analysis is thought to be an art form rather than a science.

Regarding the deep learning approaches to textual analysis, a paper claims that CNN (Convolutional Neural Networks) could be better than RNN (Recurrent Neural Networks) on catching semantics from texts while RNN is better on getting the context information. Their proposed method shows some improvement compared to previous research.

Concerning the textual analysis, there is an excessive amount of methods to analyze finance news articles. A common method is to use every single word from an article and assign weights to each one. This method assigns importance to certain words and undermines others in order to determine the overall meaning of an article. Another method is the noun phrasing approach which identifies certain phrases using a lexicon.

The AZFinText system was developed based on the noun phrasing strategy. The system aims to predict discrete number stock prices 20 minutes after an article release. With the use of a Support Vector Machine (SVM), the system shows better results than linear regression.

The AZFinText system was also tested against quantitative funds and had a 2% higher return than the best performing quant fund. The approach used textual representation and statistical machine learning methods to analyze financial news articles of similar industries.

Another paper shows that the models that work with sentiment analysis outperforms the bag-of-words model. In the paper, three approaches were used: the sentiment analysis approach, the bag-of-words approach and the sentiment priority approach. In order to produce the results, the daily stock price return prediction accuracy was compared with the Hong Kong Stock Exchange market data for the period of five years.

The importance of textual analysis is significant. Examining the 8-K documents alone, it was found that text boosts prediction accuracy over 10%. This impact is most important in the short term but persists for up to five days. Form 8-K is used to notify investors in United States public companies of important events. The information included in this form is important for big trading institutions and not for individual investors.

3. Background

Collection of data can be done in various ways. It was thought that the most accurate way to retrieve data would be to collect headlines and finance news articles. A crawler is a type of program that has the capability of indexing information found online by extracting text or other forms of media from a web page so that users can search more efficiently. Web pages are built using HTML or XHTML so in order to parse and extract information from them, scrapers need to be developed to appropriately handle these mark-up languages.

Another important issue discussed in this paper is the market sentiment. We refer to market sentiment as the anticipated price movement in a market. Upward price movement is said to be bullish while downward price movement is said to be bearish. In the last decade, market sentiment is also measured by utilizing news analytics which includes sentiment analysis on textual stories.

4. System Design

The textual analysis was performed using a lexicon and rule-based sentiment analysis tool. The sentiment analysis performed takes into account typical negations (e.g., “not good”), use of contractions as negations (e.g., “wasn’t very good”) and conventional use of punctuation that signals increased sentiment intensity (e.g., “Good!!!”).

The news websites are visited every hour since visiting them more frequently could result in a ban of the scraper’s IP address. The scraper extracts the headlines and the links of the articles, performs a sentiment analysis based on a lexicon and stores them in a database.

The system uses object-oriented programming (OOP), a type of software design in which the data type of a data structure is defined. Also, the types of operations (methods or functions) that can be applied to the data structure are defined.

An abstract class was designed as a basis for every custom website scraper. An abstraction is one of the three main principles of object-oriented programming (OOP). The process of abstraction is used to hide all the relevant object data, reduce complexity, and increase efficiency.

A custom website scraper was written for every website that data was collected from since the design of each website in terms of HTML code is different which means that one implementation could not have worked. The scraper is written using Beautiful Soup, a python package for parsing HTML and XML documents, matplotlib, vaderSentiment and several other libraries for storing the data retrieved.

The python library matplotlib was used in order to create the plots of the sentiment. Matplotlib is a plotting library for the Python programming language that provides an object-oriented API for embedding plots.

```
# An abstract class representing a finance website scraper
class FinWebScraper(ABC):
    def __init__(self, url):
        self.response = requests.get(url)
        self.website_average = WebsiteAverage(url)

    def get_soup_object(self):
        ## Debug
        ## print("HTML: %s" % (self.response.text))
        soup = BeautifulSoup(self.response.text, "html.parser")
        return soup
```

```
def validate_url(self, url, valid_host):
    o = urlparse(url)
    return valid_host == o.netloc

def remove_query_string(self, url):
    return furl.furl(url).remove(args=True, fragment=True).url

def classify_headline(self, headline):
    # Set self.sentiment
    txt_classifier = Classifier(headline)
    sentiment = txt_classifier.sentiment()
    print(sentiment)
    self.sentiment = sentiment
    self.update_avgs()

def update_avgs(self):
    # We want to avoid neutral articles
    if self.sentiment['neu'] < 1.0:
        Average.add_compound(self.sentiment['compound'])
        Average.add_neg(self.sentiment['neg'])
        Average.add_neu(self.sentiment['neu'])
        Average.add_pos(self.sentiment['pos'])
        self.website_average.add(self.sentiment['compound'])

def save(self):
    db = Database()
    db.insert_article({
        'link': self.article_link,
        'title': self.article_title,
        'neg': self.sentiment['neg'],
        'neu': self.sentiment['neu'],
        'pos': self.sentiment['pos'],
        'compound': self.sentiment['compound']
    })
```

The abstract class representing a finance news site scraper

```
import matplotlib.pyplot as plt

x = []
for date in dates:
    x.append(date)
y = []
for compound in weighted_compounds:
    y.append(compound)
fig, ax = plt.subplots()
ax.plot_date(x, y, linestyle='-')
fig.autofmt_xdate()
plt.show()
```

Creating a plot of the sentiment with matplotlib

5. Data

After performing the sentiment analysis, the compound sentiment was plotted along with the previous compound sentiments on a scale from -100 to 100. Usually, the sentiment is not bigger than 10 or less than -10 because there is a lot of neutral news taken into account.

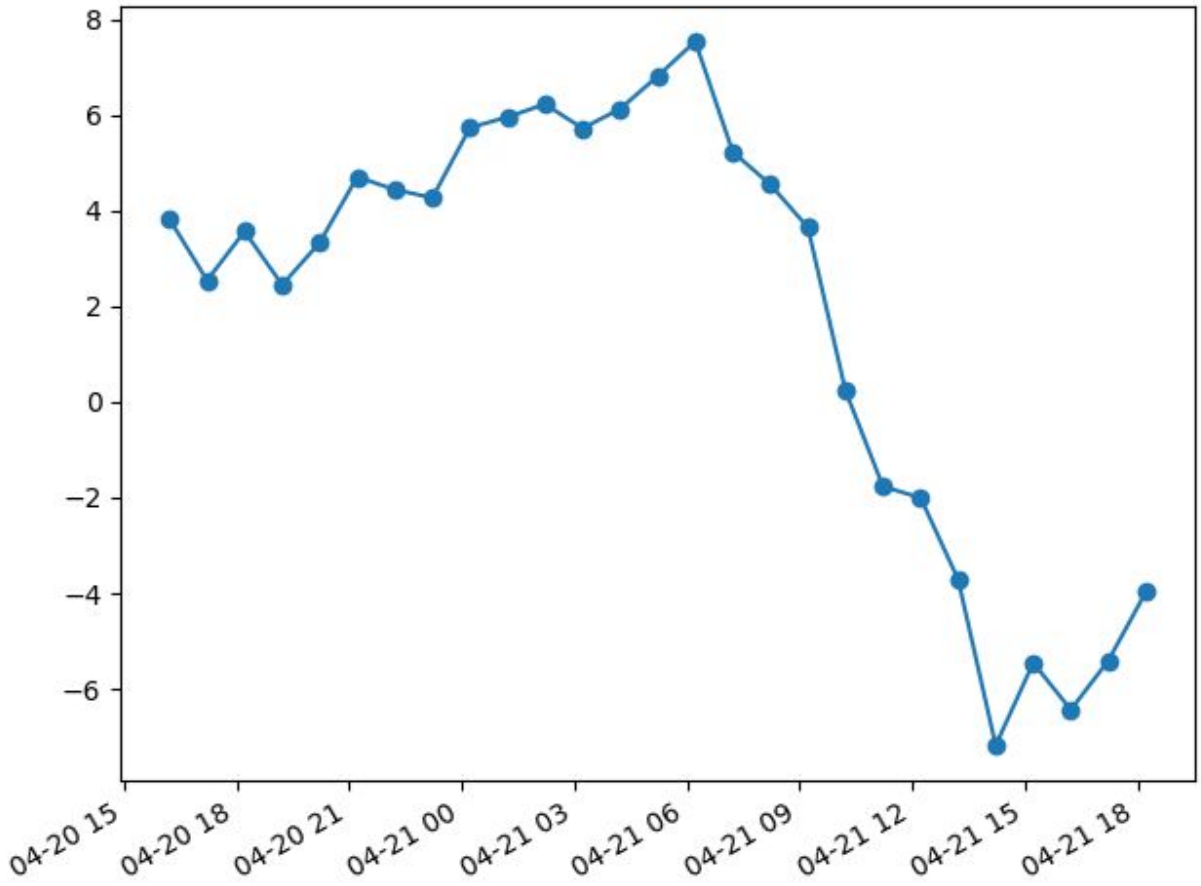

Figure 1: Sentiment plot from the 20th of April 2020 at 3:00 pm EST to the 21st of April 2020 at 6:00 pm EST

Figure 1 shows the sentiment plotted from the 20th of April 2020 at 3:00 pm EST to the 21st of April 2020 at 6:00 pm EST. It is observed that the sentiment turns negative around 9 pm on April 21st.

Figure 2: Dow Jones Industrial Average plot from the 20th of April 2020 at 12:00 am EST to the 21st of April 2020 at 12:00 am EST

Figure 2 shows the Dow Jones Industrial Average for April 20th 2020 and April 21st 2020. There are some similarities between the sentiment and this figure since both are declining for about the same period in time. Specifically, the Dow Jones Industrial Average opened at 24,095.10 and closed at 23,650.44 on that day.

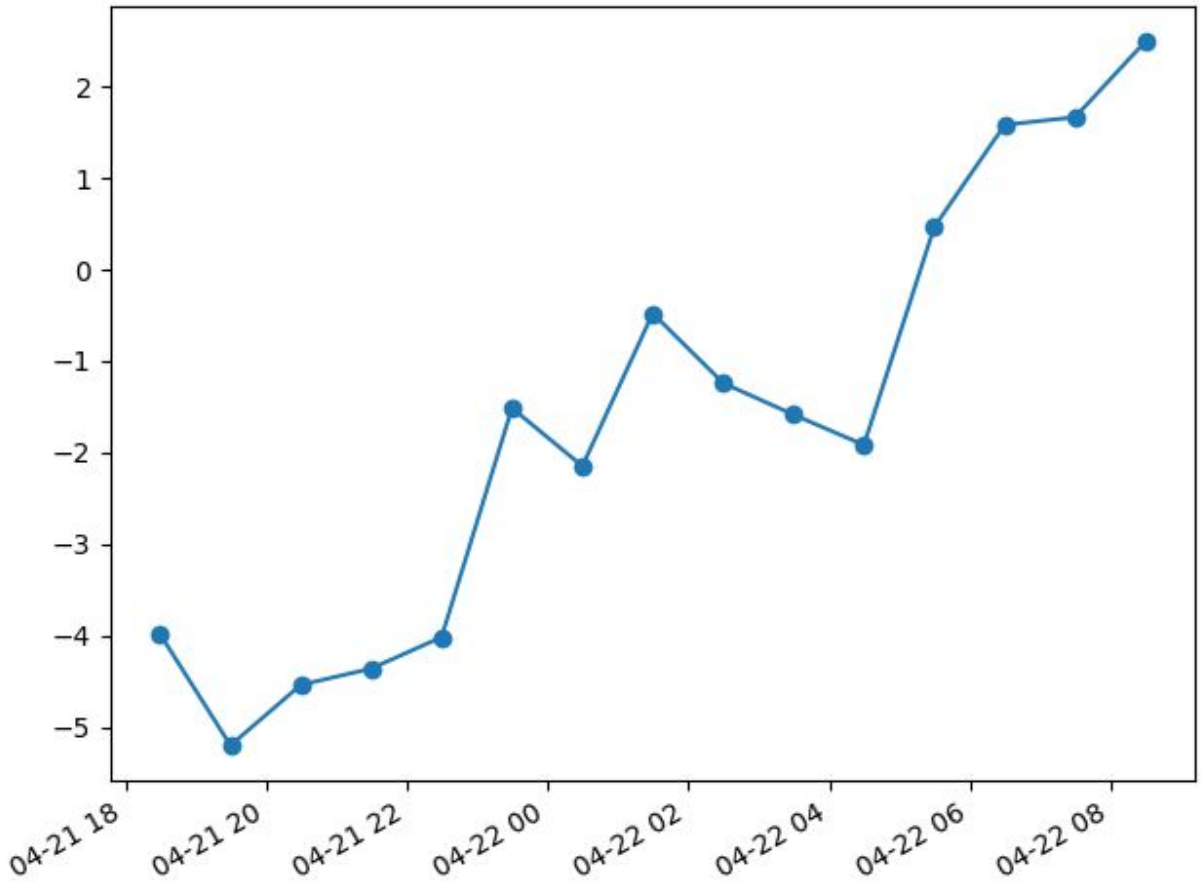

Figure 3: Sentiment plot from the 21st of April 2020 at 6:00 pm EST to the 24th of April 2020 at 8:00 pm EST

Figure 3 shows the sentiment plotted from the 21st of April 2020 at 6:00 pm EST to the 24th of April 2020 at 8:00 pm EST. It is observed that the sentiment is increasing throughout the day.

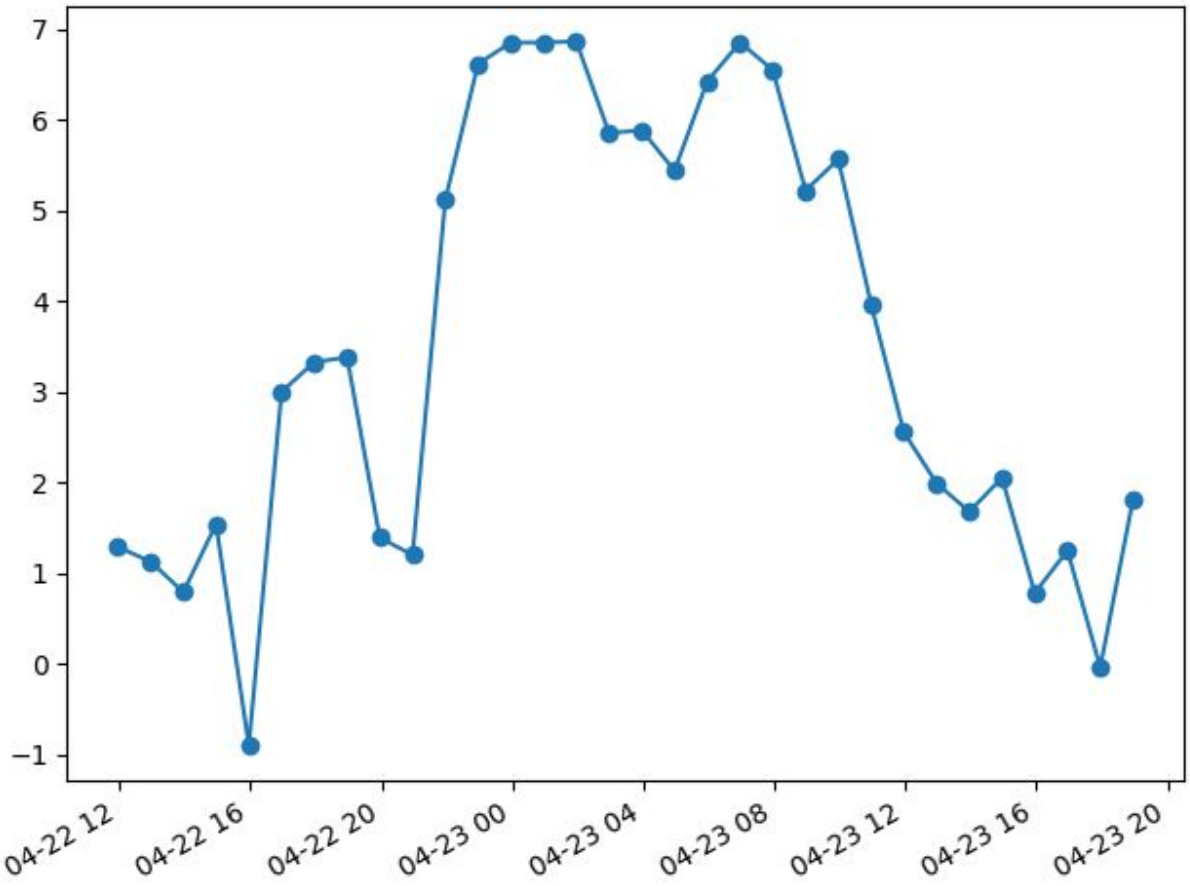

Figure 4: Sentiment plot from the 22nd of April 2020 at 12:00 pm EST to the 21st of April 2020 at 8:00 pm EST

Figure 4 shows the sentiment plotted from the 22nd of April 2020 at 12:00 pm EST to the 23rd of April 2020 at 8:00 pm EST. It is observed that the sentiment increases initially but suddenly drops. It is also observed that the volatility index (VIX) increased at approximately 12:40 pm on that day. However, the sentiment started dropping before the increase of the volatility which indicates that the news started getting worse before the market reacted.

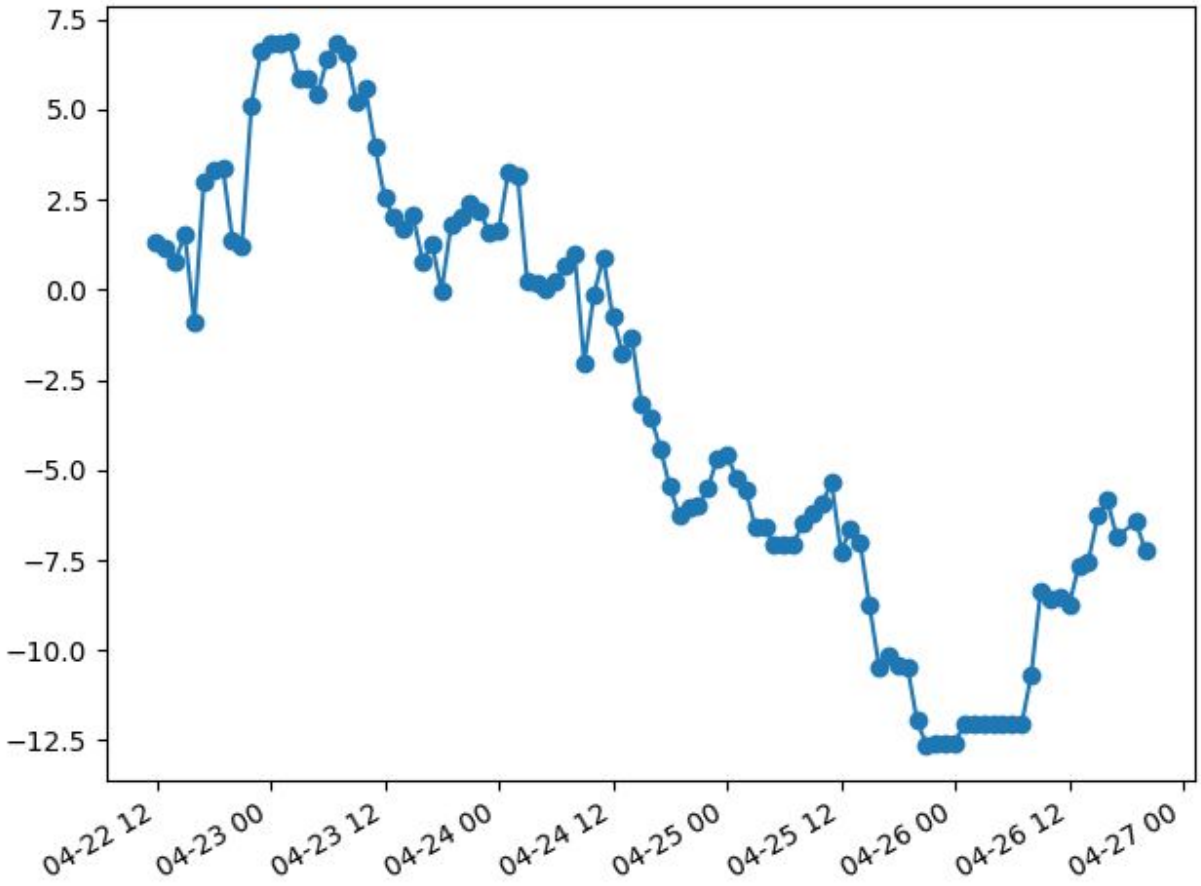

Figure 5: Sentiment plot from the 22nd of April 2020 at 12:00 pm EST to the 27th of April 2020 at 12:00 am EST

Figure 5 shows the sentiment plotted from the 22nd of April 2020 at 12:00 pm EST to the 27th of April 2020 at 12:00 am EST. It is observed that during those five days, the news articles were getting worse. Also, it can be noticed that on April 26th, the same sentiment was getting produced on an hourly basis since most websites are not getting updated on Sundays.

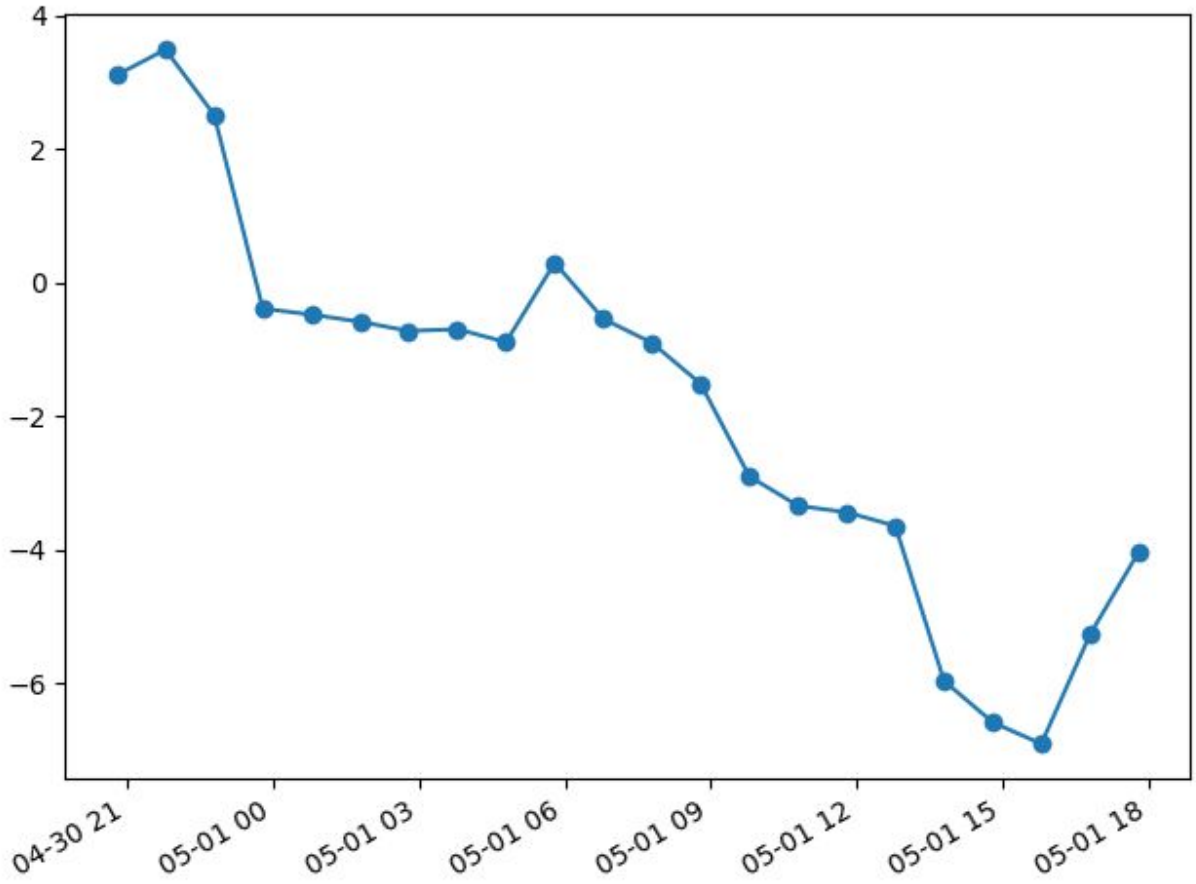

Figure 6: Sentiment plot from the 30th of April 2020 at 9:00 pm EST to the 1st of May 2020 at 6:00 pm EST

Figure 6 shows the sentiment plotted from the 30th of April 2020 at 9:00 pm EST to the 1st of May 2020 at 6:00 pm EST. It is observed that the sentiment is decreasing throughout the day. The Dow Jones Industrial Average opened at 24,585.57 and closed at 24,345.72 on that day.

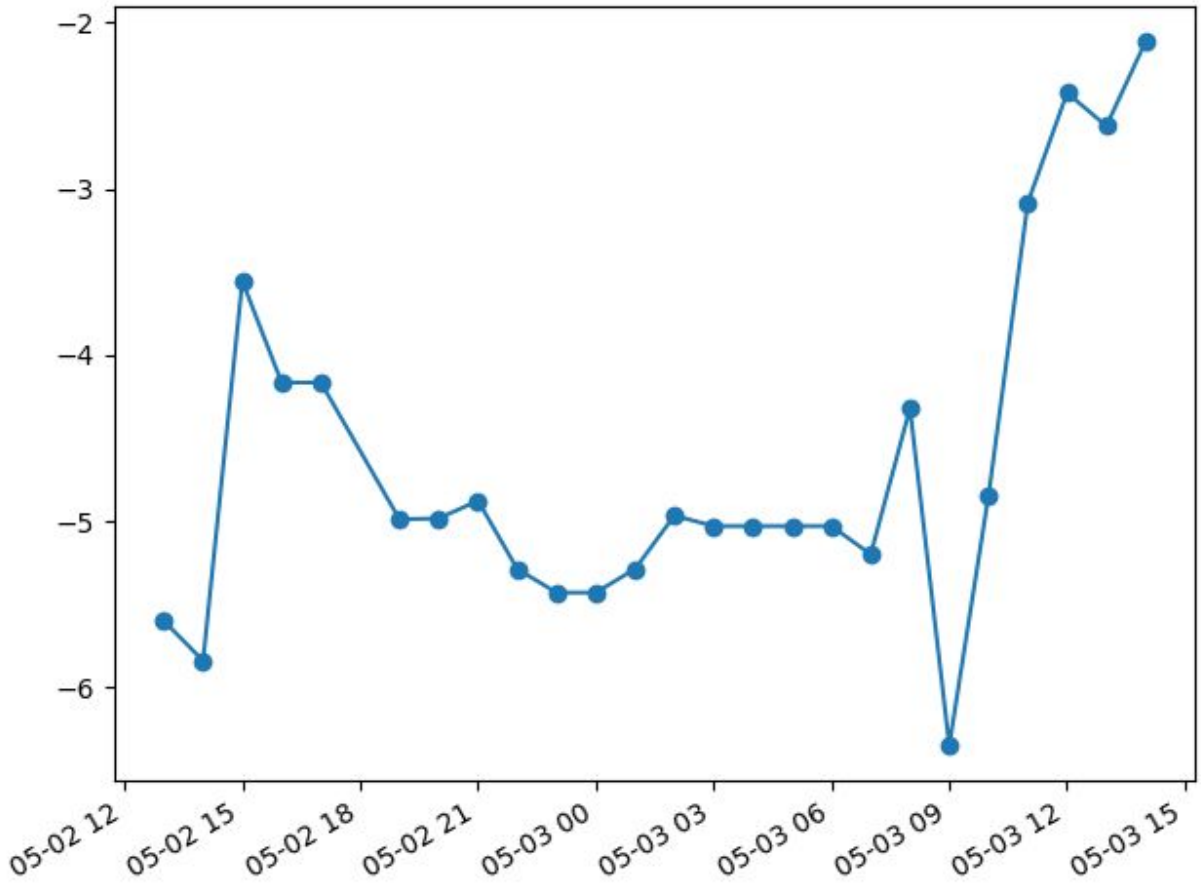

Figure 7: Sentiment plot from the 2nd of May 2020 at 12:00 pm EST to the 3rd of May 2020 at 3:00 pm EST

Figure 7 shows the sentiment plotted from the 2nd of May 2020 at 12:00 pm EST to the 3rd of May 2020 at 3:00 pm EST. It is the day after the previous figure. It is observed that the sentiment starts increasing after 9:00 pm EST. The Dow Jones Industrial Average opened at 23,581.55, or 142.14 points lower than the previous Friday, and closed at 23,749.76 on the next day.

6. Conclusion

It can be observed that the sentiment plots usually follow the major indexes. More importantly, there are some cases where the sentiment is ahead of the market. Also, the majority of headlines are neutral and do not have an impact on the sentiment. In the future, improvements can be made to the strategy utilized to calculate the sentiment and a virtual trading environment can be created to compare the performance of the sentiment to the Dow Jones Industrial Average or any other major index.

References

[1] Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. On the Importance of Text Analysis for Stock Price Prediction.

[2] Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69(1), 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

[3] Schumaker, R., & Chen, H. (2009). A quantitative stock prediction system based on financial news. Information Processing and Management, 45(5), 571–583. https://doi.org/10.1016/j.ipm.2009.05.001

[4] Schumaker, R., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), 27(2), 1–19. https://doi.org/10.1145/1462198.1462204

[5] Schwartz, R. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work: Discussion. The Journal of Finance, 25(2), 421–423. https://doi.org/10.2307/2325488

[6] M. R. Vargas, B. S. L. P. de Lima and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," 2017 IEEE International Conference on Computational