# Predicting the Direction and Magnitude of Stock Market Fluctuations from Sentiment Scores of Financial News Articles

**Bernard Wittmaack**

## Introduction and Goals

Attempting to forecast financial market fluctuations from one day or one week to the next is the basis of short-term trading. However, even with the application of sophisticated models, the practice is inherently risky due to the volubility of the market, let alone individual stocks, that occurs on shorter time scales. Therefore, short-term trading is ill-advised for the average investor.

However, what if there were an easily accessible and interpretable source of reliable information available to the amateur investor looking to predict how the market will change over the course of days to weeks? In fact, one of these prophecies may even be open right now in a buried browser tab. Every day, would-be speculators sedulously study the Wall Street Journal or the Motely Fool in hopes of catching a glimpse of a future where they are casually glancing at brochures on luxury cruises or sumptuous vacation homes instead.

Financial news sources command an army of experts who review and analyze the causes behind market trends and hazard predictions about their future states. With so many experts and news sources, it can be difficult to keep abreast of all the information published daily. Unlike following the price of a stock or an index fund, textual information on the same is generally less straightforward and more time-consuming to interpret.

Natural language processing (NLP) can be used to analyze the deluge of articles reporting on the stock market. In particular, sentiment analysis algorithms can score a piece of text on whether the writing conveyed a positive or negative sentiment on average. Although, reporting in general focuses on what happened, readers following the financial markets are often more interested in the future. Hence, many articles in this domain are written as predictions. Aggregating the scores of articles that come out each day, could a statistically significant signal be found to predict stock market fluctuations?

The goal of this project is to determine whether the sentiment scores of financial articles can be used to predict what will happen with the stock market. Ultimately, fluctuations in the closing numbers from the S&P 500 index will be related to sentiment scores of CNBC financial article descriptions. A time-series analysis will be used to find whether there is a clear signal between the sentiment analysis scores and the indexed fluctuations.

## Dataset

The list of articles to be used in this project comes from the Financial News Headlines dataset hosted on Kaggle [1] while the S&P 500 data for the same time range is easily accessible online from various sources like the Wall Street Journal [2].

Three news sources are represented in the complete dataset. Two are British news sources: The Guardian and Reuters. The third one, CNBC, is American owned. Below are descriptions of the datasets as they were given on the Kaggle page.

- Data scraped from CNBC contains the headlines, last updated date, and the preview text of articles from the end of December 2017 to July 19th, 2020.
- Data scraped from the Guardian Business contains the headlines and last updated date of articles from the end of December 2017 to July 19th, 2020 since the Guardian Business does not offer preview text.
- Data scraped from Reuters contains the headlines, last updated date, and the preview text of articles from the end of March 2018 to July 19th, 2020.

## Exploratory Data Analysis

Figure 1 shows a sample of the CNBC dataset. There are only three columns: headlines, the article's publication date and time, and its description. Note that some records are empty; these are eliminated. Another issue is that a few articles are updated and appear multiple times in the dataset. For these articles, we take only the first publication time and drop the later records.

| | Headlines | Time | Description |
|---|---|---|---|
| 0 | Jim Cramer: A better way to invest in the Covi... | 7:51 PM ET Fri, 17 July 2020 | "Mad Money" host Jim Cramer recommended buying... |
| 1 | Cramer's lightning round: I would own Teradyne | 7:33 PM ET Fri, 17 July 2020 | "Mad Money" host Jim Cramer rings the lightnin... |
| 2 | NaN | NaN | NaN |
| 3 | Cramer's week ahead: Big week for earnings, ev... | 7:25 PM ET Fri, 17 July 2020 | "We'll pay more for the earnings of the non-Co... |
| 4 | IQ Capital CEO Keith Bliss says tech and healt... | 4:24 PM ET Fri, 17 July 2020 | Keith Bliss, IQ Capital CEO, joins "Closing Be... |

*Figure 1*: Sample of CNBC raw data.

Although the Guardian and Reuters datasets are similar, the Guardian, as previously noted, does not contain the article descriptions. In addition, these sources only contain the date and not the time. To normalize all the news sources into one dataset, the Guardian descriptions are simply the headlines and only the date component of the datetime of the CNBC articles is retained. The final dataset contains 53,158 records.

## Sentiment Analysis

The first step after cleaning the data is to calculate sentiment analysis scores. Assigning a sentiment score to a piece of text can be complicated. Academic texts convey information in a different way than one might send a text to a friend. A domain specific publication might have distinct connotations to certain words than what would be accepted in common vernacular. For instance, the word “bullish” in a financial publication can denote very favorable market conditions, whereas the same word takes on a negative connotation to the layperson. In addition, a sequence of words, when taken together, may convey a strongly different sentiment than the words taken individually. Take the bigram, “pretty ugly”, where pretty intensifies the negative sentiment rather than canceling it out.

Python has a popular and versatile natural language processing (NLP) package known as NLTK (natural language took kit). As part of NLTK, a sentiment analysis algorithm, VADER (Valence Aware Dictionary and sEntiment Reasoner) which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works as well on texts from other domains [3]. The VADER tool provides a convenient way for us to assign a sentiment scores to the article headlines and descriptions text. The tool output of the VADER tool is an assigned score between -1 (very negative) to 1 (very positive) to the analyzed text.

Before applying the VADER algorithm, certain words that do not carry much meaning should be filtered out from the news headlines and descriptions. Common emotionally inert words like the articles “the” or “a” push the overall sentiment score towards zero. Hence, to boost the sentiment signal from

the text, removing these types of words is paramount. To that end, all common English stop words were removed from both the news headlines and descriptions. In addition, commonly appearing financial words such as “money” or “stock” were also removed since they convey no sentiment. Finally, CNBC often references source specific terms such as “Jim Cramer” and “Mad Money”, which also were removed. It is possible to further improve the lexicon by assigning custom sentiment scores to domain specific words such as “bullish” or “bearish” to even better characterize the intended sentiment of the article. We could go even further to identify meaningful n-grams such as “stocks rise” to improve the sentiment scores. However, these items are left as topics for future work.

After preprocessing the textual data, the VADER scores were calculated for the article headlines and descriptions in addition to an averaged score of the two. Figure 2 shows the sentiment scores of the article descriptions plotted over time. Interestingly, there is almost no overlap between sentiment scores of the three news sources. CNBC has the most positive scores, and in fact never dips below zero on any day. Reuters is generally positive, but also begins dipping into negative scores after the COVID pandemic started in early 2020. Finally, the Guardian has negative sentiment analysis scores over the entire period of the data collected. Due to these differences in average sentiment, a fair comparison between any fluctuations in the financial markets necessitates normalization of the scores by subtracting their respective means.

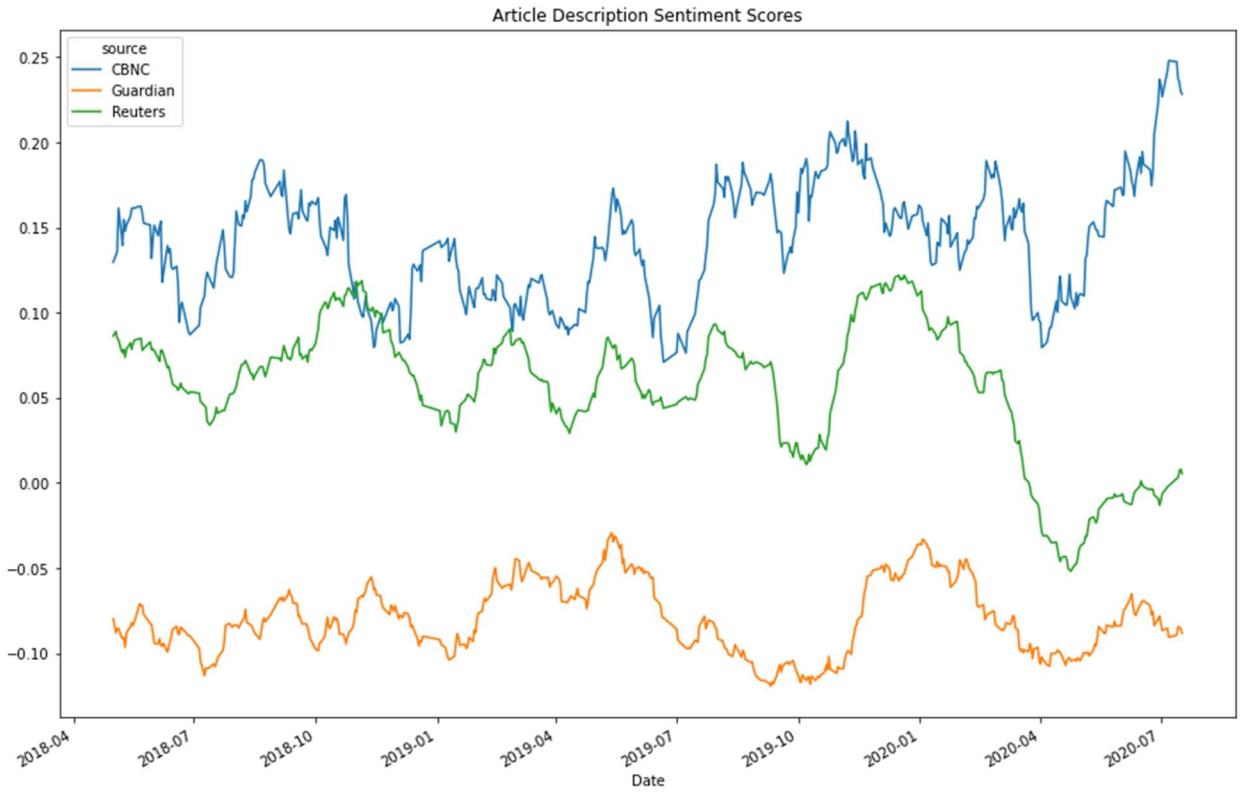


*Figure 2*: 30-day moving average of sentiment scores for each news source's article description.

## Stock Market Data

There are numerous options both domestically and internationally when it comes to quantifying the stock market. In the U.S., the Standard and Poor's 500 (S&P 500) is a stock market index that tracks the performance of 500 large companies listed on the stock exchanges in the U.S. with a market cap of near $40 trillion as of November 2021. [4] As one of the most followed and influential indexes, the S&P 500 is chosen to compare the article sentiment scores against. [2]

The sentiment scores of the article descriptions reflect changes or trends in the stock market data. Therefore, the raw S&P 500 closing prices need to be differenced in order to be compared directly with the sentiment scores. In addition, the sentiment scores need to be averaged at least by the day as there are multiple articles published on any given day. However, the S&P 500 data is only collected on non-holiday weekdays when the markets are open. Therefore, both the S&P 500 fluctuations and the article description sentiment scores are averaged on a weekly basis.

## Correlation between Sentiment Scores and Stock Market Fluctuations

The similarity of the sentiment scores and the market fluctuations is quantified with the Pearson R correlations. In all cases, the correlation is negative and not statistically significant (alpha = 0.05). In the case of the British news outlets, the correlation is almost zero (-0.05 with p = 0.56 and -0.06 with p = 0.48 for the Guardian and Reuters, respectively). For CNBC, the correlation is -0.10 and p = 0.25, which indicates a slightly better signal albeit not a statistically significant correlation. One possible reason for the better correlation for is that while both the Guardian and Reuters are British outlets, CNBC covers American markets. Hence, the correlation with the U.S. S&P500 is unsurprisingly better with a U.S. news publisher. For the remainder of this analysis, we focus only on the CNBC scores. Moreover, we will restrict the analysis to only the article description sentiment scores as they tend to better represent the true sentiment of the report.

### Lagged Correlation

Although the correlation between the CNBC sentiment scores and the market fluctuations is not statistically significant (alpha=0.05), perhaps there an optimal lag between the news publication date and the averaged weekly market fluctuations? Plotted in Figure 3, the correlations between the CNBC article description sentiment scores and the S&P 500 changes with the averaged market fluctuations differenced between -6 and + 6 weeks. The maximum correlation (Pearson R = 0.25 with p=0.01) appears at +4 weeks and is statistically significant. In other words, the CNBC article sentiment scores are most highly correlated with the index fluctuations 4 weeks in the future.

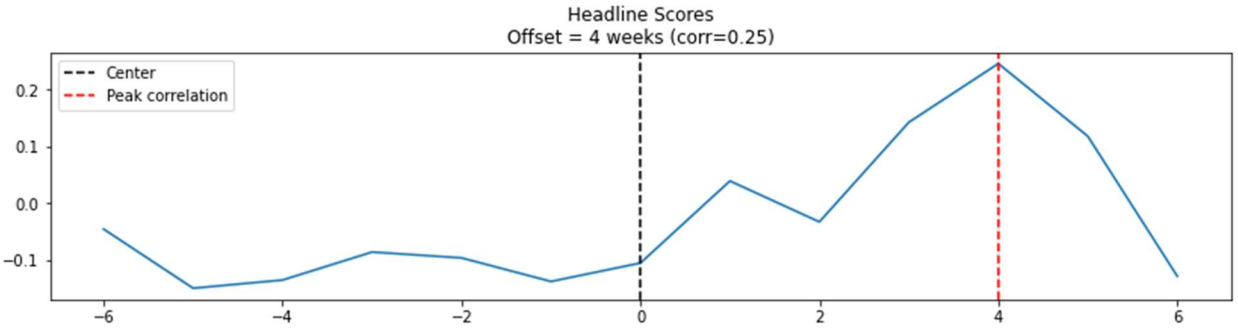


*Figure 3*: Correlation between CNBC article description sentiment scores and lagged S&P 500 fluctuations.

## Moving Average Smoothing

Although the correlation between the sentiment scores of the articles and the S&P500 significantly improved with the 4-week offset, it can still potentially be improved by using a moving average to smooth the data. A moving average window between 1 and 14 weeks was tested (Figure 4), and we found that a window size of 6 weeks yielded the highest correlation of 0.41, or over 1.5 times the correlation without the moving average.

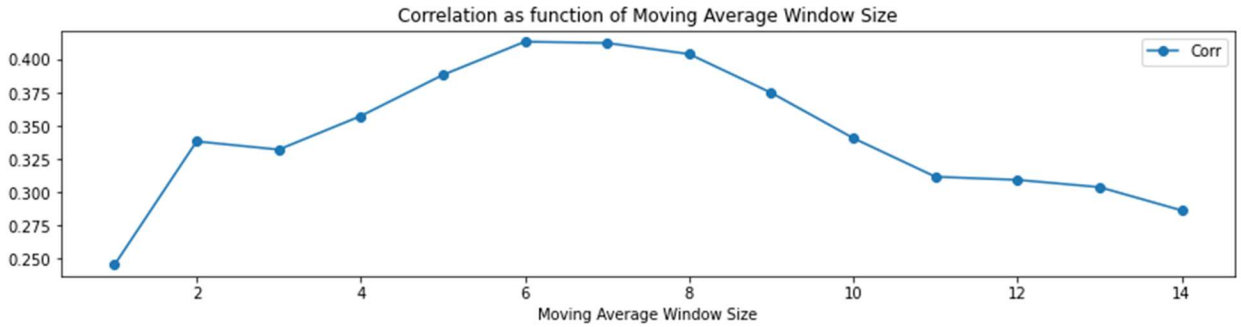


*Figure 4*: Correlation between CNBC article description sentiment scores and lagged S&P 500 fluctuations as a function of moving average window size.

## Epoch Correlations

So far, we have focused on averaged correlations between the sentiment scores and market fluctuations. However, it may be interesting to see how the correlation changes over time. Figure 5 is a heatmap that shows the correlations between the 4-week shifted S&P 500 fluctuations against the CNBC article description scores aggregated by a moving average of 6 weeks across 10 epochs. Each epoch is 11 weeks long and shows the correlations between -6 and +6 week offsets from the optimal 4-week shift. Note that, by and large, at offset = 0, the correlation is almost always positive or close to neutral, suggesting that our averaged positive correlation is meaningful. If the correlation were to jump between positive and negative, like seen when examining offset values around 0, the noise would make any predictions of the stock market fluctuations unreliable. Interestingly, some epochs have consistently positive or negative correlations across all offset values.

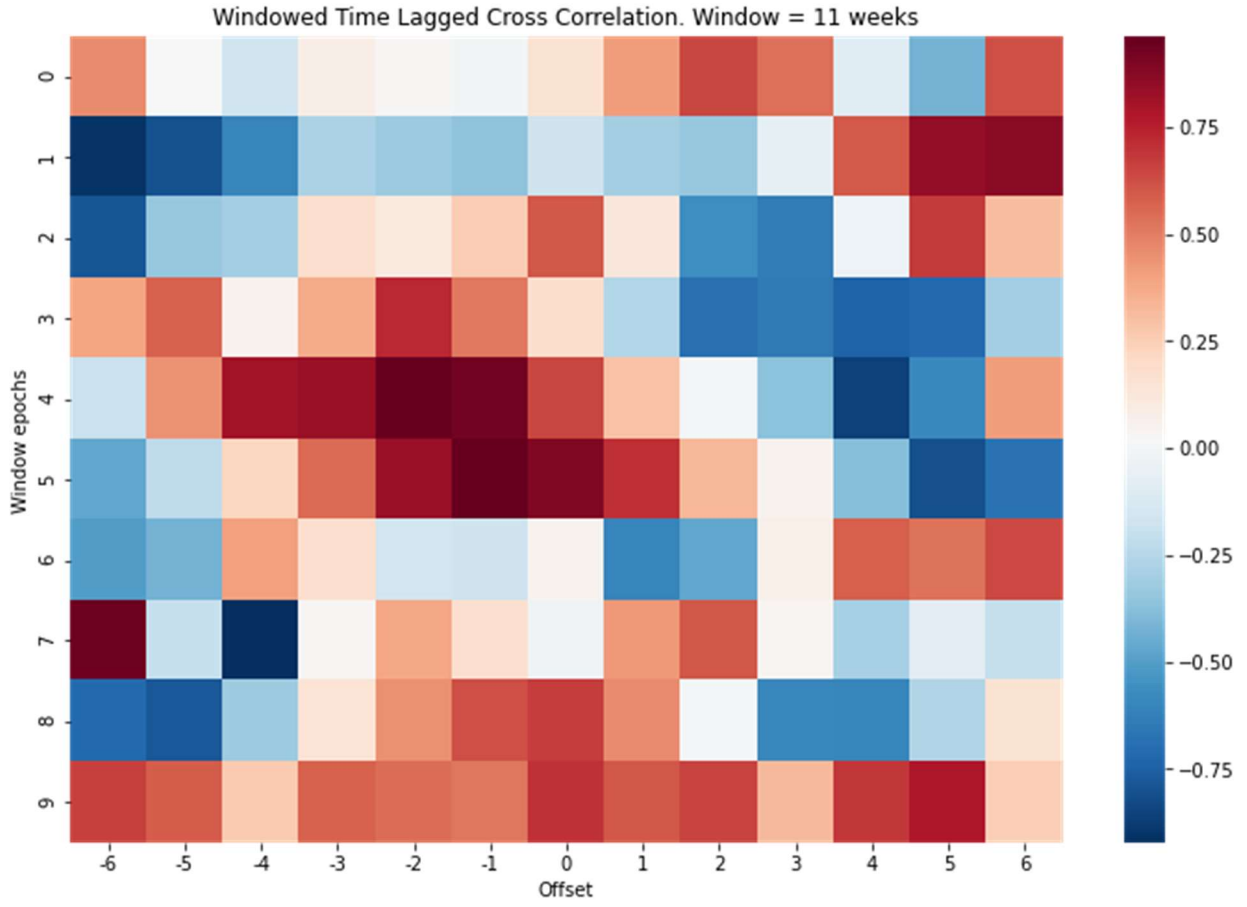


*Figure 5*: Heatmap of correlations across ten 11-week epochs.

## Time Series Modeling

Building on the above analysis of the correlations across epochs, what if there was a pattern with how the correlations varied over time. For instance, does CNBC have regular intervals where the sentiment of their articles aligns more closely with the market and other periods where the correlation is not quite as strong. Perhaps there are times of the month or year where reporting is less focused on what is currently going on in the market and rather takes on a more retrospective or predictive nature.

### Best Model Fit

With a time series model, we can look at the difference between the sentiment analysis scores of the CNBC article descriptions and the S&P 500 fluctuations to identify any patterns (Figure 6). An augmented Dicky-Fuller test indicates no evidence that the series in Figure 6 has any unit roots (p=0.40). In other words, we can say that the time series is stationary.

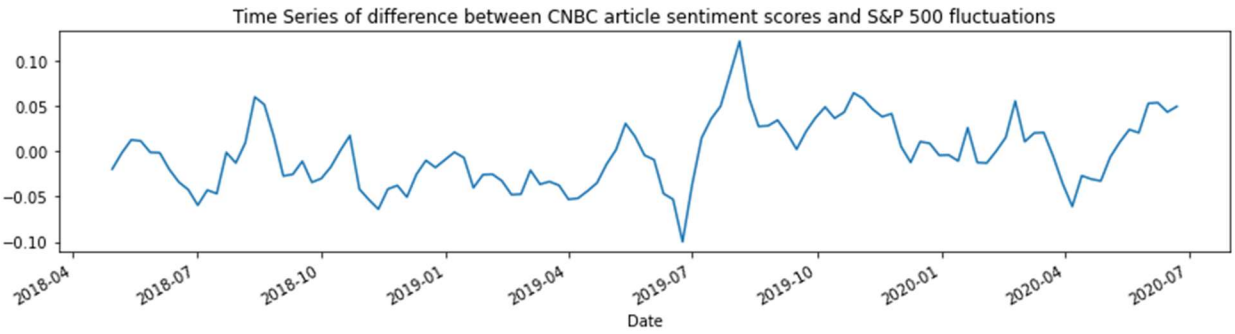


*Figure 6*: Time series of difference between CNBC article sentiment scores and S&P 500 fluctuations.

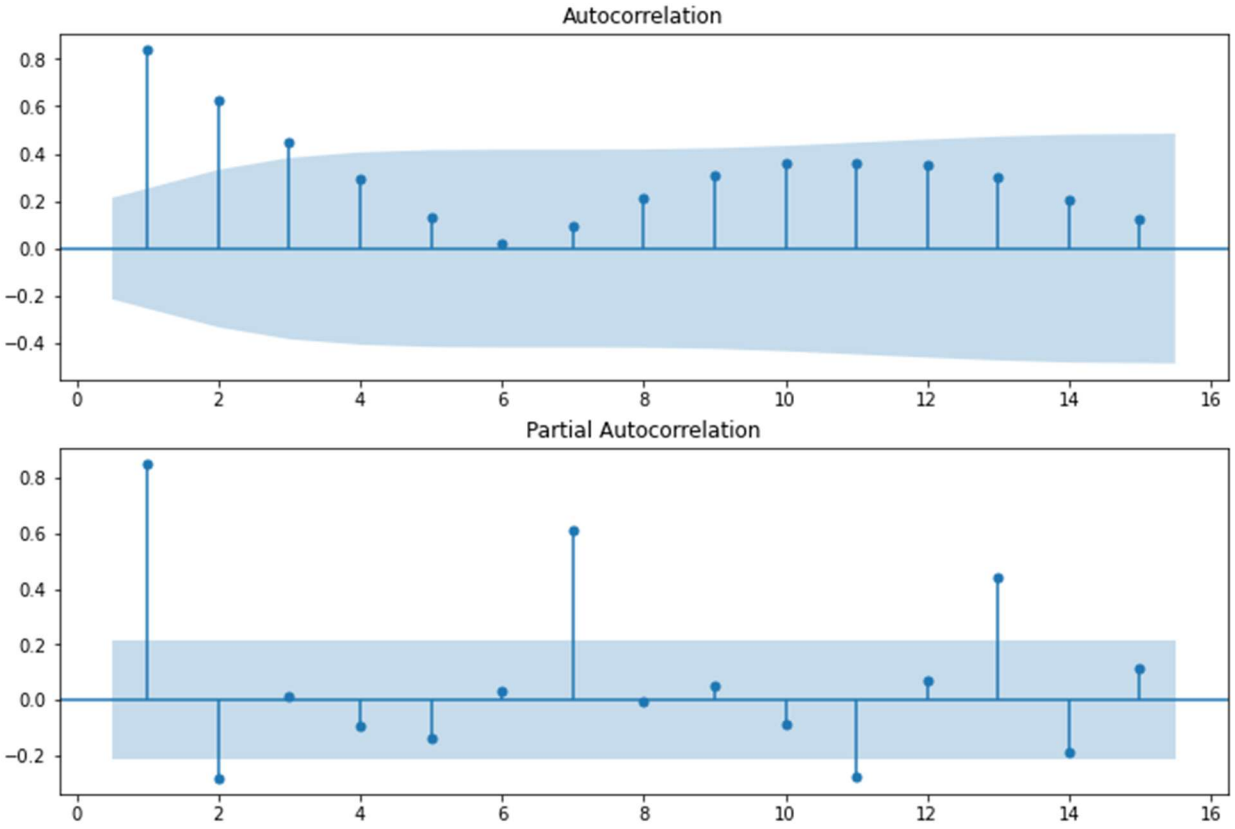


*Figure 7*: ACF and PACF correlogram plots of time series of difference between CNBC article sentiment scores and S&P 500 fluctuations.

Establishing that the time series is stationary, we calculate the autocorrelation (ACF) and partial autocorrelation (PACF) functions and plot them for lags between 1 and 16 weeks as a correlogram (Figure 7) for the first 75% of the times series, leaving the remaining quarter for testing purposes. From the plots,

it looks like there are both autoregressive and moving average components. Estimating the ARMA(p,q) time series model from the correlograms is done by observing whether there is a geometric decay or a number of discrete, significant lags. The decay of the ACF function with 3 significant lags in the PACF plot suggest an AR(3) model. However, a more definitive analysis is done with a grid search to find the best model parameters. Using both the AIC and BIC to determine goodness of a given time series candidate model's fit, we find that an ARMA(2,3) model is the winner (Table 1). Interestingly, the simpler AR(2) model is only the 2nd best as ranked by BIC.

*Table 1*: Top 6 ARMA models in order of ascending BIC.

| p | r | q | AIC | BIC |
|---|---|---|---|---|
| 2.0 | 0.0 | 3.0 | -408.368233 | -391.352515 |
| 2.0 | 0.0 | 0.0 | -398.124524 | -388.401257 |
| 1.0 | 0.0 | 1.0 | -397.754265 | -388.030998 |
| 1.0 | 0.0 | 0.0 | -394.557851 | -387.265401 |
| 2.0 | 0.0 | 1.0 | -396.144480 | -383.990396 |
| 3.0 | 0.0 | 0.0 | -396.137147 | -383.983063 |

## Forecasting

With our ARMA(2,3) model we can now forecast and test our predictions on the holdout test time series. First, it is important to establish a baseline prediction as a benchmark for our model's accuracy. In this case, we simply use the last value of the training time series and forecast, giving an MSE of 0.0008.

Two major types of forecasting with time series are multiple-step ahead and one-step ahead predictions. A one-step ahead prediction uses only the last value of the time series and then predicts the next value using the fitted model. On the other hand, a multiple-step prediction will continue making predictions past the immediately following value using the prior predicted value. Figure 8 shows the fit using these two algorithms. Each series of predicted values is shown in red on top of the actual data. The red shading represents the 95th percentile confidence interval around the prediction.

The multi-step ahead forecast for the ARMA(2,3) model initially dips and then increases, but quickly plateaus due to the lack of additional shocks. In the case of the one-step ahead predictions where

we use the latest data point in the time series rather than the last forecasted prediction, the fit is unsurprisingly much closer to the actual data (MSE = 0.00038 vs 0.00096). Compared to the baseline prediction with an MSE of 0.0008, the one-step ahead prediction has half of the error. However, the multi-step forecast preforms even worse than the naïve prediction.

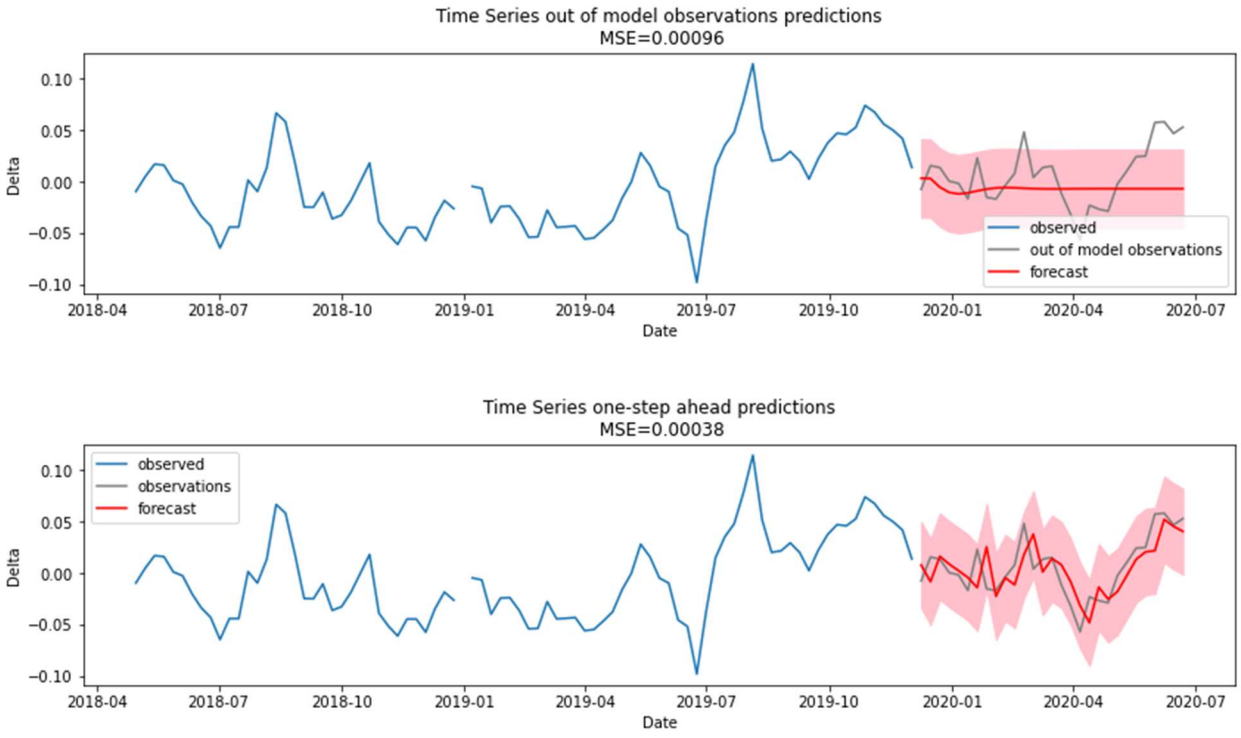


*Figure 8*: Time series forecasting using multiple steps and single-step ahead algorithms.

## Conclusions and Future Work

A relatively strong and statistically significant correlation between averaged CNBC article description sentiment scores and S&P 500 fluctuations exists for market changes 4-weeks in the future. In other words, the sentiment of the articles today can be used to predict how the index will change in approximately a month. This trend continues to, by and large, hold even when the data is broken into 10 epochs.

To reveal any potential patterns existing in the difference between the averaged sentiment scores and the index fluctuations time series, a grid search to find the best ARMA parameters was done. The best

model based on BIC and AIC was ARMA(2,3). Using a one-step ahead prediction, an MSE half of what was calculated using a naïve last-value prediction is achieved.

To potentially improve the measured correlation between the article description sentiment scores and the index fluctuations, a more sophisticated sentiment scorer could be used. In particular, the scorer could be trained on financial data with terms such as “bull market” assigned a very favorable score. Another source of improvement would be to filter out more stop words, such as location, company, or people names which do not contribute to either a positive or negative sentiment.

A more robust comparison in the error of the time series model predictions against the naïve forecast could be done with windowed comparison rather than a simple 75/25 split of the entire time series. We assumed that the time series would continue being well modeled by an ARMA(2,3). However, the nature of the time series could easily change over time, and continually fitting a new time series as data comes in would very probably result in a lower MSE for the one-step ahead forecast.

## References

[1] https://www.kaggle.com/notlucasp/financial-news-headlines

[2] https://www.wsj.com/market-data/quotes/index/SPX/historical-prices

[3] https://github.com/cjhutto/vaderSentiment

[4] https://ycharts.com/indicators/sp_500_market_cap