A COVID-19 Stock Market Prediction Model Using Financial Data and News Sentiment

Nablul Haseeb

Department of Computer Science, CUNY Hunter College

Email: nablul.haseeb63@myhunter.cuny.edu

Abstract

The spread of the COVID-19 virus has globally dominated news cycles in 2020, and has had prominent impact on the international financial markets. Officially declared a global pandemic by the World Health Organization (WHO) in March 2020, significant uptick and downturns in the financial market can be correlated to public sentiment towards future outlook on the containment of the pandemic. The goal of this paper is to analyze the effects of existing financial information and news data on the predictability of future stock prices during the global COVID-19 pandemic. Financial parameters of stocks such as daily closing price, high price, low price and market volume are considered along with sentiment scores for daily news headlines related to stocks and COVID-19. To evaluate the accuracy of the prediction model, three distinct prediction models are designed using a neural network architecture with different combinations of input parameters. The broader goal for this project is to develop and establish a financial market prediction model that can be utilized during future global pandemics to minimize the loss of financial capital.

Table of Contents

Introduction

Overview

Over the last few decades, the use of artificial intelligence has permeated different facets of the financial sector. Particularly, there has been significant interest in analyzing stock price and index patterns to predict their future values. Given that the global movement of financial capital is currently at an all time high, the dynamic feedback of computerized prediction models provide seemingly instantaneous financial guidance. Various forms of such artificial intelligence driven prediction models are used by financial institutions in making investment decisions.

Stock prices are known for being dynamic and susceptible to quick changes due to impact from financial parameters, as well as external factors. Notable financial parameters of a stock consist of its current market price (CMP), daily high and low prices, market volume and market capitalization. Records of these parameters for different stocks are accessible via stock exchanges such as the NASDAQ, New York Stock Exchange, London Stock Exchange etc. Stock prices are also often significantly impacted by external factors such as political unrest, health crises or general public sentiment.

Although there are numerous prediction models for stock prices that include or exclude different sets of these parameters, none of these techniques have proven to be fully accurate and reliable in their predictability. This can be attributed largely to the unquantifiable nature of the external factors affecting stock prices - although the financial parameters such as price and market capitalization are available in quantifiable format for processing, there is no standardized procedure for retrieving or pre-processing subjective external factors such as news articles or public sentiment from social media posts.

A major event that has dominated public sentiment, and thus indirectly the stock market, has been the global COVID-19 pandemic of 2020. For the majority of the year, stock prices have fallen with worsening news of the pandemic, and improved with hopeful news of a potential vaccine development. The public's sentiment towards the condition of the pandemic can be correlated with the general shift in the financial market. As a result, in order to predict and understand the behavior of stocks and other financial capital in future global health crises, it is

critical to understand the relationship between the external factors that have impacted stock values during the COVID-19 outbreak.

Motivation

Due to the monumental impact of the COVID-19 pandemic on the financial market during the latter part of 2019 and entirety of 2020, the goal of this project is to examine and develop a stock predict model using both financial factors as well as external factors. To analyze the external factors, news articles are mined and processed using natural language processing (NLP) techniques to establish a quantitive sentiment value. These sentiment values and other stock attributes are then fed through a neural network architecture to develop a prediction model.

A primary motivation behind this project is the ability to push the forefront of artificial intelligence use in the realm of stock prediction. Up until this point, the majority of stock prediction models have utilized historic quantitive data to make future projections. The incorporation of sentiment analysis allows for a far greater range of external factors to be considered in the prediction model, while simultaneously increasing the versatility of the model to react to real world events that can indirectly influence stock values. The development of such a comprehensive model can greatly benefit the financial sector, and change the overall landscape of dynamic financial data mining and processing.

Another major motivation behind this project is the potential ability to project financial market behavior during future global health crises. The data collected and processed during the COVID-19 pandemic can be used to correlate the effects of public sentiment with their impact on the financial market. For example, depending on the expected release date of a potential vaccine, financial institutions and governments can anticipate the expected rate of growth or decline of stocks during that time frame. This can in turn minimize the loss of financial capital, and help stabilize the job market.

Finally, this project and similar projects that analyze the impact of a global health crisis through sentiment analysis can be used to extrapolate the prediction models beyond the financial sector to other industries. For example, sentiment analyses can be conducted to predict the likelihood of business failures based on the population's consumer habits. Similarly, social media

feeds can be mined to examine the population's receptiveness towards a certain vaccine or medical product. The scope of application for sentiment analysis is vast and presents a large frontier for artificial intelligence development.

Scope

The goal of this paper is to analyze the effects of existing financial information and news data on the predictability of future stock prices during the global COVID-19 pandemic. To understand the holistic effect of the pandemic on the financial market, a wide range of (16) different stocks are considered, which are spread across (4) different financial market sectors. The study is focused on the time period ranging between October 2018 and October 2020. Pre-pandemic data is considered to establish a baseline for the prediction model prior to the market fluctuations and unrest caused by the spread of the virus.

For financial information, daily closing price, high price, low price and market volume for stocks are considered. Daily stock news is retrieved and processed to determine a quantitative sentiment score. In addition, daily top headlines are processed for a separate COVID-19 sentiment score which is also input into the prediction model. A lexical analysis approach is used to establish a sentiment score for both stock news and COVID-19 news. While Natural Language Toolkit (NLTK) with Python 3.8 is primarily used to analyze stock news data, a custom rule-based lexical model is used to analyze the news for COVID-19 sentiment. To evaluate the accuracy of the prediction model, three distinct prediction models are designed using a neural network architecture with different combinations of input parameters. The financial stock information and news data are retrieved from *Google Finance* and *New York Times*, respectively, using non-commercial licenses.

The broader goal for this project and related future works is to develop and establish a financial market prediction model that can be utilized during future global pandemics to minimize the loss of financial capital. Such a prediction model can also help reduce unemployment and help stabilize market conditions by raising investor confidence in market predictability. Finally, further research in this field can also help extend similar models to other industries outside of finance.

Review of Related Works

Stock Prediction Models

Stock price prediction models have been a focus of both academic and industrial research for several decades. Although numerous prediction models and techniques have been developed over the years, most of such models can be classified into one of the following categories:

- Fundamental Analysis
- Technical Analysis
- Efficient Market Hypothesis and Random Walk Hypothesis
- Machine Learning and Artificial Intelligence based analysis
- Sentiment Analysis based on financial news

Fundamental Analysis of stock prices relies on holistic studies of a company's profile in order to make predictions regarding its future stock value. Such in-depth analyses take into account the company's performance and profitability, along with its assets, debts, infrastructure and investments. Agrawal *et al.* states that based on Fundamental Analysis, the market price of a stock always tends to move towards its "intrinsic" or "real" value. Valuations of stocks are performed based on whether the current price is above or below this theoretical threshold. According to research performed by Agrawal *et al.*, important parameters that are often considered for Fundamental Analysis often include the Price-to-Book Ratio (P/B), the Price-to-Earning Ratio (P/E), dividend yield and Debt-to-Earning Ratio amongst others. A primary disadvantage of a stock prediction model that solely relies on Fundamental Analysis is that such a model is often hard to formalize and automate, and primarily depends on subjective human input.

Unlike Fundamental Analysis, Technical Analysis relies on historic quantitive parameters of a stock in order to predict its future value. According to Shah's research, these quantitive parameters can be represented as *indicator functions* that influence the market value of a stock. Common indicator functions used in Technical Analysis are Moving Average (MA), Exponential Moving Average (EMA), Rate of Change (ROC), Relative Strength Index (RSI) etc. Research conducted by Agrawal *et al.* finds that approximately 90% of major stock traders rely on some

form of Technical Analysis to guide their stock purchasing decisions. Despite the wide spread use of this technique, Fundamental Analysis has been criticized for failing to account for external factors that affect stock prices, and for being too variant and unpredictable depending on implementation technique (Sureshkumar *et al.*).

A subset of stock price prediction models state that the future price of a stock is entirely unpredictable based on historic data (at least to a useful extent). In his research Shah classifies this model as the Efficient Market Hypothesis (EMH), stating that in such a model the market always sets the stock's price at its *ideal* or *efficient* level, regardless of the current price. Similar to the Efficient Market Hypothesis, the Random Walk Hypothesis states that the movement of stock prices resemble a random walk pattern, without any clear deterministic trends. However Shah states that the availability and development of intricate prediction models using Machine Learning techniques challenge the EMH and Random Walk Hypothesis.

In recent decades, Artificial Intelligence driven and Machine Learning based stock prediction techniques have grown in popularity. Ganatr *et al.* developed a neural network for stock predict using an R-tool implementation. The network received closing price, stock turnover, global indices, interest rate and inflation as the network input parameters, and had the future price prediction as its output parameter. In benchmark comparison with existing Technical Analysis models, the neural network provided improved prediction accuracy. Atsalakis *et al.* used a similar neural network design along with integrated soft computing techniques to develop a prediction model with significant prediction accuracy compared to conventional statistical models.

The forefront of stock prediction models rely on sentiment analysis to quantify the market's opinion and disposition towards a stock. Such methods can also be used to establish an overall understanding of the financial environmental, and determine if investors are optimistic or weary. One of the most notable works in this area is by Bollen, who studied the public sentiment (happy, calm, anxious) derived from twitter feed, and correlated the results to the Dow Jones Industrial Index. Bollen found that using a fuzzy neural network architecture to process the data yields a strong correlation between the public sentiment and the movement of the Dow Jones. In a similar study, Zhang found a strong negative correlation between public mood states

(optimistic, pessimistic, hopeful, fearful) and the Dow Jones Average Index. Dickinson *et al.* investigated the correlation between public mood sentiment and stock prices using the Pearson correlation coefficient.

Alongside social media feeds such as from Twitter, news articles also provide a reliable source of general sentiment and opinion held by the public. Niederhoffer analyzed New York Times, and classified two decades of top headlines into 19 categories ranging from extremely good to extremely bad. His research finds that the financial market tends to have an overreaction to bad news in comparison to good news. Davis *et al.* performed a similar study researching the effects of optimistic and pessimistic language used in news on a company's future stock performance. The research concludes the following: (1) a bias exists between the writer's intension and the reader's interpretation of the news and (2) readers react strongly to news that reinforce what they have already read in other news articles. In his research, Tetlock finds that trading volume of a stock generally increases following pessimistic reports, and highly pessimistic reports are typically followed by a reversion and adjustment of market prices.

Sentiment Analysis Using NLP

The explosive growth of user-generated text on the internet has increased research interest in NLP and sentiment analysis of text over the last few decades. Accurate interpretation and classification of text can have significant impact on the frontier of artificial intelligence, and have consequential effects on research and industry. Sentiment analysis approaches can be divided into two major categories of research: (1) Lexical Approaches and (2) Machine Learning Approaches.

A lexical approach utilizes a lexicon i.e. a dictionary of pre-tagged words. Each word within the lexicon is assigned a sentiment polarity score. For example, the word "great" may be assigned a positive sentiment polarity score, while the word "horrible" may be assigned a negative one. For a lexical sentiment analysis of a text, each word in the text is searched in the lexicon, and the word's sentiment polarity score is added to the overall polarity score of the sentence (Michelle *et al.*). At the end of the process, the cumulative sentiment polarity score of the text is used to determine its degree of positivity/negativity. Lexical analyses constitute the

early stages of NLP approaches, and despite their simplicity, yield fairly accurate results for simple sentence structures (Kennedy *et al.*).

Since lexical analysis relies on the accurate mapping of words in a text to their respective sentiment polarity score in the lexicon, numerous researches have been conducted to determine the most accurate assignment of sentiment polarity scores. In their research, Hatzivassiloglou and Wiebe found that the sentiment analysis of single-phrase texts can be performed with a 80% accuracy by only using a lexicon of pre-tagged adjectives. In a similar research, Kennedy *et al.* used a pre-tagged adjective lexicon to analyze a dataset of movie reviews and found a smaller accuracy rate of 62%. Turney used internet results to determine the sentiment polarity score of words for his lexical approach, and achieved an accuracy of 65%. In related research, Kamps *et al.* decided to use the WordNet database to determine the sentiment polarity score of the words used in the lexicon. The results reported an accuracy of 64%, similar to Turney's research. In an alternate approach by Turney and Littman, the lexicon was formulated by developing the semantic orientation of words. By subtracting a word's negative semantic orientation from its positive semantic orientation, Turney and Littman were able to achieve an accuracy of 82%.

Alongside Lexical Analysis, the other primary method of sentiment analysis relies on supervised and unsupervised machine learning techniques. In machine learning approaches, classifiers are trained to determine the polarity of texts. The classifiers are trained via an initial input of feature vectors and tagged training data, and tested on text data not viewed during the training process. The selection and determination of classifiers plays a significant role in the effectiveness of the machine learning model (Michelle *et al.*). Common variants of classifiers include unigrams and n-grams (single words or a sequence of n-words within a sentence), number of positive or negative words, sentence length etc. Naïve Bayes classifier, maximum classifier and Support Vector Machine (SVM) are some of the most commonly used models (Sun *et al.*). Michelle *et al.* finds that machine learning sentiment analysis methods typically yield slightly better prediction accuracy compared to conventional lexical approaches, with accuracies ranging between 63% to 82%. However these results are often dependent on the language being analyzed and the classifier features selected.

The research in the realm of NLP and sentiment analysis is a developing field with great potential for growth. As such models become increasingly accurate, they provide the potential for much more reliable inclusion of sentiment analysis in other fields of research such as stock market predictions.

Methodology

The analysis procedure used in this project consists of data selection, data collection, sentiment analysis of textual data and machine learning via neural network. The workflow of the procedure is summarized in Figure 1:

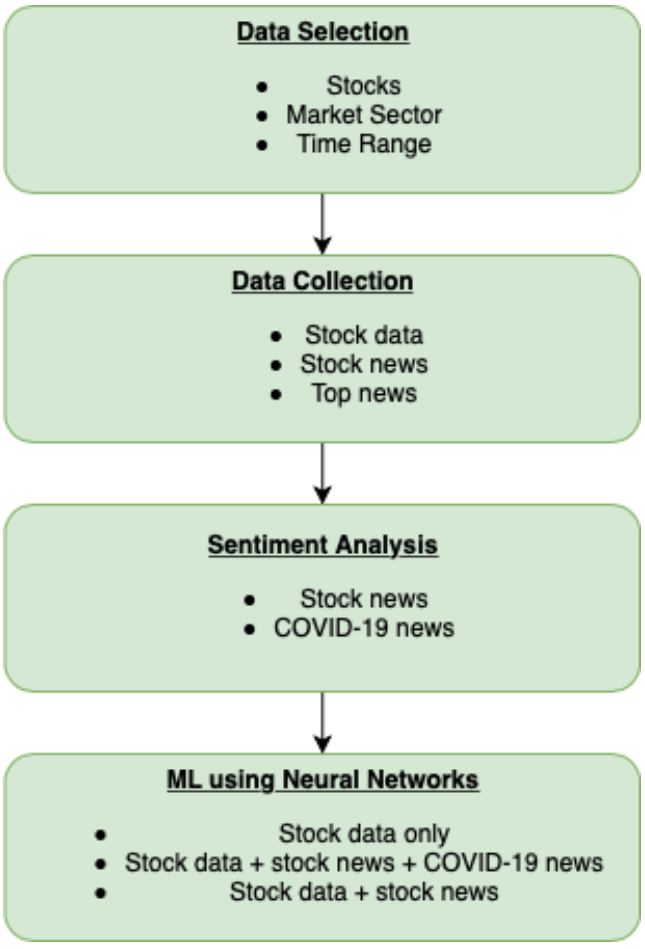

Figure 1: Methodology Workflow

Data Selection

Four different market sectors are analyzed in this project in order to establish a holistic overview of the general financial market. The market sectors correspond with to the Global Industry Classification Standard (GICS), and are as follows: Financial, Information Technology, Consumer Discretionary and Communication Services. Within each sector, four of the largest companies are considered for their stocks. While picking the stocks, consideration is also given to stocks that are frequently featured in the news in order to establish a sizable dataset for news sentiment analysis. A summary of the market sectors and stocks is shown in Table 1:

Market Sector	Stocks
Financial	Bank of America (BAC), JPMorgan Chase (JPM), Berkshire Hathaway (BRK.A), Goldman Sachs (GS)
Information Technology	Apple (AAPL), Cisco (CSCO), Intel (INTC), Oracle (ORCL)
Consumer Discretionary	Tesla (TSLA), Amazon (AMZN), Ebay (EBAY), McDonalds (MCD)
Communication Services	Facebook (FB), Google (GOOGL), Comcast (CMCSA), Verizon (VZ)

Table 1 - Market Sectors and Stocks

Since the goal of the project is to analyze the impact of news sentiment on stock prices during the COVID-19 pandemic, a suitable time range between October 31, 2018 and October 31, 2020 is considered. Data from 2018 and 2019 is considered in order to establish a baseline for the prediction model without large fluctuations of the financial market during the COVID-19 pandemic in 2020.

Data Collection

Daily financial data (closing price, high price, low price, market volume) for each stock is retrieved using Google Finance's API module between October 31, 2018 and October 31, 2020. The top 10 daily news articles for each stock is retrieved using New York Times' API module using a non-commercial license. The top 20 daily headlines for general news is similarly retrieved from New York Times for the same time range.

Sentiment Analysis

The sentiment analysis portion of the project is performed in a series of smaller steps in order to adequately pre-process the news text data retrieved from the internet. Text analytics is performed primarily using the NLTK (Natural Language Toolkit) library for Python 3.8. NLTK is a powerful Python package that provides a diverse range of NLP capabilities. In particular, the VADER module of NLTK, a rule-based model of sentiment analysis built on the NLTK platform, is used in order to process the stock news data. A separate lexical sentiment analysis is performed on the daily headlines for the COVID-19 news. A summary of the sentiment analysis workflow is shown below in Figure 1:

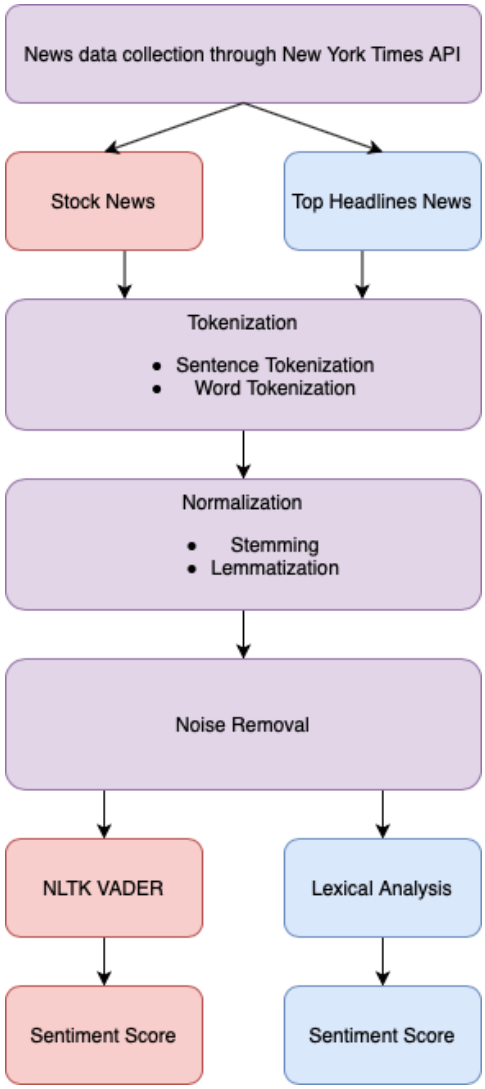

Figure 1: Sentiment Analysis Flowchart

Tokenization:

The first step of text analytics for the news data is tokenization. Tokenization breaks down a larger chunk of the text into smaller pieces. Tokenization is performed in two separate parts: (1) sentence tokenization and (2) word tokenization. Sentence tokenization breaks down a paragraph or multi-sentence headline into individual sentences, while word tokenization breaks down a single sentence into a series of individual words. Tokenization allows the analysis process to handle raw text data as a series of individual handleable input parameters. An example of a short text before and after tokenization is shown below:

Raw text: "COVID-19 wasn't improving. Public anxiousness was high!"

After sentence tokenization: ["COVID-19 wasn't improving", "Public anxiousness was high!"]

After word tokenization: [("COVID-19", "wasn't", "improving", "."), ("Public", "anxiousness", "was", "high", "!")]

Normalization:

The next step in pre-processing the news data for text analytics is normalization. Normalization is the process of converting all words into a common format for more accurate interpretation by the program. Normalization consists of several different steps including converting casing, contraction handling, numeric conversion, stemming and lemmatization. Stemming is a process of linguistic normalization which reduces words to their word root, or removes derivational affixes. Lemmatization is a similar linguistic normalization which reduces words to their correct lemmas. The above example after text normalization is as follows:

After normalization: [("COVID-19", "is", "not", "improve", "."), ("public", "anxiety", "is", "high", "!")]

Noise Removal:

After normalization, the next pre-processing step performed for text analytics is noise removal. As part of this step, all punctuations and miscellaneous symbols are removed the the text. Next, words with neural sentiments, known as stop words, are removed from the text. Examples of common stop words include "the", a", "I", "is" etc. The above example after noise removal is as follows:

After noise removal: [(“COVID-19”, “not”, “improve”), (“public”, “anxiety”, “high”)]

Lexical Analysis Using NLTK VADER:

After the pre-processing steps, Python’s NLTK library is used to perform sentiment analysis on the daily stock news data retrieved for each of the 16 stocks. NLTK provides both machine learning and lexical text analytics capabilities as part of its features. The VADER module, which provides a rule-based lexical sentiment analysis of text is primarily used in this project. VADER provides a library of words with predefined sentiment tags, as well as lexical features such as idiom check, emoticon handling, punctuations emphasis, booster words etc.

Special Lexical Analysis for COVID-19 News:

The daily top headlines are processed separately using a custom lexical analysis to determine the sentiment related to COVID-19 news. The analysis is based on the presence of *trigger words*. If a headline contains one or multiple trigger words, the text is scanned for related positive and negative *sentiment words*, and assigned a sentiment score. The positive and negative sentiments are then amplified by the presence of positive/negative *booster words*. All sentiment scores for a given headline are summed to determine its final sentiment score.

Trigger Words = [COVID, covid, COVID-19, covid-19, corona, coronavirus, virus, pandemic, quarantine]

$$FinalSentimentScore = \sum(POSsentiment) * POSbooster) + \sum(NEGsentiment * NEGbooster)$$

A list of the positive/negative sentiment words and booster words are shown in Tables 2 and 4, respectively.

	Sentiment Words							
Negative	death	-1.0	pandemic	-0.75	COVID-19	-0.75	lockdown	-0.50
	ICU	-0.9	outbreak	-0.75	covid-19	-0.75	shelter	-0.50
	hospital	-0.85	pessimistic	-0.75	corona	-0.75	restriction	-0.50
	hospitalization	-0.85	pessimism	-0.75	coronavirus	-0.75		
	fear	-0.85	COVID	-0.75	virus	-0.75		
	fearful	-0.85	covid	-0.75	quarantine	-0.50		

	Sentiment Words							
Positive	optimism	+0.85	hope	+0.85	vaccine	+0.75	trial	+0.50
	optimistic	+0.85	hopeful	+0.85	test	+0.50		

Table 2 - Positive/Negative COVID-19 Sentiment Words

	Sentiment Booster Words							
Negative	worsen	+2.00	spike	+1.85	rapid	+1.50	improvement	-2.00
	worse	+2.00	increase	+1.75	decrease	-1.75		
	surge	+2.00	again	+1.75	improve	-2.00		
Positive	improve	+2.00	progress	+2.00	decrease	-1.75		
	improvement	+2.00	increase	+1.75				

Table 3 - Positive/Negative COVID-19 Sentiment Booster Words

Neural Network

After pre-processing the news data and developing corresponding sentiment scores, the results are fed through a neural network system in order to use a machine learning approach to design the stock prediction model. For each day between October 31, 2018 and October 31, 2020, the following input parameters for each stock are considered for the neural network:

1. Daily closing price of stock	5. Three-day average news sentiment for the stock
2. Daily high price of stock	6. Three-day average news sentiment for the market sector containing the stock
3. Daily low price of stock	7. Three-day average news sentiment for all market sectors
4. Daily market volume	8. Three-day average news sentiment for COVID-19

To analyze the effects of the news sentiment scores on the overall prediction accuracy of the model, three separate sequential neural networks are considered:

A. 4-layer NN with input parameters 1, 2, 3 & 4
B. 4-layer NN with input parameters 1, 2, 3, 4, 5, 6, 7 & 8
C. 4-layer NN with input parameters 1, 2, 3, 4, 5, 6 & 7

Each of the three neural networks are modeled through python's Keras library, with the geometries shown in the figures below. A batch size of 10, and epoch of 150 is used for training the networks, with ReLU and Sigmoid activations at the hidden and final layers, respectively.

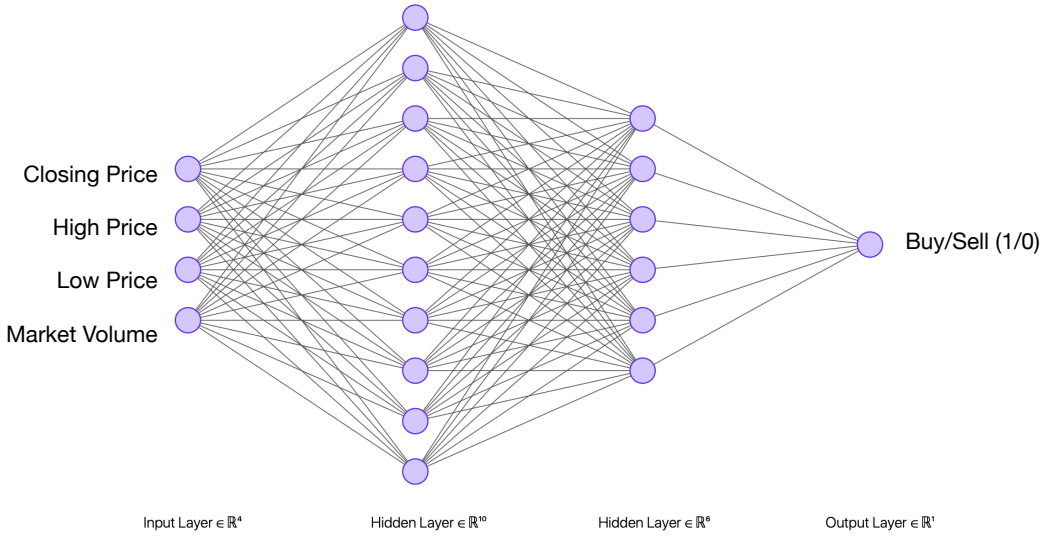

Figure 2: Neural Network (A)

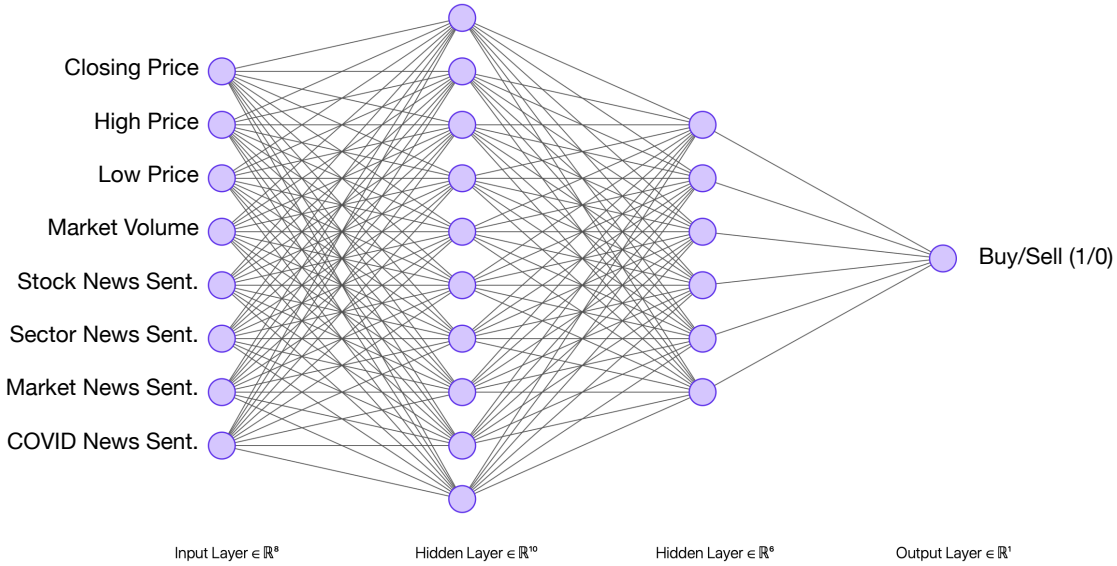

Figure 3: Neural Network (B)

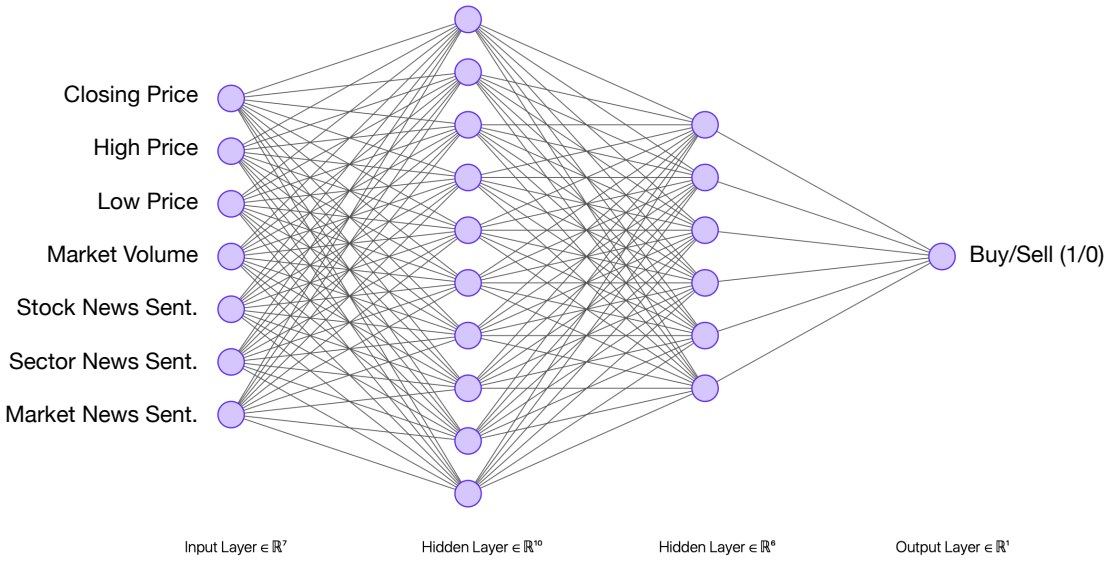

Figure 4: Neural Network (C)

The output of the NN determines whether the price for the stock being considered will decrease or increase the following day (buy vs. sell). Each network is initialized with random weights, and then trained over 80% of the available data for each stock. Finally, each model is tested over the remaining 20% of the data for accuracy.

Results

Stock Parameters and Market Sentiment

This portion of the Results section presents the relationships and correlations between stock parameters such as stock price and volume, and the news sentiment scores outlined in the Methodology portion of this paper. In order to juxtapose the different variables, they are first determined for each day between October 31, 2018 and October 31, 2020, and then normalized within a range of 0 and 1. The normalization step is taken in order to standardize the input going into the neural networks, as well as to develop a comprehensive method for representing the relationships. The following relations are compared in this paper:

- Technical Analysis: Stock Price & Stock Volume
- Sentiment Analysis: Stock Price & Stock Sentiment
- Sentiment Analysis: Stock Price & Market Sector Sentiment

- Sentiment Analysis: Stock Price & COVID-19 Sentiment
- Sentiment Analysis: Market Sector Sentiment & COVID-19 Sentiment
- Sentiment Analysis: Overall Market Sentiment & COVID-19 Sentiment

Technical Analysis: Stock Price & Stock Volume:

In a sentiment-less technical analysis, stock prices are compared with market volumes for each day. Figures 5a through 5d shown below present these relationships for (4) representative stocks [AAPL, AMZN, TSLA & JPM], with the remainder of the stocks and their corresponding relations shown in Figure A1 in Appendix A.

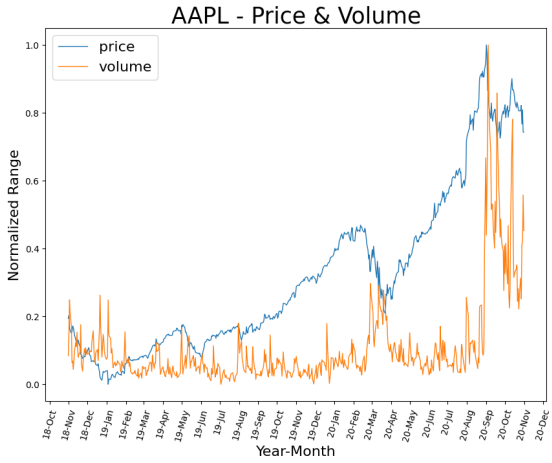

Figure 5a

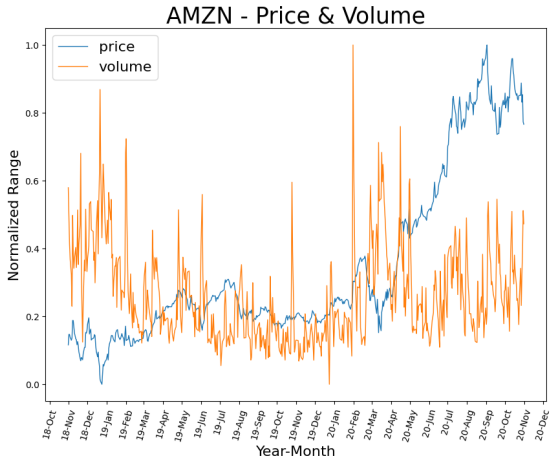

Figure 5b

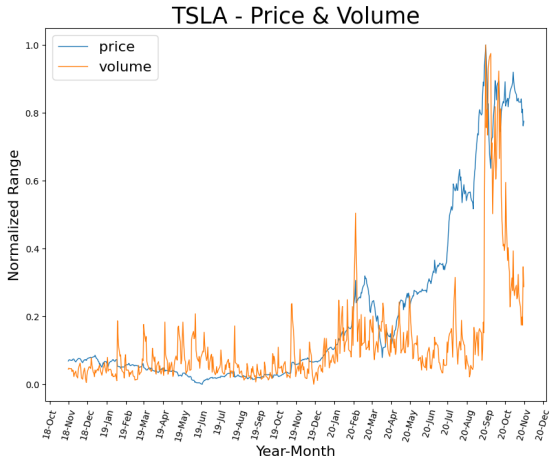

Figure 5c

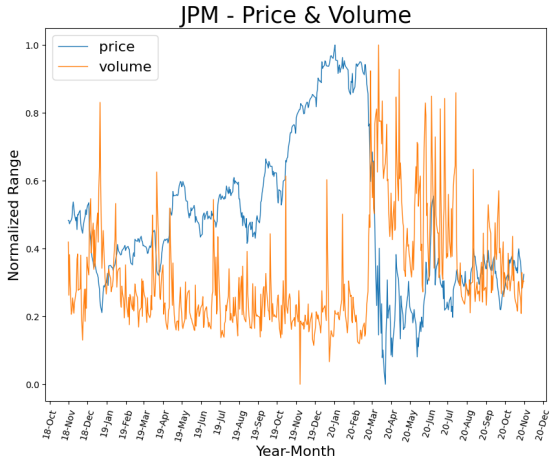

Figure 5d

Figure 5: Stock Price and Stock Volume

Sentiment Analysis: Stock Price & Stock Sentiment:

Next, the prices of each stock is compared with the stock's news sentiment scores for each day. Figures 6a through 6d shown below present these relationships for (4) representative stocks [BAC, GS, JPM & MCD], with the remainder of the stocks and their corresponding relations shown in Figure A2 in Appendix A. For days lacking sufficient news to establish a sentiment score, the average of the last and next available scores is used. This approach is evident through the multiple flat lines shown in Figure 6c for JPM.

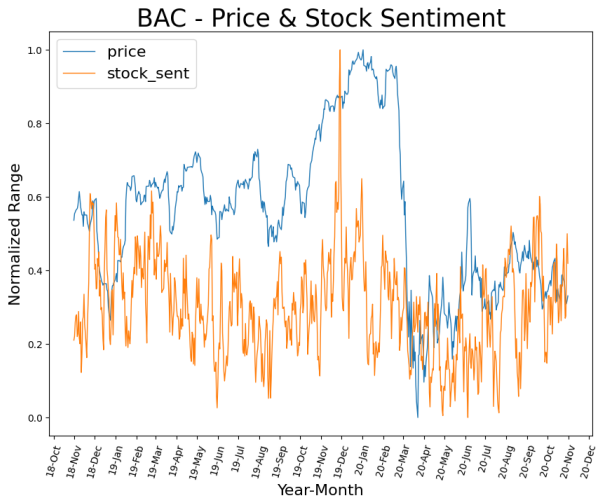

Figure 6a

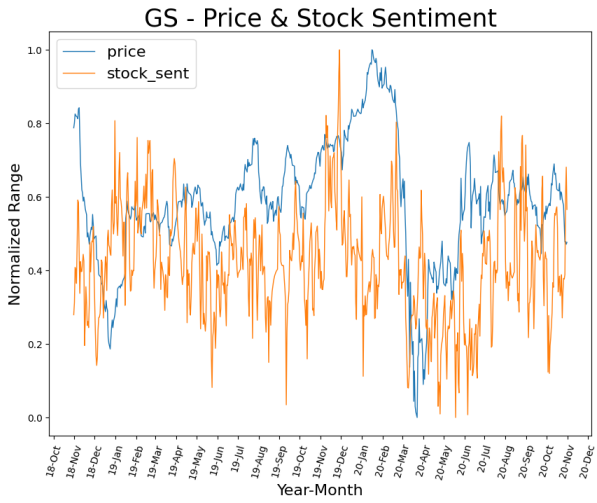

Figure 6b

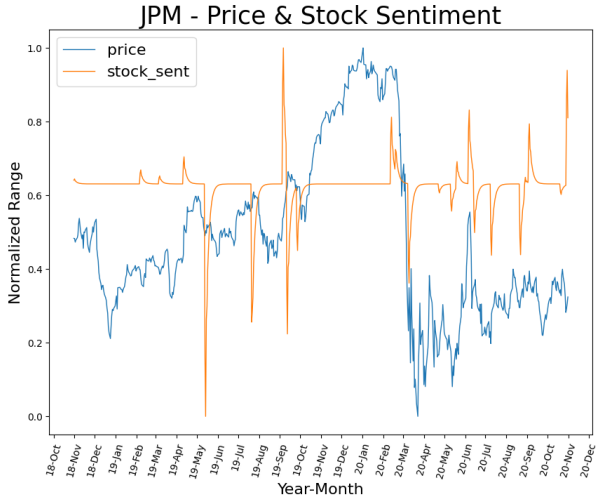

Figure 6c

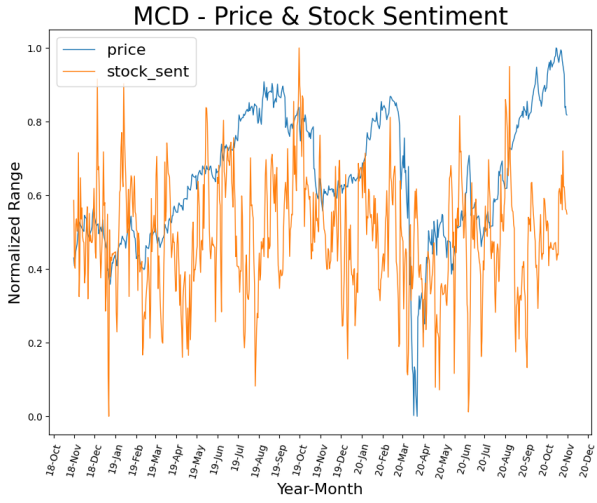

Figure 6d

Figure 6: Stock Price and Stock Sentiment

Sentiment Analysis: Stock Price & Market Sector Sentiment:

As part of the next comparison, the prices of each stock are compared with the news sentiment scores of its corresponding market sector. For example, the stock price of Bank of America Corporation (BAC) is plotted with the news sentiment score of the finance sector. In addition, the news sentiment score for the overall market, averaged across the (16) stocks considered in this paper, is also shown in each plot. Figures 7a through 7d shown below present these relationships for (4) representative stocks [BAC, EBAY, JPM & TSLA], with the remainder of the stocks and their corresponding relations shown in Figure A3 in Appendix A.

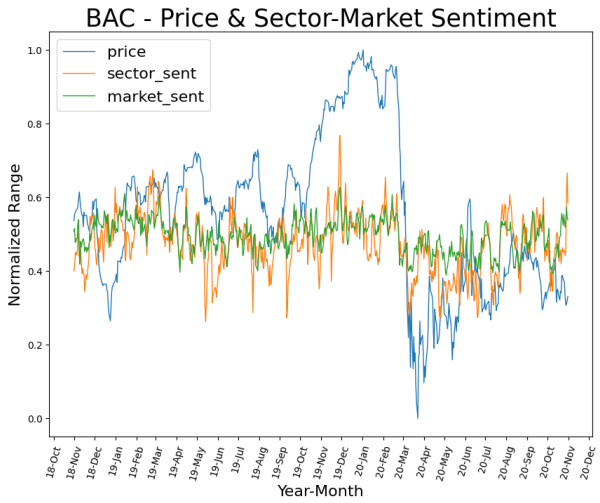

Figure 7a

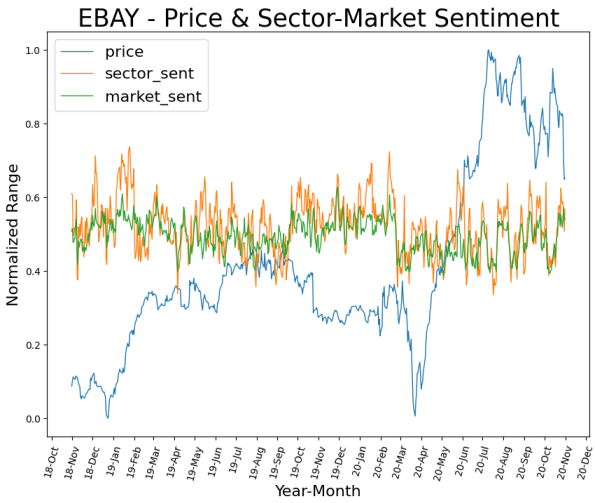

Figure 7b

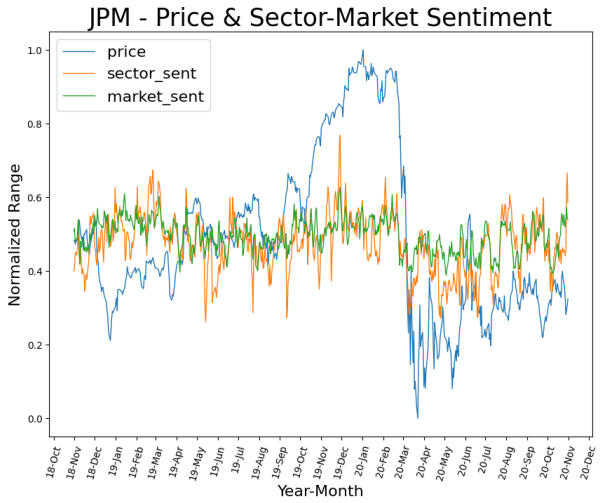

Figure 7c

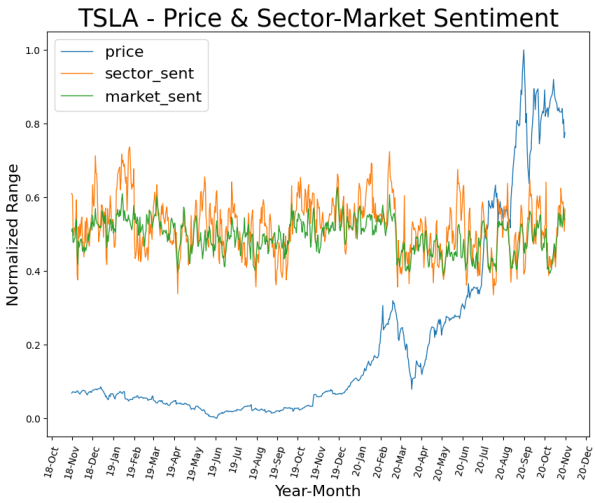

Figure 7d

Figure 7: Stock Price and Market Sector Sentiment

Sentiment Analysis: Stock Price & COVID-19 Sentiment:

Next, the prices of each stock are compared with the COVID-19 news sentiment scores for each day. The COVID-19 news sentiment scores are determined from a list of daily top news headlines, and are constant for each comparison with stock prices. Figures 8a through 8d shown below present these relationships for (4) representative stocks [AMZN, FB, GOOG & VZ], with the remainder of the stocks and their corresponding relations shown in Figure A4 in Appendix A.

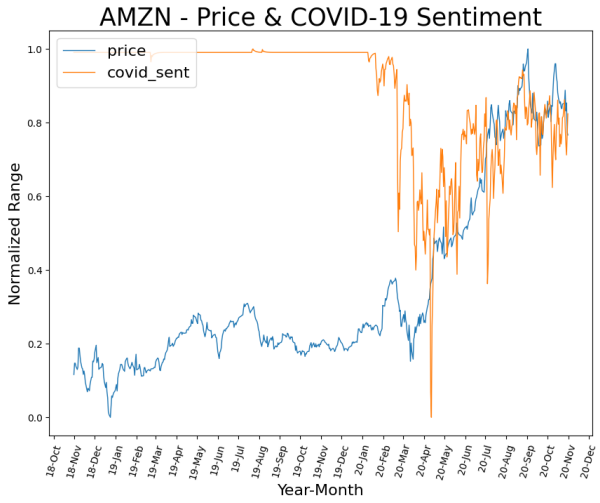

Figure 8a

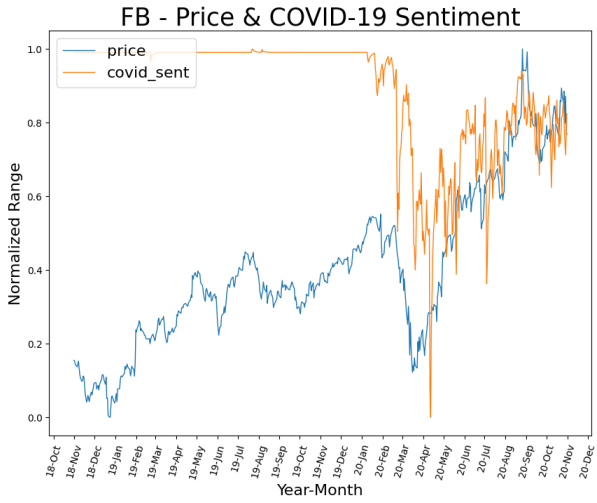

Figure 8b

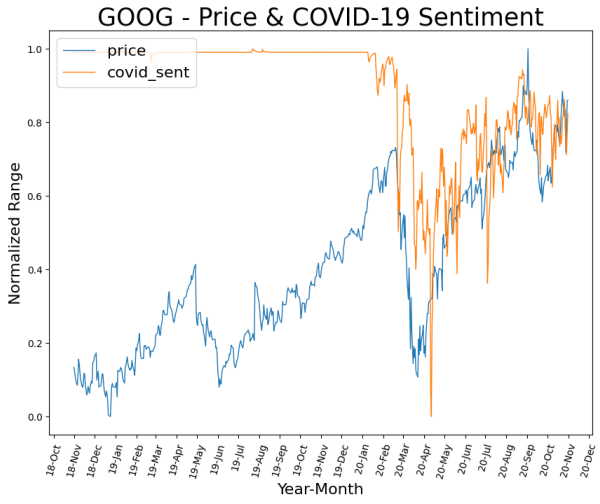

Figure 8c

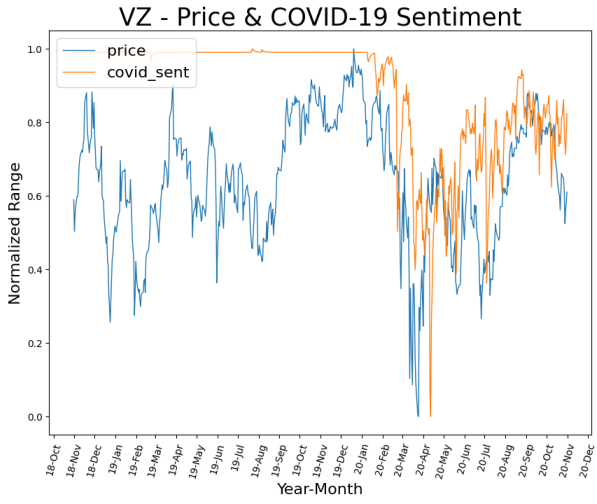

Figure 8d

Figure 8: Stock Price and COVID-19 Sentiment

Sentiment Analysis: Market Sector Sentiment & COVID-19 Sentiment:

For the next comparison, the news sentiment scores for each of the four market sectors [Finance, Information Technology, Consumer Discretionary & Communications] are compared with the COVID-19 news sentiment scores for each day. Figures 9a through 9d shown below present these relationships for the four market sectors.

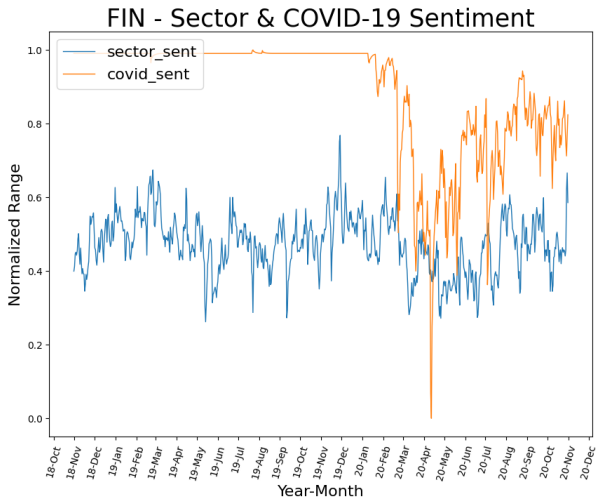

Figure 9a

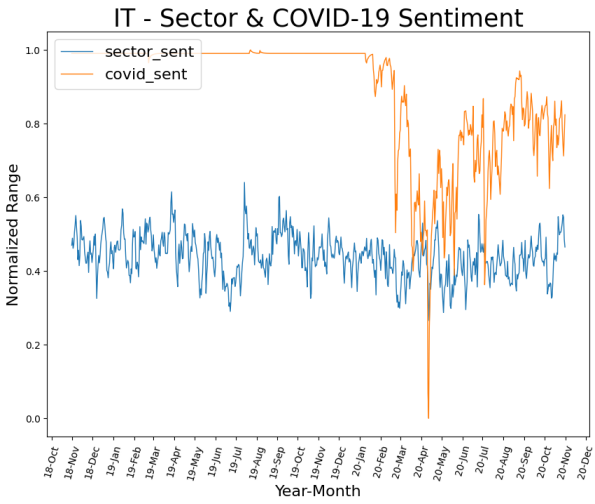

Figure 9b

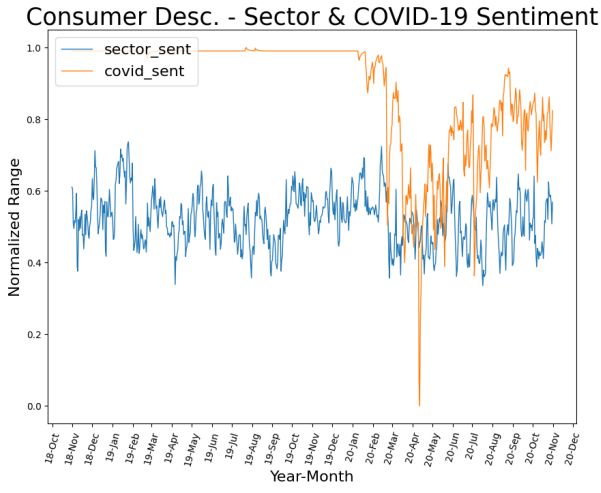

Figure 9c

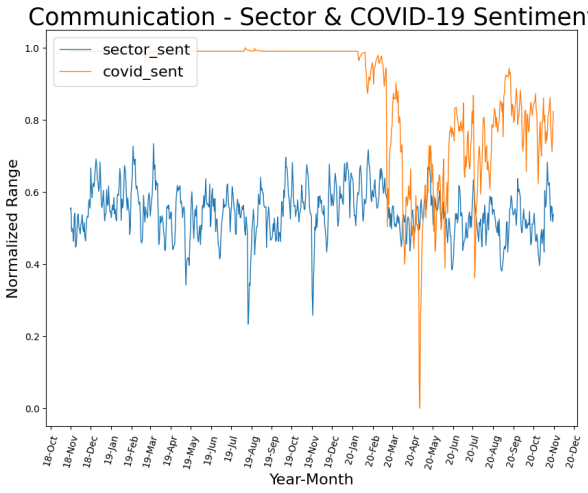

Figure 9d

Figure 9: Market Sectors and COVID-19 Sentiment

Sentiment Analysis: Overall Market Sentiment & COVID-19 Sentiment:

In this last analysis, the news sentiment scores for the overall market, averaged across the (16) stocks considered in this paper, are compared with the COVID-19 news sentiment scores for each day. The relationship is presented below in Figure 10.

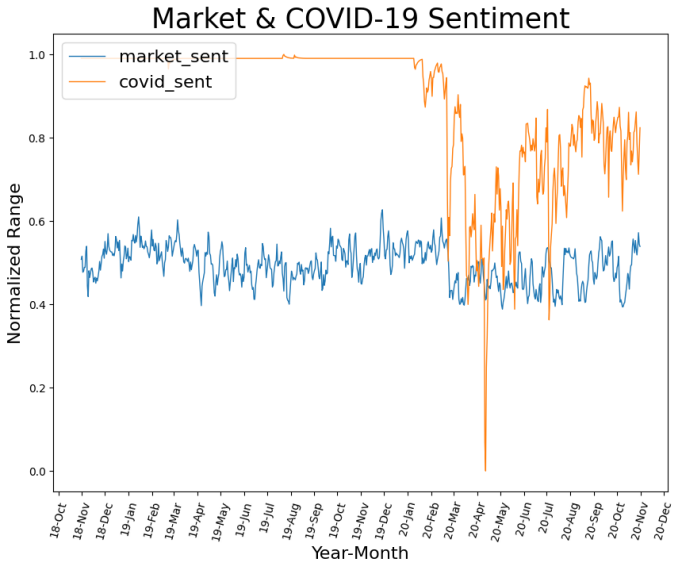

Figure 10 - Market Price and COVID-19 Sentiment

Stock Prediction Model

Three separate sequential neural network based prediction models are developed using the information discussed in the previous section. The input parameters and architecture of the models are reviewed in depth under the Neural Network section of Methodology in this paper. The performance of each neural network model is presented below in Table 4:

	Neural Network (A)	Neural Network (B)	Neural Network (C)
BAC	48.65%	58.25%	52.67%
JPM	47.97%	63.91%	57.43%
BRK.A	50.00%	56.90%	50.68%
GS	50.68%	54.05%	51.35%
AAPL	48.65%	49.25%	47.97%

	Neural Network (A)	Neural Network (B)	Neural Network (C)
CSCO	52.03%	56.30%	53.20%
INTC	55.41%	57.40%	55.80%
ORCL	46.62%	50.68%	48.65%
TSLA	56.08%	59.85%	53.38%
AMZN	51.35%	52.78%	49.32%
EBAY	49.32%	48.65%	51.35%
MCD	54.05%	57.94%	49.32%
FB	49.10%	55.41%	52.70%
GOOG	53.38%	56.90%	52.03%
CMCSA	45.95%	55.66%	53.38%
VZ	51.35%	56.08%	54.01%

Table 4 - Neural Network (A, B & C) Performance Comparison For Stocks

Table 5 and 6 shown below present the performance of each neural network model on the different market sectors and overall market, respectively. The values of Table 5 are calculated by averaging the performance of the models for each stock in the market sector, while the values in Table 6 show the average performance of the models across all (16) stocks.

	Neural Network (A)	Neural Network (B)	Neural Network (C)
Finance	49.33%	58.28%	53.03%
Information Tech.	50.68%	53.41%	51.41%
Consumer Disc.	52.70%	54.81%	50.84%
Communication	49.95%	56.01%	53.03%

Table 5 - Neural Network (A, B & C) Performance Comparison For Market Sectors

	Neural Network (A)	Neural Network (B)	Neural Network (C)
All Market Sectors	50.67%	55.63%	52.08%

Table 6 - Neural Network (A, B & C) Performance Comparison For All Market Sectors

Discussion

As indicated in some of the prior works presented in the Related Works section of this paper, and evident by the technical analysis presented in the Results section, the price of a stock and its corresponding market volume are not strongly correlated. Using 10-day segments of data, and calculating the Spearman's Correlation for each segment, the average correlation between stock prices and market volumes across all (16) stocks is +0.07. Spearman's Correlation, where -1.00 indicates a strong negative correlation and +1.00 indicates a strong positive correlation, is used as an analysis metric to account for non-linear correlations in data. It is important to note that not all stocks had poor correlation between price and market volume. Figure 5a and 5c show stronger correlation between the two parameters for AAPL and TSLA, with correlation coefficients of -0.10 and -0.16, respectively.

Unlike market volume, stock prices demonstrate a stronger correlation when compared with news sentiment scores determined from news related to the stock. Using the same procedure to determine the Spearman's Correlation outlined above, the average correlation between stock prices and stock sentiment scores across all (16) stocks is +0.12. The correlation varies between each stock, with the strongest being for EBAY at +0.18, and the weakest being for VZ at -0.01 The correlation calculation is dependent on availability of news data regarding the stock. For example for JPM, as shown in Figure 6c, there were few news headlines regarding the company between October 31, 2018 and October 31, 2020, leading to a poor correlation score of +0.05.

Next, the Results section presents the relationship between stock prices and news sentiment scores averaged across the different market sectors. The average Spearman's Correlation between stock prices and market sector sentiment scores across all (16) stocks is +0.11. The results show a poorer correlation between stock price and market sector news sentiment scores, compared to between stock prices and stock sentiment scores. This is likely due to the presence of a variety of companies within the same market sector, thus leading to a balancing affect on the overall news sentiment for the market sector. For example BAC and JPM, both part of the same market sector (Finance), show different correlations of +0.04 and +0.11.

The next portion of the Results section presents the relationship between stock prices and news sentiment regarding COVID-19 retrieved from daily top headlines. Correlations for

COVID-19 news sentiments are only calculated after January 2020 due to a lack of relevant COVID-19 news prior to this time. The average Spearman's Correlation between stock prices and COVID-19 sentiment scores across all (16) stocks is +0.17. Similar to the correlation between stock prices and stock news sentiment scores, the relationship between stock prices and COVID-19 sentiment scores varies significantly between stocks. For example, JPM demonstrates the strongest relationship with a correlation of +0.31, while ORCL demonstrates the weakest correlation of +0.08. This variation can be attributed to the diverse impact of COVID-19 on different companies and market sectors. For example between January 2020 and April 2020, AMZN demonstrates a correlation of -0.12, suggesting an inverse relationship between AMZN's stock price and sentiment towards the virus. This may be related to the considerable spike in online retailing caused by the onset of global quarantine measurements. MCD on the other hand demonstrates a positive correlation of +0.24 within the same time frame, suggesting the adverse impact COVID-19 had on the Consumer Discretionary market sector due to quarantines.

The remaining two portions of the Results section present the relationship of market sector sentiments and overall market sentiments with COVID-19 sentiments. The correlation coefficients for COVID-19 sentiment and each of the four market sectors i.e. Finance, Information Technology, Consumer Discretionary and Communication are +0.10, +0.02, -0.06 and -0.11, respectively. The correlation between COVID-19 sentiment and the overall market sentiment is -0.06. The correlations between market sector sentiments and COVID-19 sentiment is relatively weak, since the presence of a variety of companies within the same market sector likely leads to a balancing affect on the overall news sentiment for the market sector. For example, as discussed earlier, the correlation between AMZN's news sentiment and COVID-19 new sentiment has an inverse relationship in the beginning of 2020, while MCD has a positive correlation.

The three stock prediction approaches outlined in the second part of the results section indicate that the consideration of news sentiment plays a impactful role in the accuracy of the prediction model. Model A, which only considers the stock price (closing, high & low) and market volume demonstrated an average prediction accuracy of 50.67% across all (16) stocks, with a maximum prediction accuracy for TSLA at 56.08%. Model C, which considers all the

parameters of Model A in addition to sentiment scores of stocks, market sectors and overall market provided an average prediction accuracy of 52.08%, with a maximum prediction accuracy for JPM at 57.43%. Model B, which considers all the parameters of Model C in addition to a COVID-19 news sentiment score performed the best with an average prediction accuracy of 55.63%, where JPM demonstrated the highest prediction accuracy of 63.91%. Considering news sentiment scores of stocks, market sectors and overall market yielded on average an improved prediction accuracy of 1.41%, while considering a COVID-19 news sentiment score provided an additional accuracy of 3.55% (4.96% total). Based on these results, it is clear that including sentiment analysis as part of a stock prediction model yields on average more reliable and accurate results.

Conclusion

Overview

This paper has reviewed a stock prediction model using sentiment analysis from news articles. In particular, the paper analyzed the impact of the global COVID-19 pandemic on stock prices, and the effects of public sentiment on the financial market. The paper examined the price, market volume and news sentiment for (16) different representative stocks spread across (4) market sectors between October 2018 and October 2020. The stock prediction model was developed using a neural network machine learning architecture and compared against the real market values. Along with the price, market volume of the stocks, four additional parameters were analyzed: (1) a three-day average news sentiment for the stock (2) a three-day average news sentiment for the market sector (3) a three-day average news sentiment for all market sectors and (4) a three-day average news sentiment of COVID-19. Three separate neural network models were developed: (A) only stock parameters as input (B) stock parameters in addition to parameters 1, 2, 3 and 4 mentioned above and (C) stock parameters in addition to parameters 1, 2 and 3 mentioned above.

Based on the results presented in the paper, it is evident that sentiment analysis improves the accuracy of the stock prediction model. The overall average accuracy of the model using only stock price data as input was 50.67%. The overall average accuracy of the model using stock

price data in addition to parameters 1, 2, 3 and 4 mentioned in the above paragraph was 55.63%. Finally, the overall average accuracy of the model using stock price data in addition to parameters 1, 2 and 3 mentioned in the above paragraph was 52.08%. From the (16) stocks analyzed, JPM and CMCSA showed the highest and lowest accuracies in the prediction model, respectively. Between the 4 market sectors analyzed, Finance and Consumer Discretionary showed the highest and lowest accuracies, respectively.

On average, using sentiment analysis information yields an 1.41% accuracy improvement over a purely technical analysis approach. Including sentiment analysis information for news explicit to COVID-19 yielded an additional 3.55% (4.96% total) improvement in accuracy. Therefore, financial market prediction models during future global pandemics can utilize a sentiment analysis approach to strengthen the quality of market forecast. This can in turn help minimize the the loss of financial capital due to uncertainty in the market, and help stabilize investor confidence. Similar models can be extended to other situations involving general market unrest such as during times of war and political unrest. An improved predictability of the financial market can also result in lower unemployment rates and faster rebound periods from economic downturns.

Future Works

In this paper, only sentiment analysis from news articles was considered as part of the stock prediction model. However to get a more holistic overview of public sentiment, future works can consider user generated social media posts. While news articles give insight into the most important and discussed topic for a given time frame, social media feeds provide a true window into the public's sentiment on those topics. In addition, the study of different groups on social media can diversify the input to the prediction model and thus increase the model's overall resiliency to boundary conditions.

In addition, more advanced sentiment analysis techniques can be implemented in future works to improve the accuracy and reliability of the sentiment analysis process. Bidirectional Encoder Representations from Transformers (BERT), developed by Google in 2018, provides improved NLP capabilities over traditional models. BERT incorporates the capability to process

all tokens within the input text at once, and provides contextual sentiment scores based on the location and placement of words within the text. Topic Modeling, another advanced NLP technique, can be used as a text pre-processor prior to the sentiment analysis phase to improve contextual categorization of textual data.

The neural network used in this project comprised of a simple 4 layer architecture and standard sequential back propagation mechanism. For future works, more intricate learning models can be employed to improve the accuracy of the prediction models. For example, General Fuzzy Neural Networks (GFNN) can be used to process the sentiment analysis prior to feeding into the main neural architecture for better NLP performance. Similarly, the neural architecture can be tuned to improve resiliency and reduce overfitting. Finally, distributed systems such as HDFS architecture along with MapReduce algorithms can be implemented to improve run-time and scale up the prediction model to account for much larger data sets.

In order to make the prediction model shown in this project industry ready, it needs to be capable of pulling news articles dynamically from the internet and providing real-time feedback on stock prices. In future works, the prediction model presented in this paper can be hosted on cloud computing services such as AWS with an online user interface in order to provide dynamic feedback capabilities. However it is important to note that such systems often involve a running cost for both the cloud service usage and real-time news collection, which were out of scope for the work presented in this paper.

References

Agrawal, S., Jindal, M., Pillai, G. N., "Momentum Analysis based Stock Market Prediction using Adaptive Neuro-Fuzzy Inference System (ANFIS)", Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, Vol. 1, March 17 -19, 2010.

Annett, Michelle, and Grzegorz Kondrak. "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs." Department of Computer Science, University of Alberta.

A.K. Davis, J.M. Piger, L.M. Sedor, Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases, Federal Reserve Bank of St. Louis Working Paper Series, 2006.

Atsalakis, G. S., Valavanis, K. P., "Surveying stock market forecasting techniques – Part II: Soft computing methods", Expert Systems with Applications, Vol. 36, pp. 5932-5941, 2009.

Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8 (2011)

Dickinson, Brian, and Wei Hu. "Sentiment analysis of investor opinions on twitter." Social Networking 4.03 (2015): 62.

Ganatr, A., Kosta, Y. P., "Spiking Back Propagation Multilayer Neural Network Design for Predicting Unpredictable Stock Market Prices with Time Series Analysis", International Journal of Computer Theory and Engineering, Vol. 2, No. 6, 1793-8201, 2010.

Hatzivassiloglou, V., Wiebe, J.: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In: Proceedings of the 18th International Conference on Computational Linguistics, New Brunswick, NJ (2000)

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Kamps, J., Marx, M., Mokken, R.J.: Using WordNet to Measure Semantic Orientation of Adjectives. In: LREC 2004, vol. IV, pp. 1115–1118 (2004)

Kennedy, A., Inkpen, D.: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. Computational Intelligence, 110–125 (2006)

L. Zhang, Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, pp. 130, 2013.

P.C. Tetlock, Giving content to investor sentiment: the role of media in the stock market, J. Finance 62 (2007) 1139–1168.

P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, J. Finance 63 (2008) 1437–1467.

Shah, Vatsal H. "Machine Learning Techniques for Stock Prediction." http://www.ijarcs.info, 2007, International Journal of Advanced Research in Computer Science.

Sureshkumar, K. K., Elango, N. M., "An Efficient Approach to Forecast Indian Stock Market Price and their Performance Analysis", International Journal of Computer Application, Vol. 34, No.5, 2011.

Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL, Philadelphia, PA, July 2002, pp. 417–424 (2002)

Turney, P.D., Littman, M.L.: Measuring Praise and Criticism: Inference of Se- mantic Orientation from Association. ACM Transactions on Information Systems, 315–346 (2003)

V. Niederhoffer, The analysis of world events and stock prices, J. Business 44 (1971) 193–219.

Appendix A - Charts

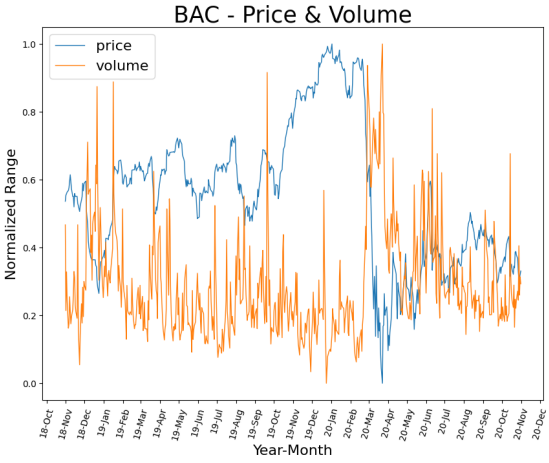

Figure A1-a

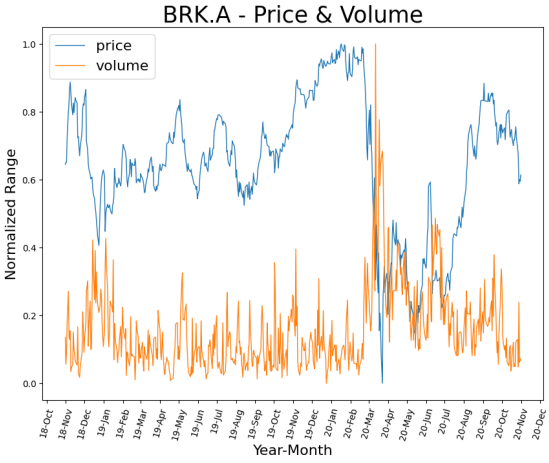

Figure A1-b

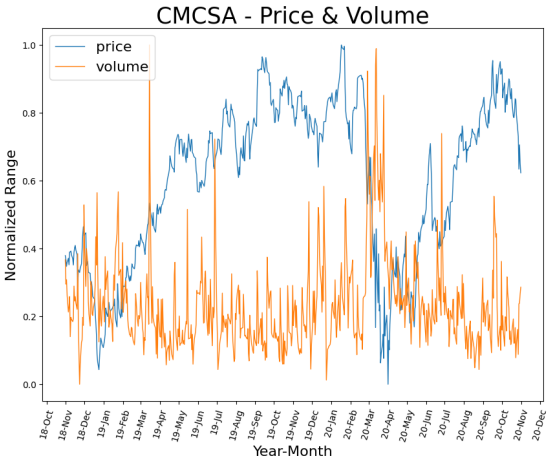

Figure A1-c

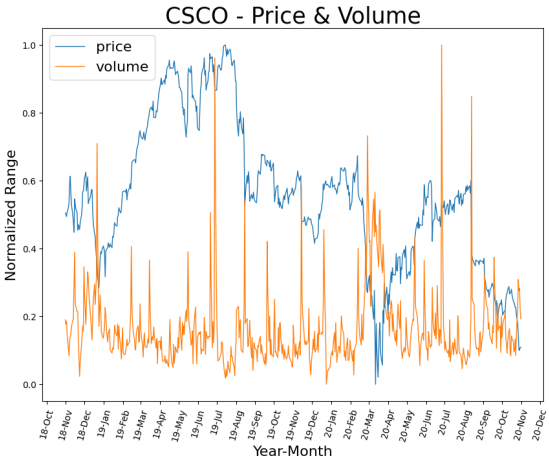

Figure A1-d

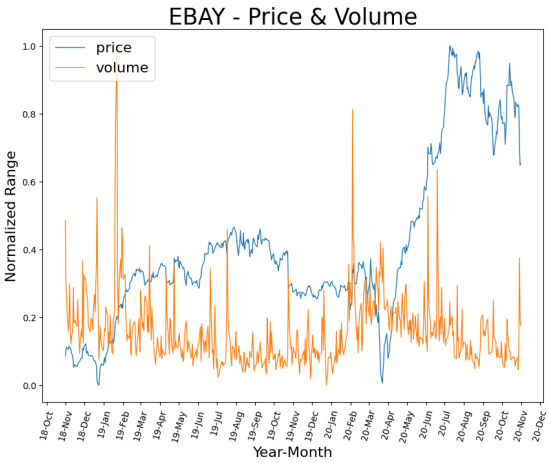

Figure A1-e

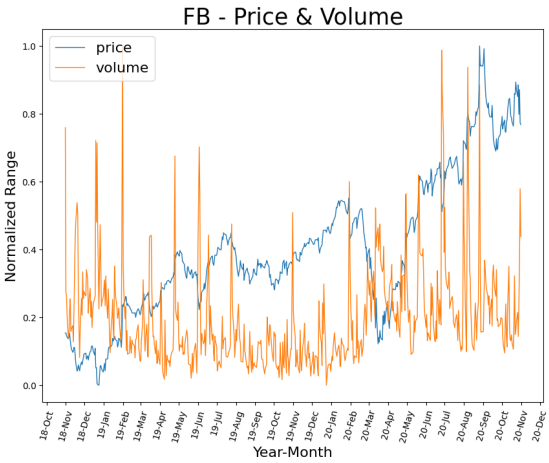

Figure A1-f

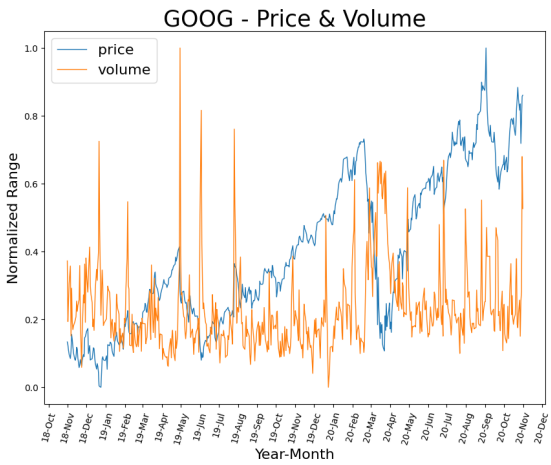

Figure A1-g

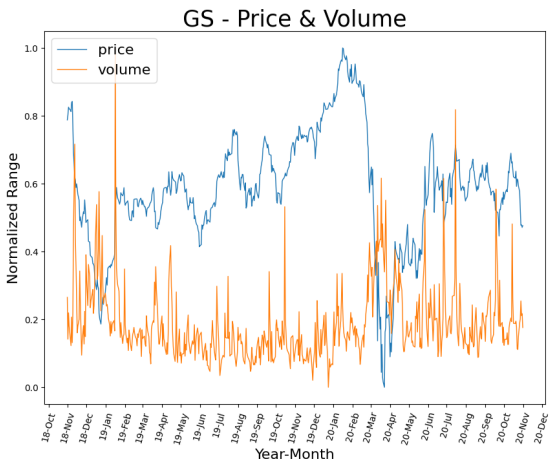

Figure A1-h

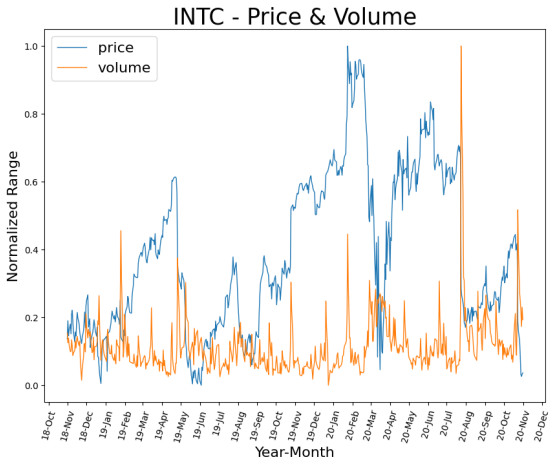

Figure A1-i

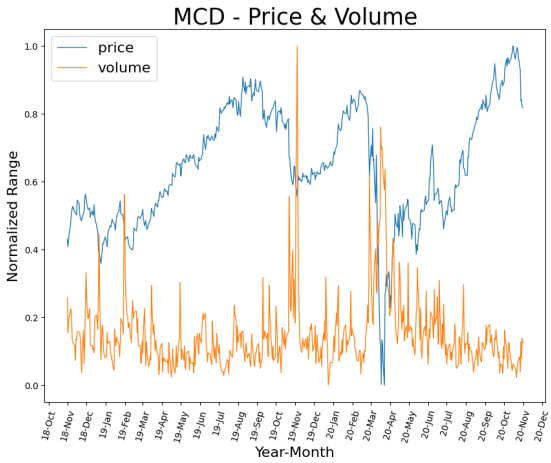

Figure A1-j

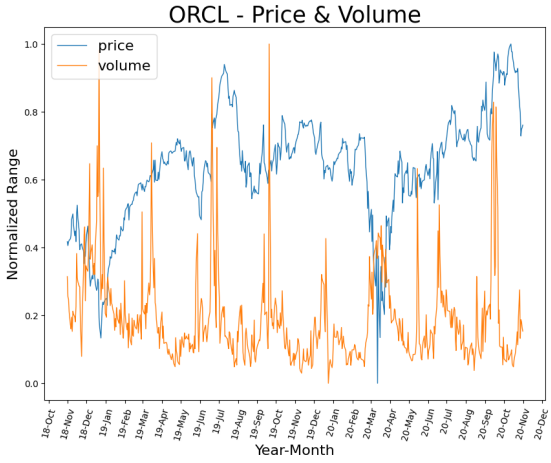

Figure A1-k

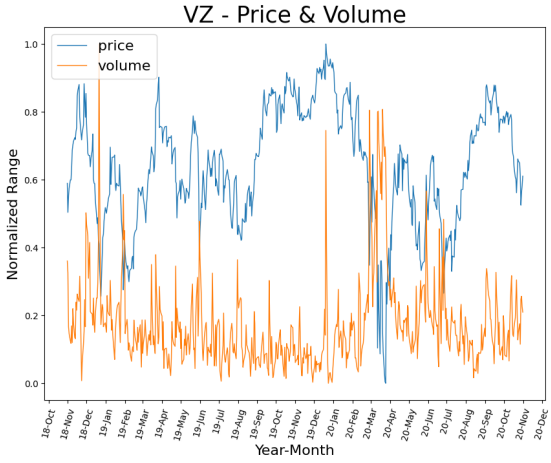

Figure A1-l

Figure A1: Stock Price and Stock Volume

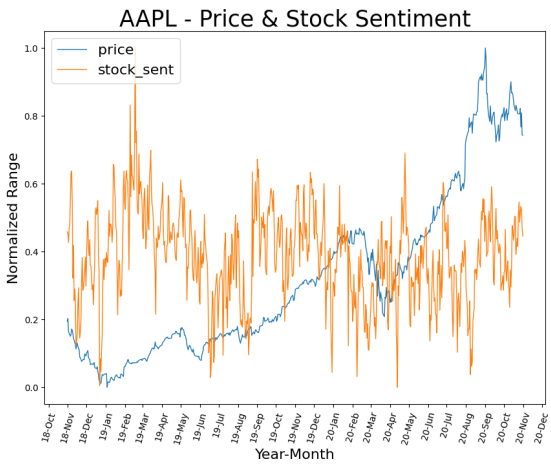

Figure A2-a

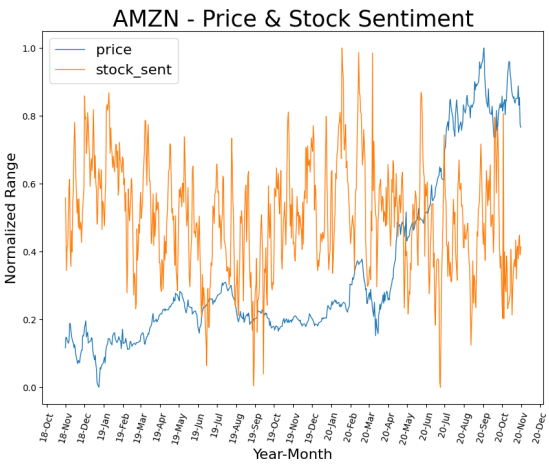

Figure A2-b

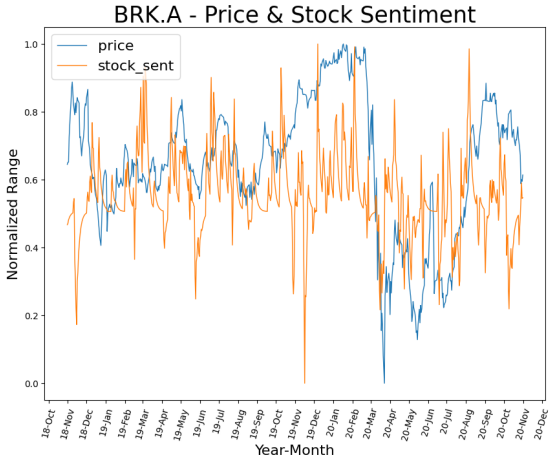

Figure A2-c

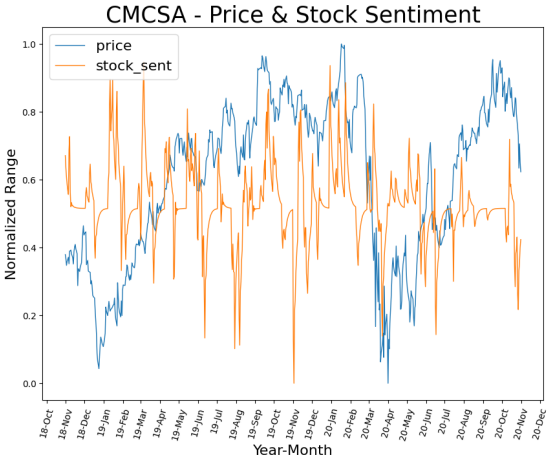

Figure A2-d

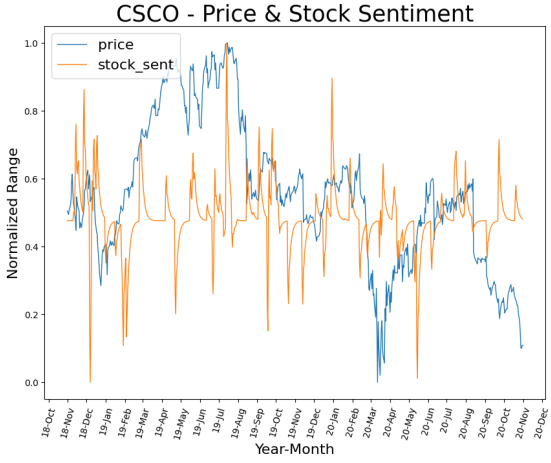

Figure A2-e

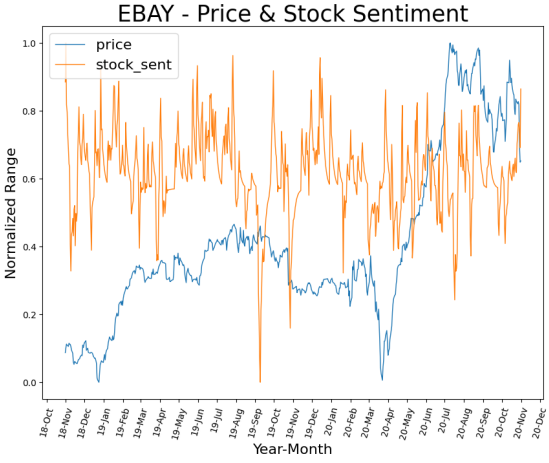

Figure A2-f

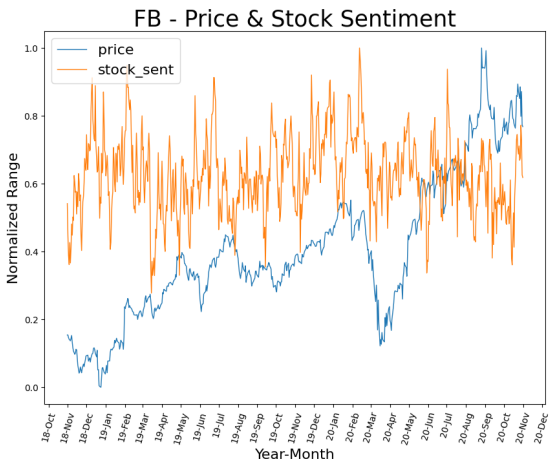

Figure A2-g

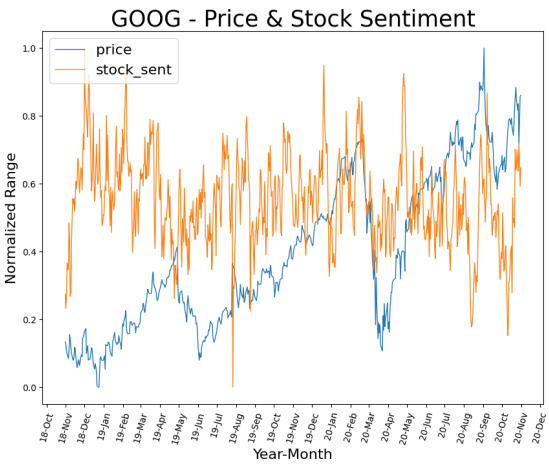

Figure A2-h

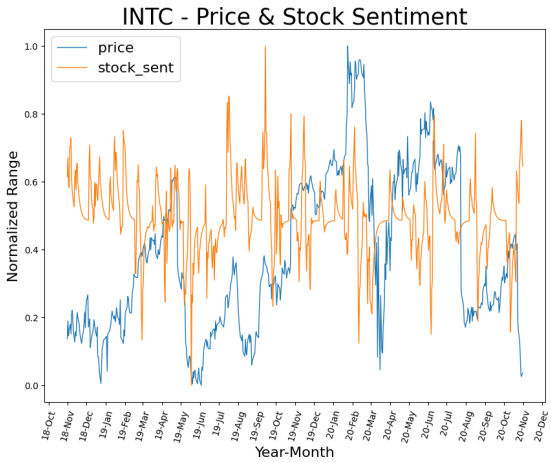

Figure A2-i

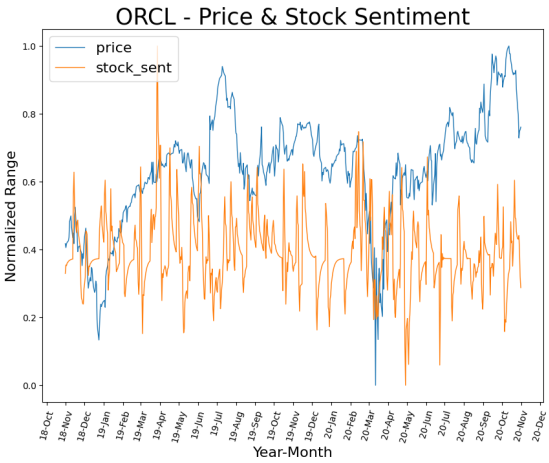

Figure A2-j

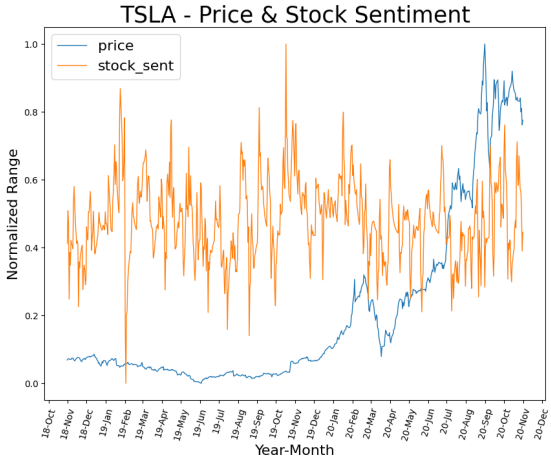

Figure A2-k

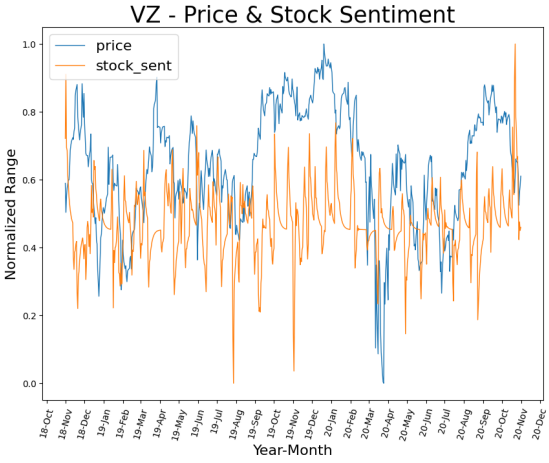

Figure A2-l

Figure A2: Stock Price and Stock Sentiment

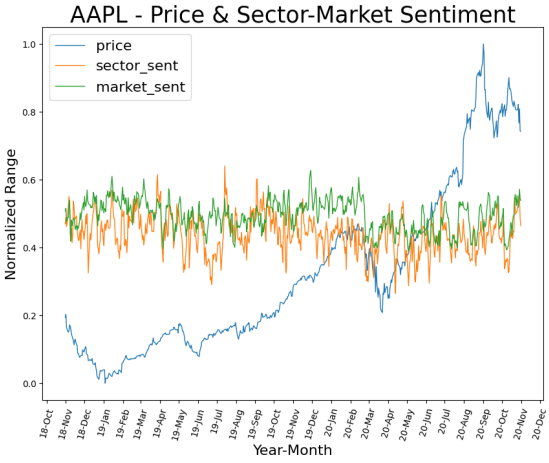

Figure A3-a

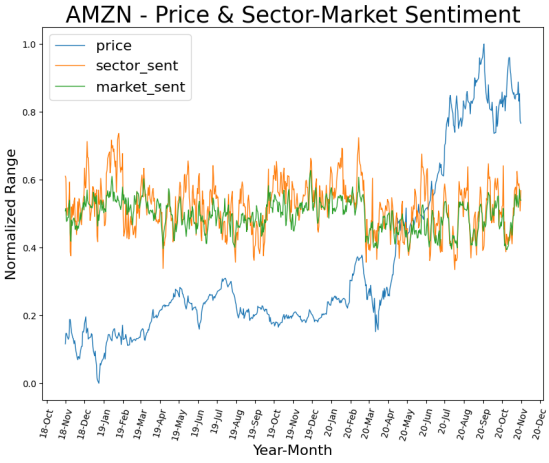

Figure A3-b

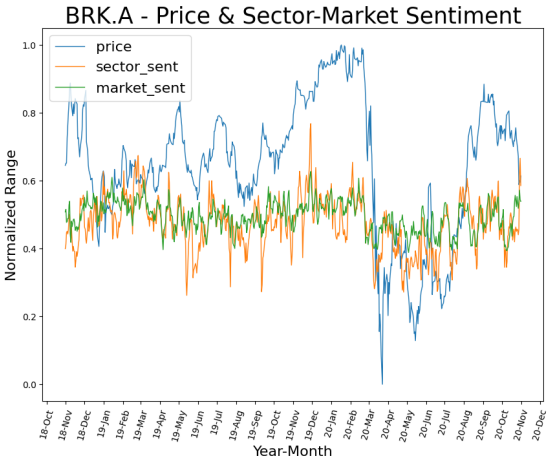

Figure A3-c

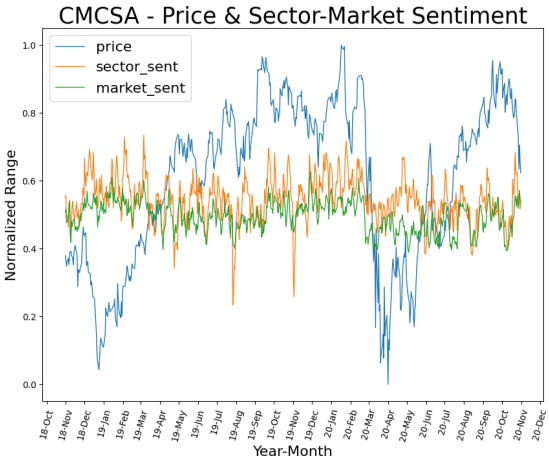

Figure A3-d

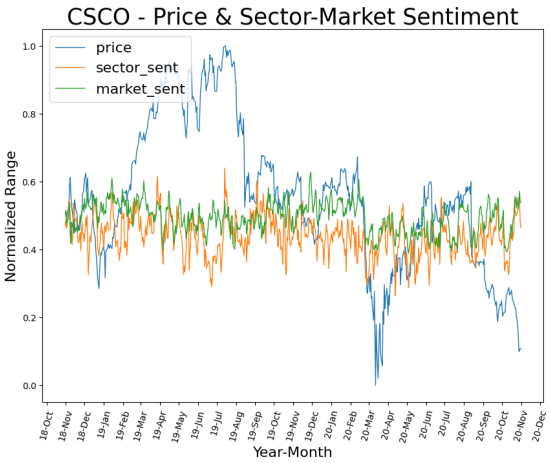

Figure A3-e

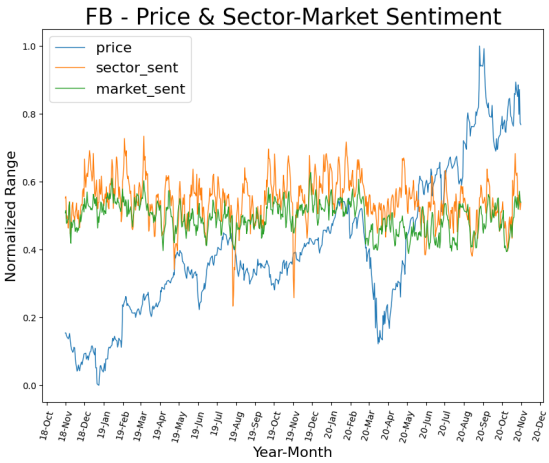

Figure A3-f

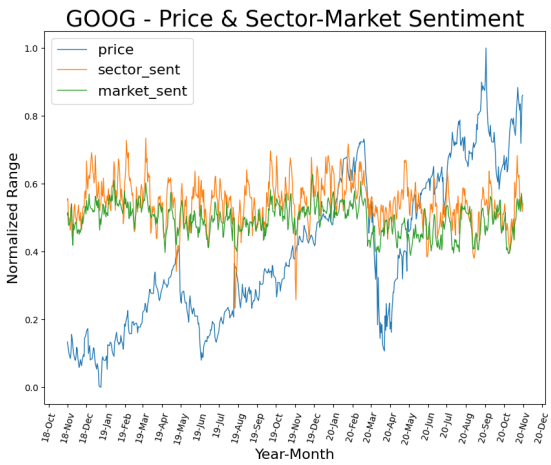

Figure A3-g

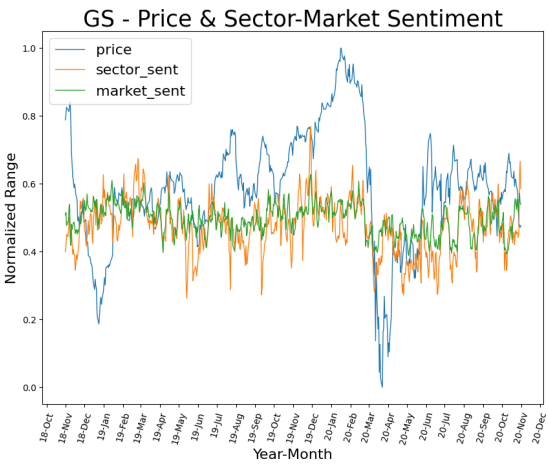

Figure A3-h

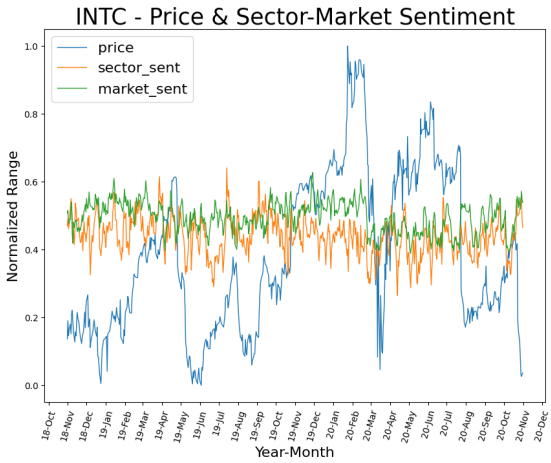

Figure A3-i

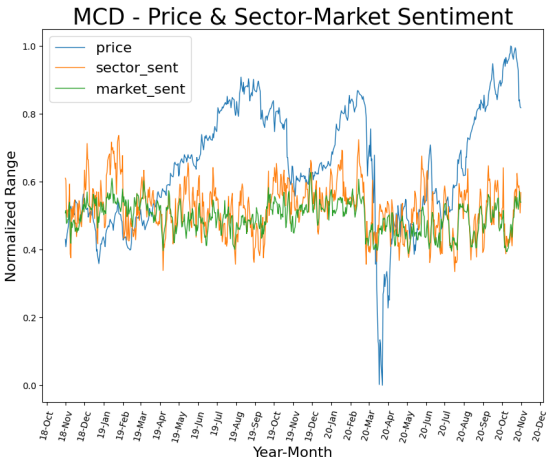

Figure A3-j

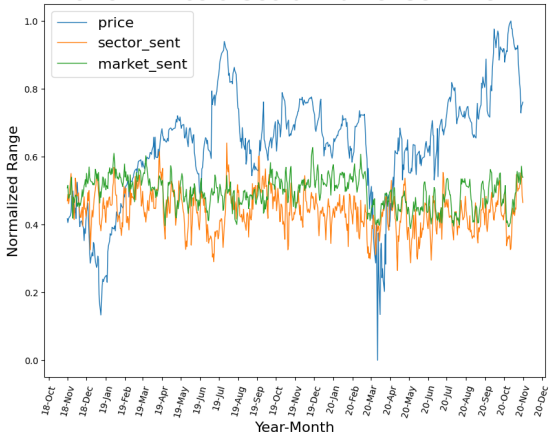

Figure A3-k

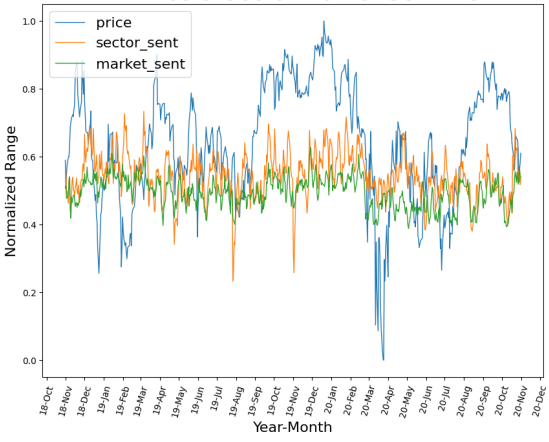

Figure A3-l

Figure A3: Stock Price and Market Sector Sentiment

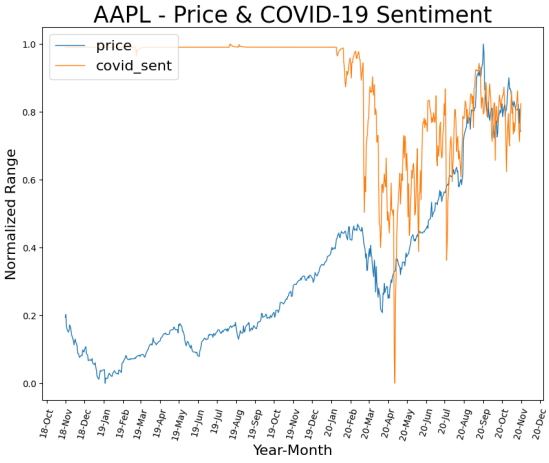

Figure A4-a

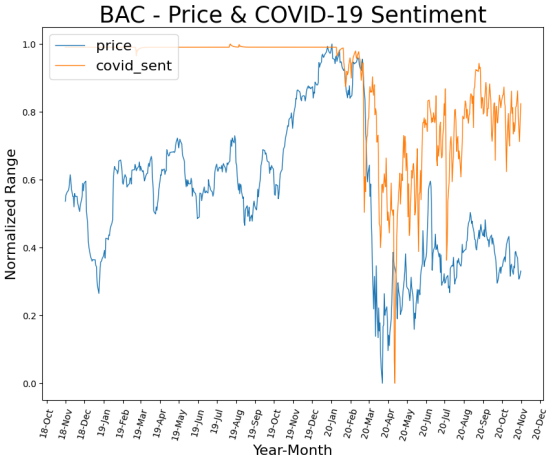

Figure A4-b

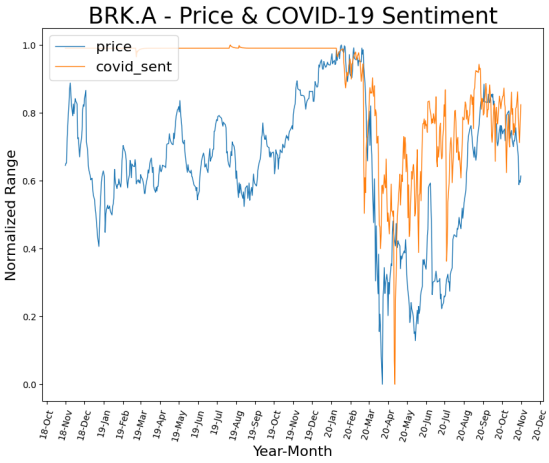

Figure A4-c

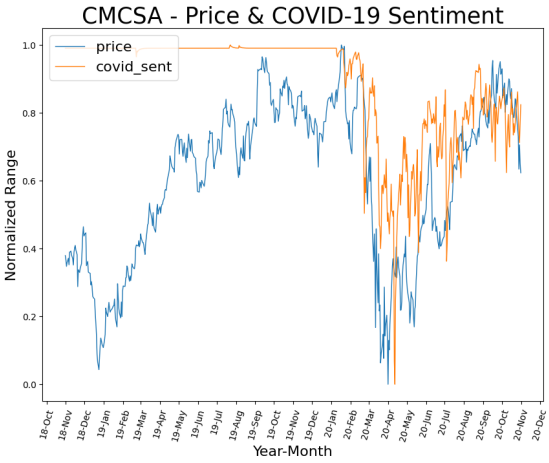

Figure A4-d

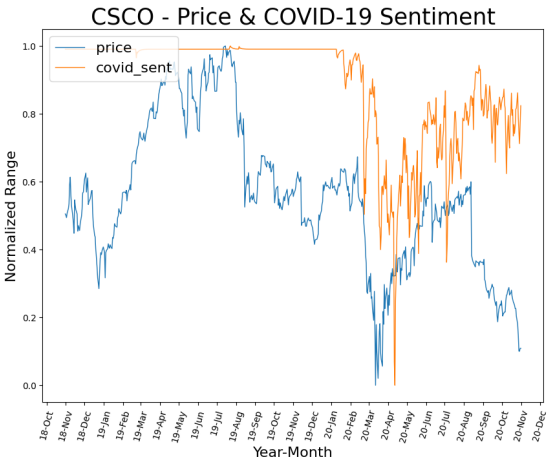

Figure A4-e

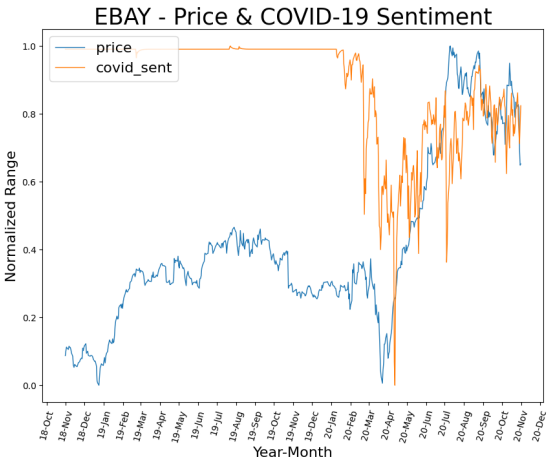

Figure A4-f

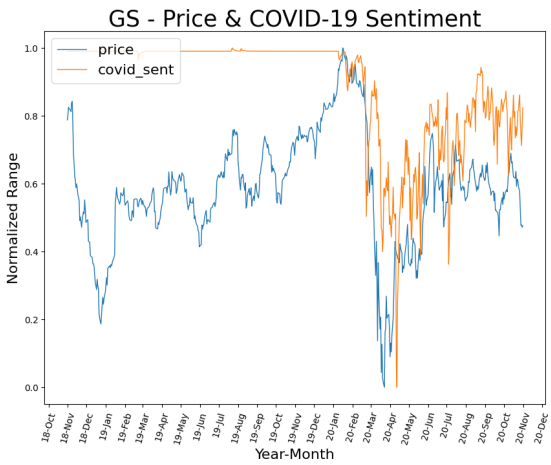

Figure A4-g

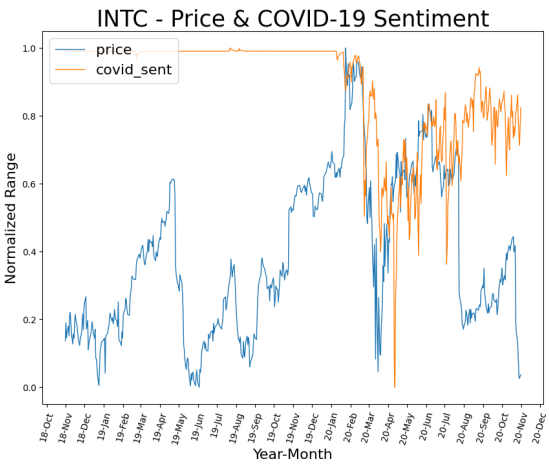

Figure A4-h

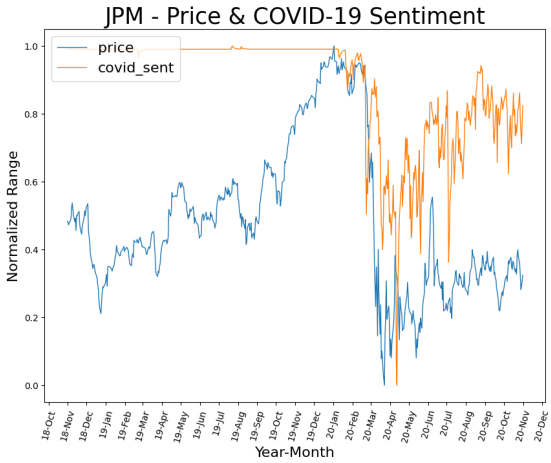

Figure A4-i

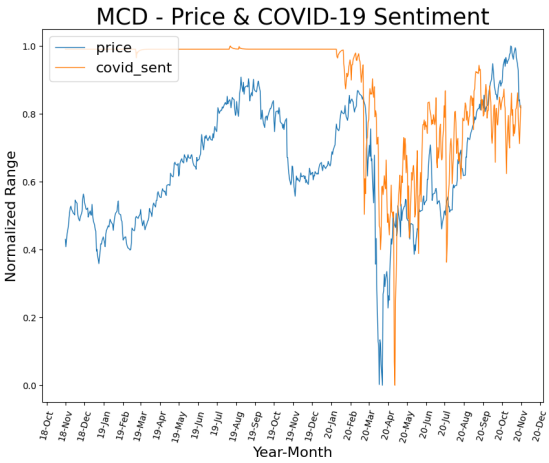

Figure A4-j

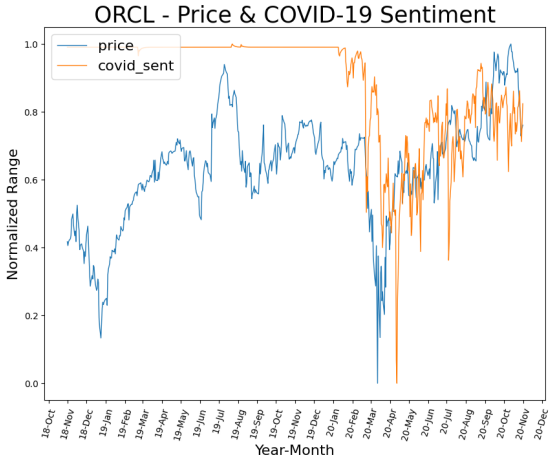

Figure A4-k

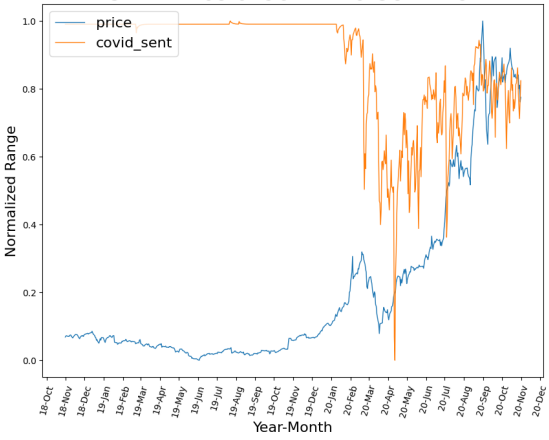

Figure A4-l

Figure A4: Stock Price and COVID-19 Sentiment

Appendix B - Code

The following python 3.8.0 code files can be accessed at:

https://github.com/nablul/COVID-Stock-Sentiment

- README.md
- config.py
- get_stock_news.py
- get_top_news.py
- process_stock_data.py
- sentiment_analysis.py
- process_data.py
- neural_network_train.py
- neural_network_test.py
- correlation.py
- plotter.py