Spring 6-1-2021

Stock Prediction using Natural Language Processing Sentiment Analysis on News Headlines During COVID-19

Mina Ibrahim
m.youssef@aucegypt.edu

The American University in Cairo

School of Business

Stock Prediction using Natural Language Processing
Sentiment Analysis on News Headlines During COVID-19

A Thesis Submitted to

Department of Management

in partial fulfillment of the requirements for

the degree of Master of Science in Finance

by

Mina Ibrahim Gerges Ibrahim

under the supervision of Dr. Medhat Hassanein

January/ 2021

Abstract

Stock Prediction using Natural Language Processing Sentiment analysis on News During Covid-19

Stock prediction based on NLP sentiment analysis is one of the most researched topics due to the revenues they generate for investors. Researchers have used various tools to achieve this, especially fundamental and technical analysis based on historical data helped to achieve this target. Due to the technological advancement and abundance of data, the introduction of machine learning tools accelerated that approach. However, as the public mood affects the stock market, the need for another analysis emerged. Natural language processing sentiment analysis on data from various sources was able to capture public events and moods. NLP is one of the most effective tools since covering the public moods, and capturing the sentiment is the main driver for stock markets. In this research, NLP sentiment analysis shall be applied to news to predict United States technology stock companies and indices during COVID-19 using a natural language toolkit. The contribution of this is the research is creating a model for predicting the technology companies listed in the United States market during the crisis. The model is achieving over 61% accuracy and could be highly improved by adding other resources of news.

Keywords: Stock Prediction, NLP, Covid-19

Table of Contents

List of Abbreviations

ML: Machine Learning

AI: Artificial Intelligence

NLP: Natural Language Processing

ARMA: Autoregressive Moving Average

ARIMA: Autoregressive Integrated Moving Average

GARCH: Generalized Autoregressive Conditional Heteroskedasticity

SVM: Support Vector Machine Technique

NN: Neural Networks

RNN: Recurrent Neural Network

CNN: Convolution Neural Network

LSTM: Long short-term memory

EMH: Efficient Market Hypothesis

API: Application Programming Interface

NLTK: Natural Language Toolkit

NYT: New York Times

KNN: k-Nearest Neighbors

List of Figures

List of Tables

Chapter 1 Introduction

Predicting stock performance accurately has been one of the most active fields of research in the financial sector due to its high returns in stock market trading and minimizing risks. Stock prediction has been developing and improving over the years, utilizing different concepts as well as techniques. To predict stock, two major streams are used either Fundamental analysis or Technical analysis. By fundamental analysis, the research is done based on the company parameters such as revenues, losses, CEO profile, and other factual bases of the company itself. On the other side, technical analysis is focused on the trading of the stock and its trends. Finally, the decision is being made regardless of the company factors—one of the most traditional directions in the technical analysis. Technical analysis for stock performance has been widely researched using a statistical approach utilizing different theories. Along with technological advancement, many techniques emerged, such as AI, including different ML approaches. On the other side, technology advancement did not only help with tools and techniques but with data and abundance of information. For example, due to technological advancement, researchers were able to collect news from different channels in terms of text and data sets. Also, it increased their ability to manipulate the data and formalize it to test and analyze it. Moreover, the introduction of easier programming languages and the exponential development of computing power of the hardware has accelerated the research in the field dramatically. These have enabled researchers to improve their statistical models in order to outperform the existing ones in terms of stock prediction. However, statistical models have always ignored the effect of the news unless they build up trends or the effect of the news show up on the stock performance. At the same time, ML tools have been introduced. These circumstances have encouraged researchers to do text analysis or NLP and predict its effect on the stock directly. Such motivation was proved by analyzing public moods

through twitter and its correlation with the Dow Jones index (Bollen, Mao, & Zeng, Twitter mood predicts the stock market, 2011). Adding to this, the introduction of different machine learning tools improved the prediction process. For sure, it comes with a few problems that will be discussed in the following lines. These drawbacks have been avoided with different artificial intelligence techniques. Moving forward, researchers were able to improve their predicting models by using different data set sources and different frames of times. Data set sources helped to provide the different frequency of news as well as more reliability as it gives more reliability of the sentiment. Moreover, even the data collected has been perceived from a different perspective in order to evaluate them efficiently in terms of which part has been evaluated, either the headline or body or evaluating both with different weights. However, the previous models are general and focusing on normal time frames. Few are focusing on finding the effect and predicting the performance during unprecedented events such as the epidemic of Covid-19 (Baker, et al., 2020). The significance of having a stock predicting model during a crisis, especially a pandemic, will help to optimize the investment portfolio according to the risks arising**.** In order to study stock prediction models on the US stock market technology companies during covid-19, several areas of predicting models will be investigated in this paper, including statistical-based models, the emerging of ML algorithms and their impact in stock prediction, text-analysis or NLP sentiment analysis models including different techniques, and time frames data sets.

The paper is organized as follows: Chapter II literature review is illustrated, chapter III includes the research Objective, Chapter IV describes the Methodology, Chapter V includes the Data description, Chapter VI will summarize the research results, Chapter VII is the conclusion and model evaluations, and finally Chapter VIII will explain the future work.

Chapter 2 Literature Review

2.1 Stock Prediction using Statistics

Stock prediction models have been an interest to many researchers, especially those with statistical knowledge, as statistics gave tools to find a pattern and predict future performance. The statistical approach is following the technical line since it is analyzing the stock performance in the absence of the fundamental company data. Researchers have adopted many different statistical tools such as ARMA, ARIMA, GARCH, and sometimes combine between different tools. In their research, Rounaghi & Zadeh were able to use the ARMA model to predict both London exchange and S&P 500 for monthly and annual returns and conducted a comparison of the results, which showed the S&P500 model outperformed the London exchange model (Rounaghia & Zadehb, 2016). An ARIMA model was used by Devi, Sundar, and Alli to forecast stock performance for companies at the New York Stock Exchange (Devi, D.Sundar2, & Alli, 2013). From another side, researchers have used the GARCH model to predict the volatility (Awartani & Corradi, 2005) as well as the stock performance (Herwartz, 2017) and others compared using ARCH & GARCH models (Engle, 2001). Researchers have even used different algorithms to mix these tools. For example, the generalized expectation-maximization algorithm used to combine ARMA and GARCH, which resulted in a model the outperformed other conventional ARMA-GARCH models (Tang, Chiu, & Xu, 2003). Though statistics provided tools to predict stock markets, pattern recognition was the main advantage of machine learning algorithms (Shah, Isah, & Zulkernine, 2019).

2.2 Introduction of Machine Learning in Stock Prediction

The advancement of technology has progressed the stock prediction models using different machine learning techniques. Machine learning is defined as "a field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). The machine learning approach has allowed finding different patterns for different time frames and various financial tools. One of these tools that machine learning helped was stock prediction. Machine learning algorithms are divided into two main categories. The first category is supervised learning, which is "providing it with input and matching output patterns" (Kachare, Kharde, & Dongare, 2012). The second model is unsupervised learning, in which systems are required to identify patterns in the absence of previous data (Kachare, Kharde, & Dongare, 2012). Another mode is reinforced learning, in which the system gets rewarded based on the action performed in the specific environment (Kachare, Kharde, & Dongare, 2012). The last mode explored by the paper is backpropagation, which uses supervised learning to calculate the error based on the given input and the corresponding output. The system evaluates its performance and recalculates to minimize the error factor (Kachare, Kharde, & Dongare, 2012). On the other hand, there is a recent mode, which is semi-supervised learning [source][the introduction of semi-supervised learning]. ML algorithms showed high accuracy in pattern recognition that was used in different fields in the financial sector, such as bankruptcy prediction (Odom & Sharda, 1990). Many other fields utilized ML, such as credit scoring (West, 2000), corporate failure prediction for banks (Lin & Chen, 2008) and many other applications, and finally, stock predictions. As one example, Das & Padly, in their research, used backpropagation and SVM for forecasting futures prices of Indian stock and came up that SVM has better accuracy than backpropagation (Das & Padhy, 2012). Another ML algorithm used was NN for stock prediction, which achieved almost 90% accuracy on short term

predictions of 10 days (Schoneburg, 1990). Also, LSTM, one of the most used machine learning techniques, was utilized for stock prediction along with CNN and RNN and concluded that CNN is the best technique used along with sliding window (Selvin, R, Menon, K.P, & Gopalakrishnan, 2017). Some researchers moved with LSTM to Attention-based LSTM (At-LSTM) to make a prediction on directional changes for both S&P 500 index and individual companies' stock price. The model involves of RNN to analyze the news text (Liu, 2018). However, machine learning tools helped in the predictions of stock markets depending on the closing price of the stock except those for high-frequency trading. Due to this fact, it lacked to capture the mid-session effect or the breaking news events. For this reason, researchers were motivated to capture this effect by analyzing the text in news and microblogs to capture the mid-session impact drivers. This has led the researcher to take machine learning tools for further step to predict stocks performance by analyzing the news, which impacts the market performance. This has created a new line of research, which is a stock prediction based on natural language processing (NLP).

2.3 **Role of NLP in Stock Prediction during Covid-19**

NLP has enabled researchers in many fields for better prediction, especially taking one step further in predicting stock performance by analyzing the news that drives the market trends, especially in unprecedented times such as during Covid-19. The impact of Covid-19 and the policy measures taken to control have affected the stock market more than even the Spanish flu (Baker, et al., 2020). Moreover, the impact of Covid-19 was in different time frames and scales on countries based on when it spread and how lethal it was (Costola, Iacopini, & Santagiustina, 2020). Moreover, some studies found out there is a correlation between the number of Covid-19 infections and market and news variables (Mamaysky, 2020). Due to the data abundance and technology advancement, NLP sentiment analysis has been widely used along with the emerging of digital

news and microblogging platforms such as Twitter and Facebook. Researchers have used NLP sentiment analysis in many fields, such as predicting movie success (Asur & Huberman, 2010). It has been used to predict general mood and, as a result, reflect on the economy (Bollen, Mao, & Xiao-Jun, Twitter mood predicts the stock market, 2011). Moreover, many other researchers have used untraditional methods to classify the sentiment, either positive or negative such as using emoji (Eisner, Rocktaschel, Augenstein, Bosnjak, & Riedel, 2016). For the many applications it has been used, the emerging of NLP implementation in the financial sector increased in recent years.

2.4 Importance of NLP in financial research

According to EMH, stock prices reflect the information available to investors (Fama, 1970). Based on this theory, stock predictions are a reflection of the news and the moods of the public. Understanding the news and translating them into sentiment would help in stock predictions. For this, using NLP in the financial sector and stock predictions specifically has been developing drastically lately using different machine learning algorithms.

2.5 Methods of Sentiment Analysis

Due to the massive usage of NLP sentiment analysis applications, researchers have started to use it heavily in stock prediction in different approaches. Different methods and approaches were used to understand the text as well as predict the behavior of the interested in the field. The models are being updated and supported with different algorithms to maximize the accuracy of the prediction. Some researchers used Machine learning approaches by training classifiers using annotated data, following this by running the target data and predicting the stock trend. Even training the classifiers took different forms through the earlier mentioned annotated data. By annotated data, it is meant it was manually annotated positive or negative. Another form is using trained libraries such as NLTK in python. The third option is using lexicon-based. By lexicon-

based, it means giving different weights for different words or text used (Oliveira a, Cortez a, & Areal, 2017). Even in the lexicon-based, there were several types of the lexicon used, for example, the generic lexicon (Deng, Mitsubuchi, Shioda, & Shimada, 2011), financial lexicon (Garcia, 2013), and microblog financial lexicon (Oliveira a, Cortez a, & Areal, 2017).

2.6 **Applying NLP sentiment Analysis on Microblogs VS News**

Stock prediction based on NLP Sentiment analysis on the text was made in two main streams: the first is news text and the second stream is the text generated on microblogs. There is a correlation between microblogs in terms of volume or sentiment and stock market movement (Ruiz, Hristidis, Castillo, Gionis, & Jaimes, 2012). Some researchers preferred using microblogs text to predict stock as they capture public moods, which reflect on stock markets (Bollen, Mao, & Zeng, Twitter mood predicts the stock market, 2011). Also, Twitter sentiment analysis is used to predict indices and big market cap companies with high accuracy (Rao & Srivastava, 2012). Other research even applied sentiment analysis and tweet volume for cryptocurrency prediction (Pant & Neupane, 2018). Some have dependently manly on microblogging as the main source of data, such as Twitter (Mittal & Goel, 2012). In general, researchers believe that social media is more affecting the microeconomic level rather than the news affecting the macro-economic level. Based on that, News NLP sentiment analysis was strongly investigated using different techniques and algorithms. Stock prediction for S&P500 and other companies was made using attention-based LSTM (At-LSTM) on directional changes utilizing RNN to encode the news text (Liu, 2018).

Analyzing news for stock prediction was done in many forms. One of the forms is analyzing the body of the article itself (Vargas, de Lima, & Evsukoff, 2017). Another research analyzed the headlines of the news (Kirange & Deshmukh, 2016). However, the news headlines are more

powerful in predicting than the content of the articles themselves (Radinsky, Davidovich, & Markovitch, 2012).

2.7 **Contribution**

The research will analyze news headlines to create a predicting stock system during Covid-19 as guidance for stock prediction during disasters. The research stands out as it is analyzing the news headlines of 2020 in which the pandemic of covid-19 spread. Applying NLP sentiment analysis on 50,000 News headlines will result in positive, neutral, and negative sentiment signals. The research extends beyond this point to use this sentiment to predict the stock trends in the following days accordingly. The research will apply this methodology to the five technology companies in the US market, along with the S&P500 Index.

Chapter 3 Objective & Methodology

3.1 Objective

The objective of this research is to develop a system that analyzes news text to predict the stock trends and take the right position to maximize returns for investors based on available data. This shall be done using NLP sentiment analysis. The analysis generates signals which are either positive, negative, or neutral. Based on the signals, a decision to buy or sell is placed in favor of the investor to maximize profit.

3.2 Methodology

The methodology will be utilizing the source of data for news. The news data collected is headlines of news articles published in the New York Times through an API using python code. It is also a must to match the frequency as well as the duration of the data for the same stock. As the effect of covid-19 is investigated, the research will focus on the date starting January 2020 to capture the impact of covid-19 introduction in the news until the end of the year. The data collected shall be examined using the word cloud analysis on both the headlines and the keywords. The target ensures that covid-19 reached the headlines and impacting our analysis. After this step, the headline text shall be tokenized and cleaning. Tokenizing the text is breaking down the sentence into words. This step shall be followed by removing punctuation and stop words. After this step, the tokenized words shall be joined and ready for sentiment analysis. By which natural language toolkit library used in python shall assign sentiment score for each headline. As there are several headlines per day, the sentiment score for the day shall be the sum of the sentiment scores of the day. On the other hand, stock data shall be retrieved for the same dates. Using machine learning,

the model shall be trained on the sentiment scores and historical stock performance. Accordingly, we will measure the accuracy of the predicting model.

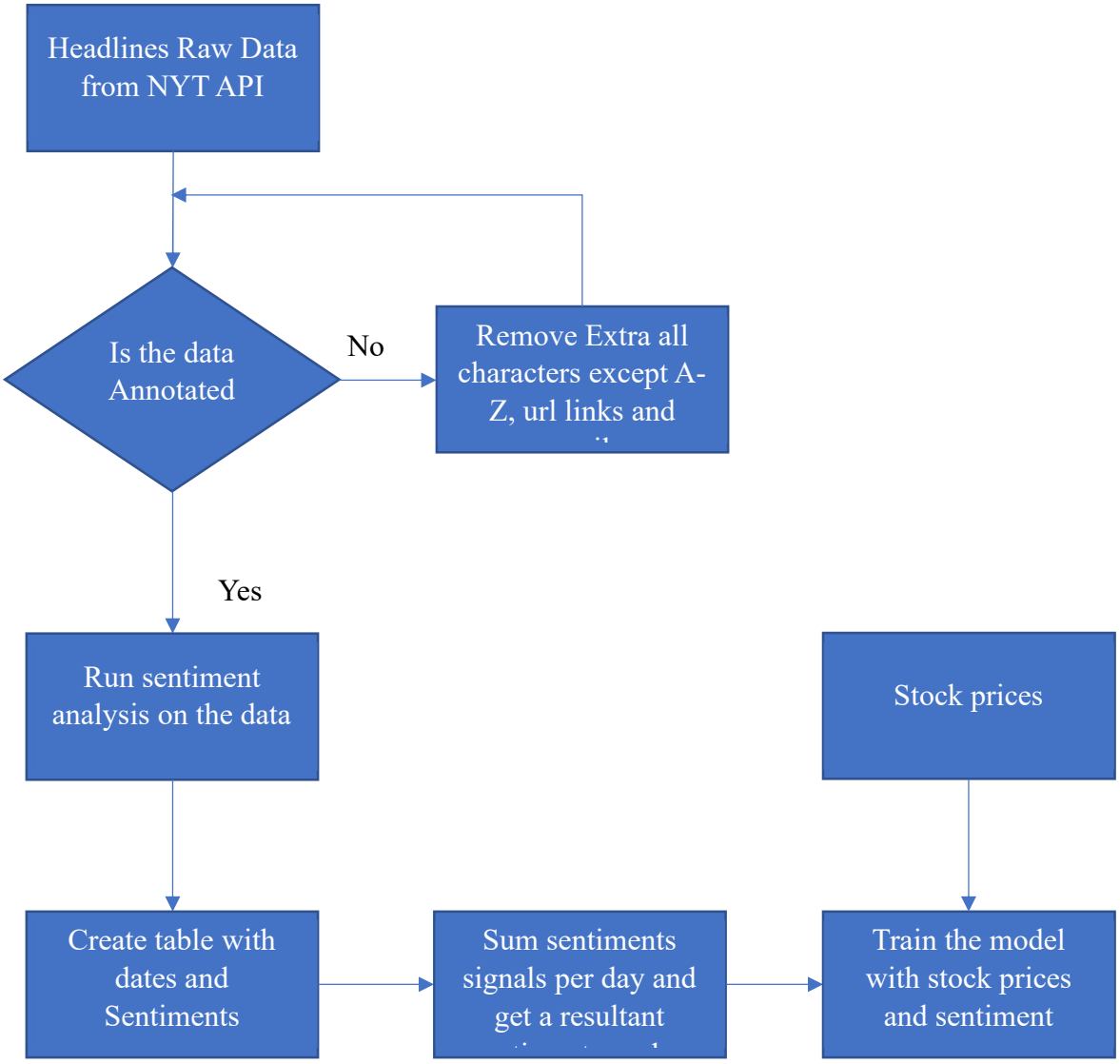

Figure 1: Flowchart describing the research methodology

3.3 Limitations

In this research, there was some limitation the shall be discussed in details in the following lines.

3.3.1 Data Limitations

In this analysis, we will use the news from the New York Times. It was hoped that microblogging using the Twitter microblogging platform added to this analysis. However, due to the high cost of retrieving microblogging tweets data, especially it is covering the whole year of 2020, it was limited to the news collected through the New York Times API.

3.3.2 Hardware Limitations

The trend in NLP sentiment analysis is deep learning algorithms that require high GPU computing power. However, due to the limitation of using personal hardware, the research is focused on using Natural Language Toolkit (NLTK) library for the sentiment analysis of the news headlines.

3.3.3 Stock Data Companies

Though there are many US-listed companies, we decided to focus on the technology companies in the US stock markets. It assumed that the technology sector was highly influenced during this period.

Chapter 4 Data Description

5.1 Types of Data Gathered

The data collected is divided into two main categories: the closing stock price and the news to be analyzed. In this chapter, the data shall be described. Also, the handling of the data shall be explained in detail.

5.2 News Headlines Data

On the other side, the news is the New York Times articles headlines for the same time frame covering different fields. The data is being extracted using python. The news data is extracted through New York Times articles headlines using NYT API.

Table 1: January 2020 headlines news sample

	headline	date	doc_type	material_type	section	keywords
0	Already Had Plenty of Trump 2020?	2020-01-02	article	Op-Ed	NaN	['Presidential Election of 2020', 'United Stat...
1	Why Did One-Quarter of the World's Pigs Die in...	2020-01-02	article	Op-Ed	NaN	['Pigs', 'Agriculture and Farming', 'Pork', 'L...
2	Coast Guard Suspends Search for 5 Missing Afte...	2020-01-02	article	News	NaN	['Rescues', 'Maritime Accidents and Safety', '...
3	N.B.A. Superstars, Growth and Lockouts: The Da...	2020-01-02	article	News	NaN	['Basketball']
4	In Rose Bowl Victory Over Wisconsin, Oregon Sh...	2020-01-02	article	News	NaN	['Football (College)', 'Rose Bowl (Football Ga...

The total number of headlines is 55,403, with an average of 4,613 headlines per month. In addition to the headlines, the news data includes the date, the document type, either article or media, material types, section, and keywords.

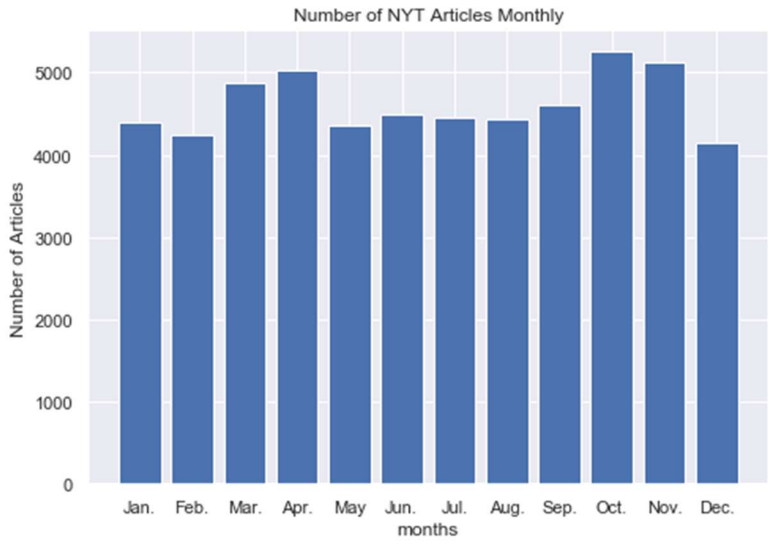

Figure 2: Number of NYT Articles Monthly

Table 2: Number of headlines per month

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
No. of NYT Headlines	4,390	4,240	4,883	5,019	4,347	4,492	4,459	4,439	4,609	5,257	5,114	4,154

Table 3: News Data Description

Website	Number of Articles	Covid-19 related Article	The average number of letters	Maximum number of letters	Method
New York Time	55,403	6,120	80	165	API

A word cloud analysis is performed to understand the most prominent news. The word cloud analysis is a presentation for the most repetitive word in the text and presenting them in bigger sizes than other words. The word cloud analysis is applied to every month's headlines. The target of the analysis is to understand how often Covid-19 showed up in the news. The first step is analyzing the word cloud figures based on the news headlines. The following step is analyzing the word clouds presented by the analysis of the keywords.

Based on the cloud analysis applied to the headlines, the Covid-19 impact is claimed to be between February 2020 and the following months. A second approach is taken to understand the headlines is applying the same word cloud analysis but on the keywords provided by the New York Times instead of focusing on the news headlines.

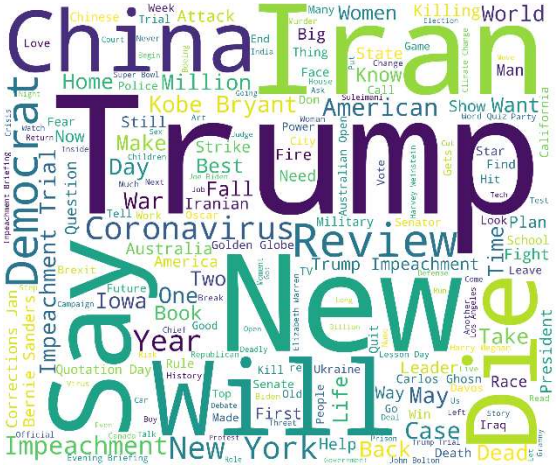

Figure 4: Sample Word Cloud analysis for Jan. articles headlines before Covid-19 Spread

Figure 3: Sample Word Cloud analysis for Jan. articles keywords before Covid-19 Spread

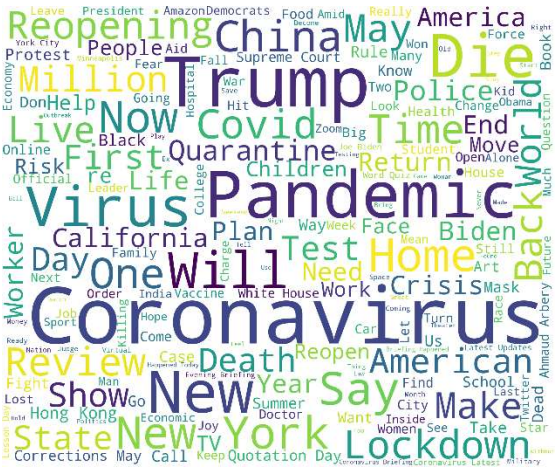

Figure 5: Sample Word Cloud analysis for May articles headlines after Covid-19 Spread

Figure 6: Sample Word Cloud analysis for May articles Keywords after Covid-19 Spread

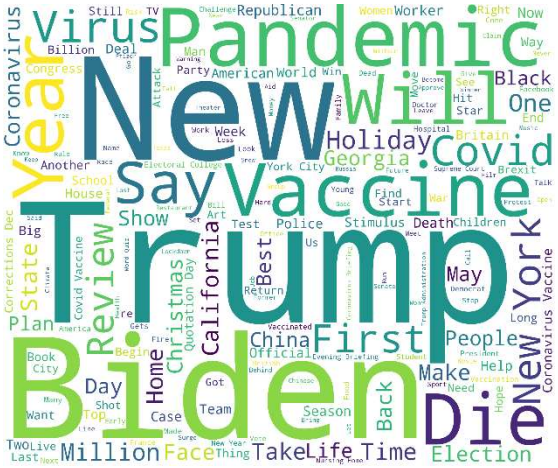

Figure 7: Sample Word Cloud analysis for Dec. articles headlines showing the decline of Covid-19 news headlines in comparison to US elections

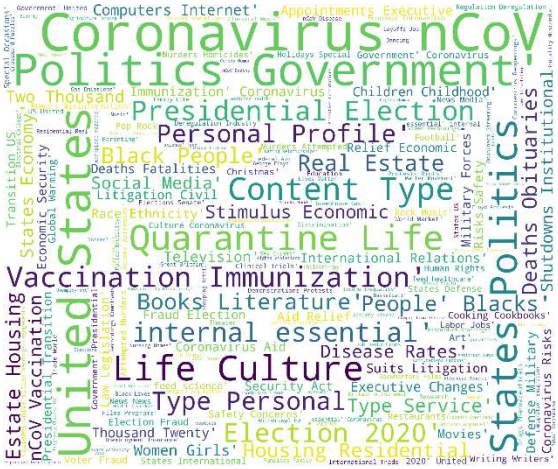

Figure 8: Sample Word Cloud analysis for Dec. articles headlines showing the decline of Covid-19 news headlines in comparison to US elections

After analyzing the two categories of images between headlines and keywords, it is noticed that the main difference is that the left side includes specifics such as names of persons, states, events, etc. However, the right-hand side is representing keywords. The word clouds are representing categories and general grouping. Tracing Covid-19 in the titles and keywords was done with many synonyms such as “Covid-19”, “coronavirus,” “ncov,” “pandemic,” Covid,” “Quarantine,” and “virus.” Most importantly, both groups have Covid-19 and related words showing starting in February 2020 until the end of the year. Towards the end, the US elections become dominant on the news along with Covid-19.

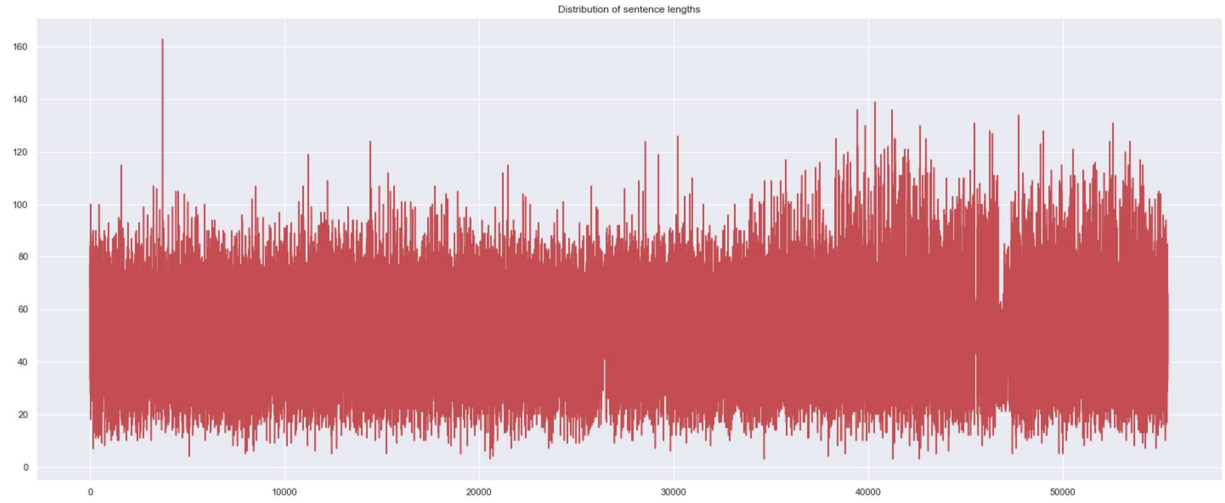

Figure 9: News Headline Sentences Length Line plot

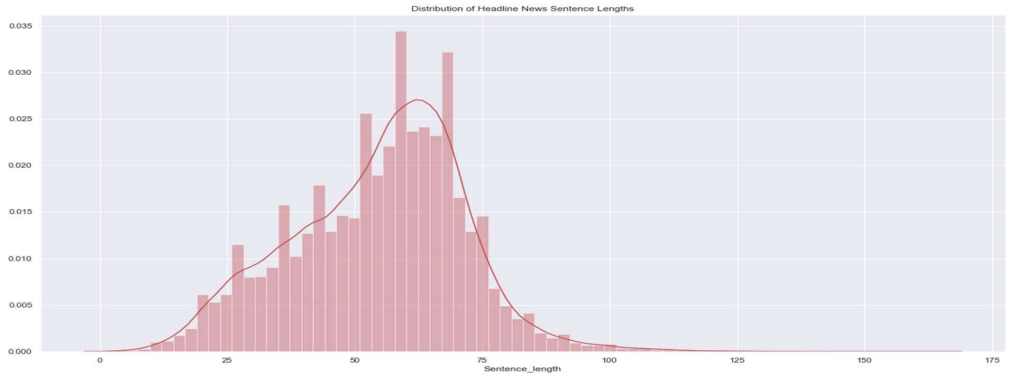

Figure 10: News Headline Sentence Length Distribution

5.3 **Data Processing**

After importing the data from the New York Times API, the data is manipulated by the python program and cleaned. The cleaning starts with tokenization, which is breaking down the sentences into words. The next step is turning all letters to lower case. The target of this step is to avoid any effect of the capital letter on the machine learning evaluation. The second step is converting the words to the lowest grade by removing “ion,” “ing,” etc. The following step is removing all the

stop words and questions word that will not affect the sentiment of the sentence. The stop words, for example, are “this,” “that,” etc. Also, removing all question words starting with’ w’.

Already Had Plenty of Trump 2020?
Why Did One-Quarter of the Worldâ€™s Pigs Die in a Year?
Coast Guard Suspends Search for 5 Missing After Fishing Boat Sinks Off Alaska
N.B.A. Superstars, Growth and Lockouts: The David Stern Years
In Rose Bowl Victory Over Wisconsin, Oregon Shows Rebuild Needs Work
Where Darth Vader Gets His Strength
Don Larsen, Yankee Who Pitched Only Perfect Game in World Series History, Dies at 90
No Corrections: Jan. 2, 2020
India Cold Wave Breaks Records, Shuts Schools and Makes Bad Air Worse
Quotation of the Day: Pika-Who? How Canadaâ€™s Military Reacted to a Pokí©mon Go Invasion

Figure 11: News Headlines before Cleaning

```
['alreadi plenti trump',
 'one quarter world pig die year',
 'coast guard suspend search miss fish boat sink alaska',
 'n b superstar growth lockout david stern year',
 'rose bowl victori wisconsin oregon show rebuild need work',
 'darth vader get strength',
 'larsen yanke pitch perfect game world seri histori die',
 'correct jan',
 'india cold wave break record shut school make bad air wors',
 'quotat day pika canada militari react pok mon go invas']
```

Figure 12: News Headlines after cleaning

Natural language toolkit library in python is utilized to analyze the cleaned news and give sentiment to each news. NLTK library uses more than 50 lexical resources and corpora to analyze text (Bird, Loper , & Klein, 2009).

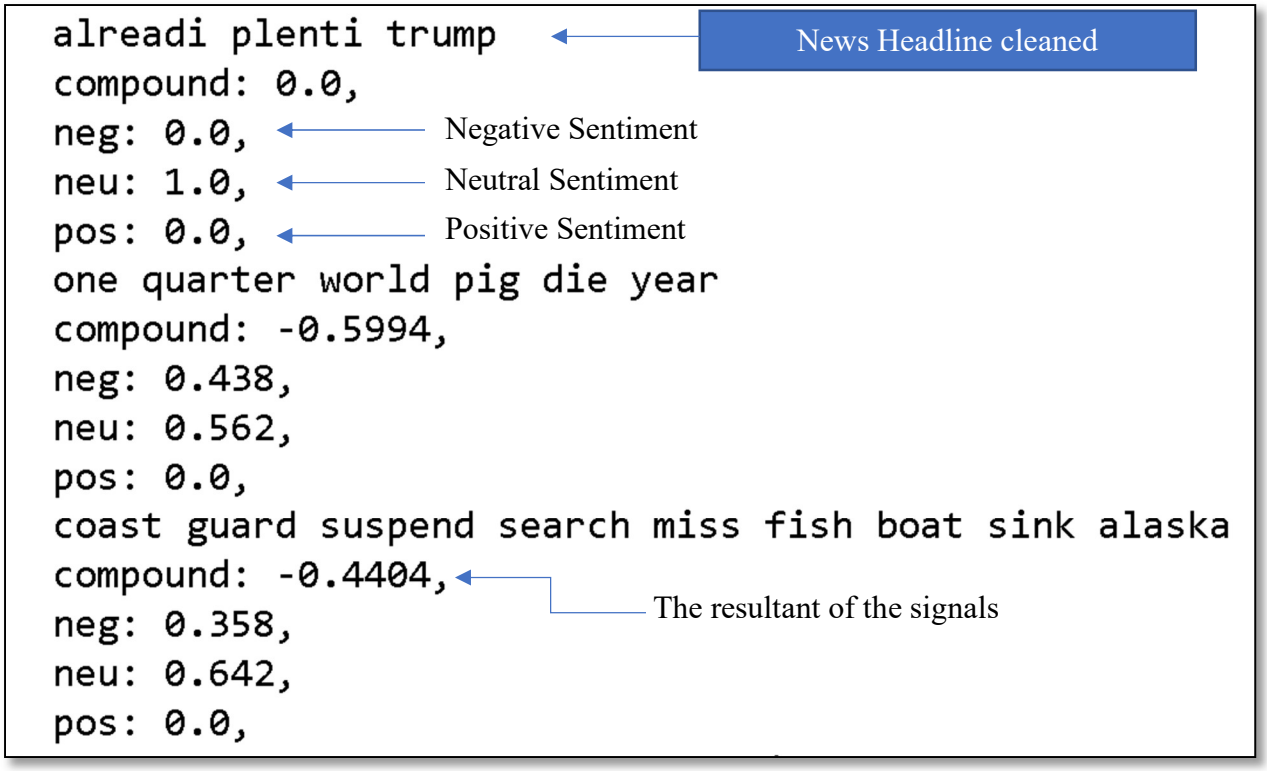

Figure 13: News Headlines Sentiment Sample

We assign a threshold of 0.2. For every headline, if the sum of the sentiment is more than 0.2, the sentiment score shall be one, and the sentiment is positive. If the sum of the sentiment is less than -0.2, the sentiment score shall be -1, and the sentiment is negative. And finally, between -0.2 and 0.2, it is considered the sentiment score shall be 0, and the sentiment is Neutral.

Table 4: Cleaned Headline News with Sentiment Score Assigned

	Date	cleaned_hlnews	Original_headline	sentiment	sentiment_score
0	2020-01-02	alreadi plenti trump	Already Had Plenty of Trump 2020?	neutral	0
1	2020-01-02	one quarter world pig die year	Why Did One-Quarter of the World's Pigs Die in...	negative	-1
2	2020-01-02	coast guard suspend search miss fish boat sink...	Coast Guard Suspends Search for 5 Missing Afte...	negative	-1
3	2020-01-02	n b superstar growth lockout david stern year	N.B.A. Superstars, Growth and Lockouts: The Da...	positive	1
4	2020-01-02	rose bowl victori wisconsin oregon show rebuil...	In Rose Bowl Victory Over Wisconsin, Oregon Sh...	neutral	0
5	2020-01-02	darth vader get strength	Where Darth Vader Gets His Strength	positive	1
6	2020-01-02	larsen yanke pitch perfect game world seri his...	Don Larsen, Yankee Who Pitched Only Perfect Ga...	neutral	0
7	2020-01-02	correct jan	No Corrections: Jan. 2, 2020	neutral	0
8	2020-01-02	india cold wave break record shut school make ...	India Cold Wave Breaks Records, Shuts Schools ...	negative	-1
9	2020-01-02	quotat day pika canada militari react pok mon ...	Quotation of the Day: Pika-Who? How Canada's M...	neutral	0

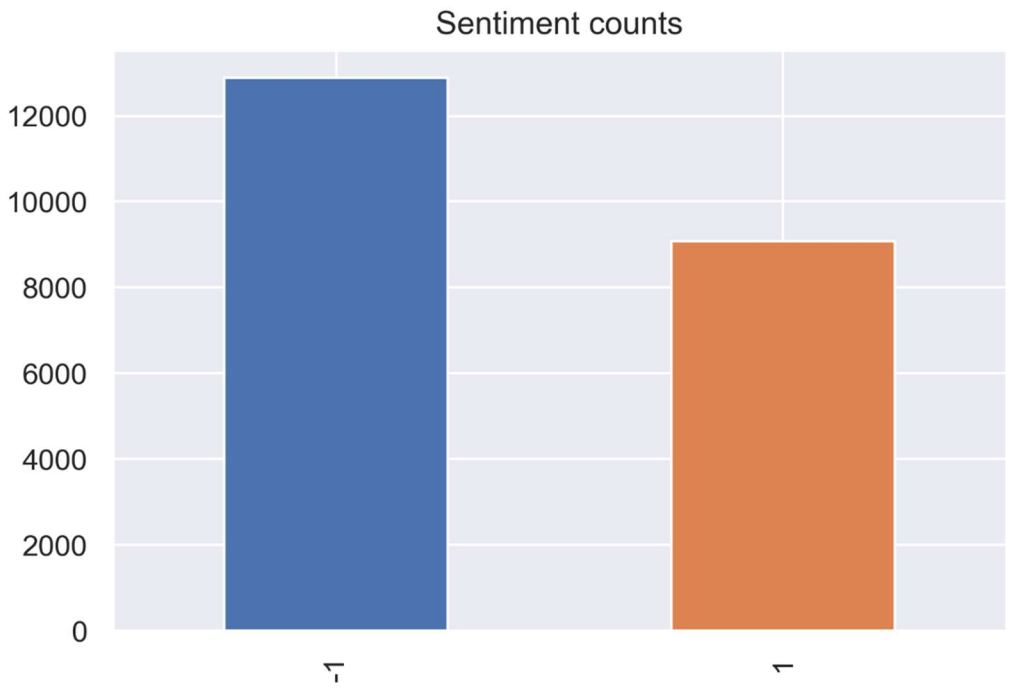

Figure 14: Sentiment Count

Table 5: Sum of Sentiment Scores per Day Sample

	Date	sentiment_score
0	2020-01-02	3
1	2020-01-03	-30
2	2020-01-04	-19
3	2020-01-05	-24
4	2020-01-06	-13
5	2020-01-07	-8
6	2020-01-08	-21
7	2020-01-09	-22
8	2020-01-10	-33
9	2020-01-11	-17
10	2020-01-12	-1
11	2020-01-13	-9
12	2020-01-14	-10
13	2020-01-15	-27
14	2020-01-16	-30
15	2020-01-17	-5

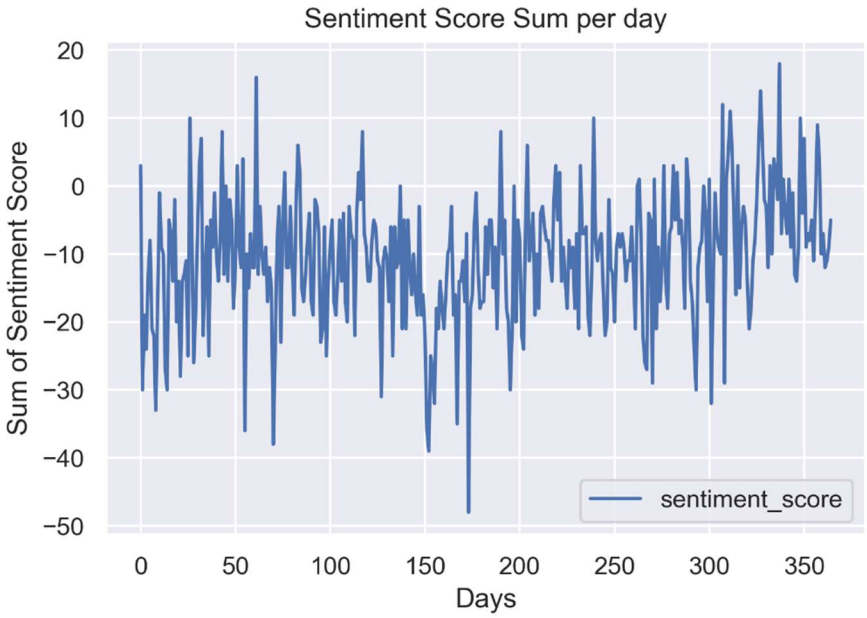

Figure 15: Sum of Sentiment Scores per Day

5.4 **Stock Prices Data**

The stock prices data shall be collected using python code. They are the adjusted stock price of Facebook, Amazon, Apple, Netflix, Microsoft, Google, and S&P500 from 1st Jan 2020 to 31st Dec 2020. For our NLP sentiment analysis, we will use three modes to compare to the returns, the percent of change, and log returns.

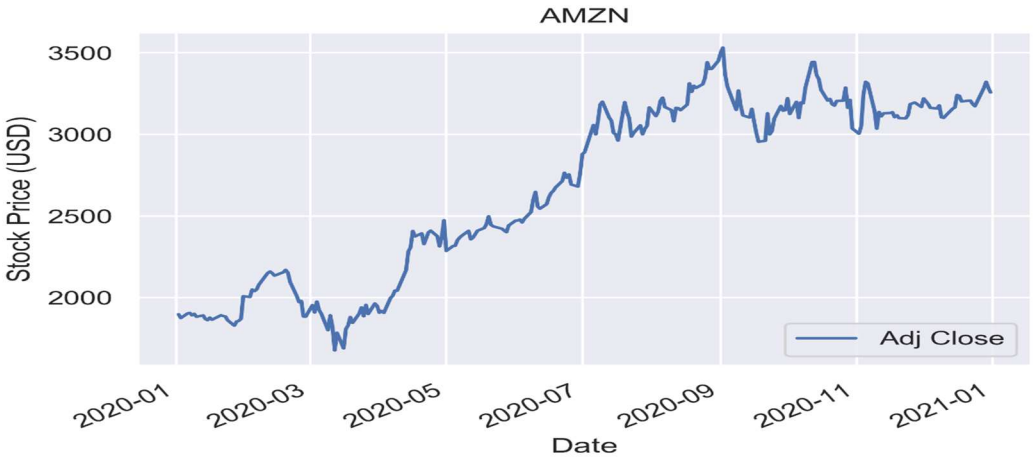

Figure 16: Sample for Amazon Stock Performance during 2020

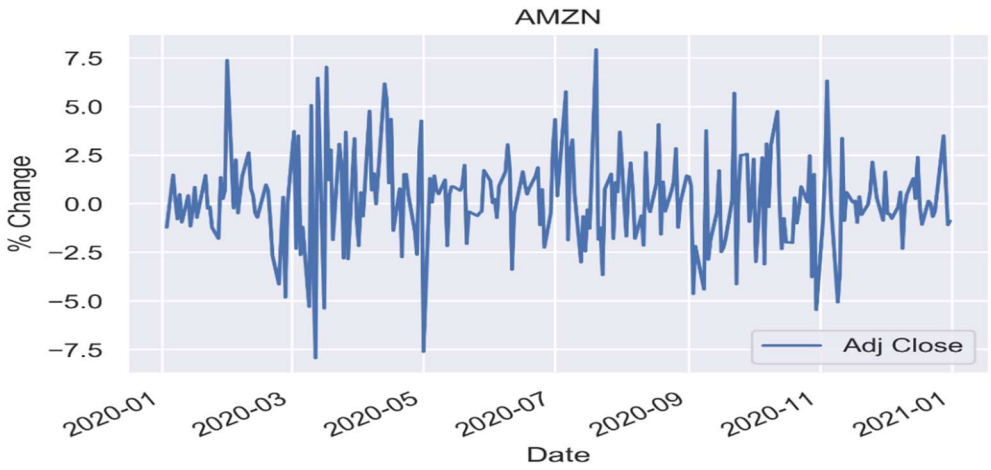

Figure 17: Sample for Amazon Stock % Change daily during 2020

Figure 18: Log Return of Amazon Stock

Table 6: Stock Data Sample for five companies and an Index

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY
2020-01-02	209.779999	74.444603	1898.010010	329.809998	158.936279	1368.680054	318.914307
2020-01-03	208.669998	73.720840	1874.969971	325.899994	156.957260	1361.520020	316.499451
2020-01-06	212.600006	74.308266	1902.880005	335.829987	157.362961	1397.810059	317.706909
2020-01-07	213.059998	73.958794	1906.859985	330.750000	155.928177	1395.109985	316.813568
2020-01-08	215.220001	75.148521	1891.969971	339.260010	158.411835	1405.040039	318.502075

Table 7: Stock Daily Percentage of Change Data Sample for five companies and an Index

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY
2020-01-03	-0.529126	-0.972216	-1.213905	-1.185532	-1.245165	-0.523134	-0.757212
2020-01-06	1.883360	0.796824	1.488559	3.046945	0.258478	2.665406	0.381504
2020-01-07	0.216365	-0.470300	0.209156	-1.512666	-0.911767	-0.193165	-0.281184
2020-01-08	1.013801	1.608636	-0.780866	2.572943	1.592822	0.711776	0.532966
2020-01-09	1.431095	2.124069	0.479927	-1.061135	1.249297	1.049792	0.678051

Table 8: Stock Log Returns Sample

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY
2020-01-03	-0.005305	-0.009770	-0.012213	-0.011926	-0.012530	-0.005245	-0.007601
2020-01-04	0.002683	-0.003868	-0.003217	0.002054	-0.007493	0.005272	-0.003798
2020-01-05	0.010671	0.002035	0.005779	0.016034	-0.002456	0.015788	0.000005
2020-01-06	0.018658	0.007937	0.014776	0.030014	0.002581	0.026305	0.003808
2020-01-07	0.002161	-0.004714	0.002089	-0.015242	-0.009159	-0.001934	-0.002816

In the following step, we add the sum of sentiments per day as a column at the end of each table to analyze using machine learning algorithms. Also, some of the days were not reported as they were holidays. It could have been interpolated. However, it was

5.5 **Indices Data**

Using python code, the indices under S&P500 were collected to understand the performance of different fields during Covid-19. It is notice

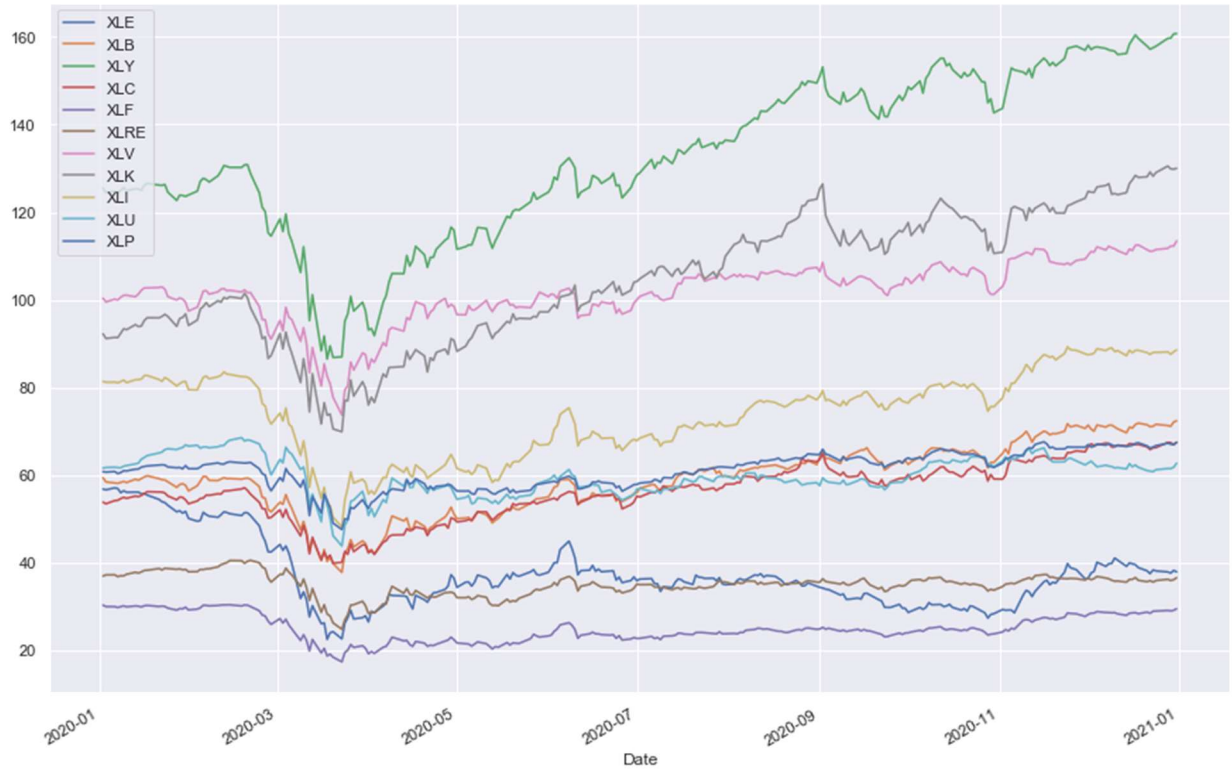

Figure 19: Indices Performance during 2020

The above graph shows the indices' performance during covid-19. However, it doesn't show the comparison between them as a percent of change. It is required another graph showing the percentage of change cumulative.

Figure 20: Percent Cumulative Change of Indices

As per this graph, it is clear that index XLE, which is the Energy index, and XLF, which is the financial index, are the most impacted sectors with a negative cumulative percentage at the end of 2020. On the other side, the XLK index, which is the technology index, has a positive cumulative portage of change. This means the technology outperformed its initial performance in 2020.

5.6 **Tables of Stock Prices Data and Sentiment Scores**

Table 9: Stock Prices and Sentiments Sum per Day

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY	sentiment_score
2020-01-02	209.779999	74.444603	1898.010010	329.809998	158.936279	1368.680054	318.914307	3
2020-01-03	208.669998	73.720840	1874.969971	325.899994	156.957260	1361.520020	316.499451	-30
2020-01-04	209.980001	73.916649	1884.273315	329.209991	157.092494	1373.616699	316.901937	-19
2020-01-05	211.290003	74.112457	1893.576660	332.519989	157.227727	1385.713379	317.304423	-24
2020-01-06	212.600006	74.308266	1902.880005	335.829987	157.362961	1397.810059	317.706909	-13

Table 10: Stock Percent of Change and Sentiments Sum per Day

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY	sentiment_score
2020-01-03	-0.529126	-0.972216	-1.213905	-1.185532	-1.245165	-0.523134	-0.757212	-30
2020-01-04	0.275036	-0.382536	-0.313084	0.225293	-0.743951	0.539713	-0.377640	-19
2020-01-05	1.079198	0.207144	0.587738	1.636119	-0.242736	1.602559	0.001932	-24
2020-01-06	1.883360	0.796824	1.488559	3.046945	0.258478	2.665406	0.381504	-13
2020-01-07	0.216365	-0.470300	0.209156	-1.512666	-0.911767	-0.193165	-0.281184	-8

Table 11: Stock Log Returns of Change and Sentiments Sum per Day

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY	sentiment_score
2020-01-03	-0.529126	-0.972216	-1.213905	-1.185532	-1.245165	-0.523134	-0.757212	-30
2020-01-04	0.275036	-0.382536	-0.313084	0.225293	-0.743951	0.539713	-0.377640	-19
2020-01-05	1.079198	0.207144	0.587738	1.636119	-0.242736	1.602559	0.001932	-24
2020-01-06	1.883360	0.796824	1.488559	3.046945	0.258478	2.665406	0.381504	-13
2020-01-07	0.216365	-0.470300	0.209156	-1.512666	-0.911767	-0.193165	-0.281184	-8

5.7 Stock Prices Data with binary Positive and Negative Sentiment

Two columns were created at the end of the table. The target of creating the two columns is to breakdown the sentiment score column into two columns of positive and negative sentiments for easier classification analysis by machine learning.

Table 12: Stock prices Log Returns with Binary Positive and Negative Sentiment

Date	FB	AAPL	AMZN	NFLX	MSFT	GOOGL	SPY	DAL	sentiment_score	Postive	Negative
2020-01-02	209.779999	74.444603	1898.010010	329.809998	158.936279	1368.680054	318.914307	58.634808	3.0	1	0
2020-01-03	208.669998	73.720840	1874.969971	325.899994	156.957260	1361.520020	316.499451	57.661533	-30.0	0	1
2020-01-04	209.980001	73.916649	1884.273315	329.209991	157.092494	1373.616699	316.901937	57.529114	-19.0	0	1
2020-01-05	211.290003	74.112457	1893.576660	332.519989	157.227727	1385.713379	317.304423	57.396694	-24.0	0	1
2020-01-06	212.600006	74.308266	1902.880005	335.829987	157.362961	1397.810059	317.706909	57.264275	-13.0	0	1

5.8 Statistics Analysis for the Matrix

	FB	AAPL	AMZN	NFLX	MSFT	GOOGL
count	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000
mean	0.083981	0.259837	0.224751	0.201170	0.145584	0.101049
std	2.599345	2.747854	2.278260	2.640684	2.513610	2.204871
min	-14.252998	-12.864693	-7.922081	-11.138862	-14.739028	-11.634152
25%	-1.114772	-1.047354	-0.971557	-1.168840	-1.000011	-0.763990
50%	0.181941	0.287138	0.167537	0.178536	0.238390	0.253393
75%	1.363606	1.572069	1.408638	1.612465	1.134388	1.289051
max	10.234995	11.980826	7.929524	11.608717	14.216898	9.241147

	SPY	DAL	sentiment_score	Postive	Negative
count	364.000000	364.000000	364.000000	364.000000	364.000000
mean	0.082281	0.143096	-10.469780	0.120879	0.862637
std	1.908140	4.596068	9.726659	0.326435	0.344704
min	-10.942365	-25.992439	-48.000000	0.000000	0.000000
25%	-0.534184	-1.856853	-17.000000	0.000000	1.000000
50%	0.217346	-0.217503	-10.000000	0.000000	1.000000
75%	0.926334	2.036795	-5.000000	0.000000	1.000000
max	9.060328	21.017103	18.000000	1.000000	1.000000

Figure 21: Statistics Analysis for matrix

Chapter 5 Results

In this research, three models were used to predict the stock prices for various companies and an index. The model's data are different in terms of the type of data used. The first model is built on adjusted stock prices. The second model is a daily parentage of change for the adjusted stock price. And the third model was using the log-returns of the same companies. In machine learning applications, various machine learning techniques were used, which regression models are, including logistic regression and random forest regressions. And for the sentiments

Table 13: The Models Matrix and Results using 2020 News Headlines

Aspect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Data	55,000 Headlines News During 2020 without Neutral					
Predicted items	Stock Prices	Percent of Change	Log Returns	Stock Prices	Percent of Change	Log Returns
Sentiment	Sum of Sentiment	Sum of Sentiment	Sum of Sentiment	Positive & Negative Sentiments Separately	Positive & Negative Sentiments Separately	Positive & Negative Sentiments Separately
Best MAPE Achieved	4.76-5.12%	> 50%	> 50%	4.6-5.2%	> 50%	> 50%
Ticker Achieved on	XLRE, XLU, XLP	NA	NA	XLRE, XLU, XLP	NA	NA

Model 1 and Model 4 results are shown in Appendix A.

On the other side, the same models were repeated using news headlines for ten years, which is around 755,000 headlines, and did not improve the results dramatically, as shown in the below table.

Table 14: The Models Matrix and Results using 10 Years News Headlines

Aspect	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Data	750,000 Headlines News During 2011-2020 without Neutral					
Predicted items	Stock Prices	Percent of Change	Log Returns	Stock Prices	Percent of Change	Log Returns
Sentiment	Sum of Sentiment	Sum of Sentiment	Sum of Sentiment	Positive & Negative Sentiments Separately	Positive & Negative Sentiments Separately	Positive & Negative Sentiments Separately
Best MAPE Achieved	5.68 - 9.18%	> 50%	> 50%	5.8-9.3%	> 50%	> 50%
Ticker Achieved on	XLRE, XLU, XLF	NA	NA	XLRE, XLU, XLF	NA	NA

The models' results are matching figure 20, where XLRE, XLU, and XLP indices were the most affected indices during Covid-19. These indices had the highest predicting accuracy during Covid-19 since they were the most affected. On the other hand, the Technology indices had the lowest accuracy in comparison since they outperformed their entry point of the year.

Chapter 6 Conclusion

Many models were created to predict stock price trends. The models developed were depending on the statistical algorithms, which are technical analysis of the stock performance. On the other side, due to the abundance of data and technology advancement, it availed more techniques and algorithms to understand and predict the stock price using NLP sentiment analysis of new and social media. However, most research was focusing on one side of the equation. This research is a trial of using NLP sentiment analysis on NEWs applying different machine learning models. The best predicting models were using the stock adjusted price and the Random Forest regression. And the best predictable tickers under this model are the real estate Index (XLRE), Gold Index (XLU), and Health index XLP.

Chapter 7 Further Research

A research improvement can be made by creating a matrix of different industry stocks and using the same techniques on the same field companies. Also, the research can be improved by combining microblogging data in the same model giving different weights for the news and microblogging. Another improvement recommended is to test the news headlines' impact on the stock market modeling by comparing two models: One of the models using the news headlines and the other one without the usage of headlines and comparing the accuracy improvement.

References

Asur, S., & Huberman, B. (2010). Predicting the Future with Social Media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492-499.

Awartani, B., & Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. *International Journal of Forecasting 21*, 167-183.

Baker, S., Bloom, N., Davis, S., Kost, K., Sammon, M., & Viratyosin, T. (2020). The Unprecedented Stock Market Reaction to Covid-19. *The review of asset pricing studies journal ranking*, 742–758.

Bird, S., Loper , E., & Klein, E. (2009). *Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.* O'Reilly Media Inc.

Bollen, J., Mao, H., & Xiao-Jun, Z. (2011). Twitter mood predicts the stock market. *Journal of computational science, 2*(1), 1-8.

Bollen, J., Mao, H., & Zeng, X.-J. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 1-8.

Costola, M., Iacopini, M., & Santagiustina, C. (2020). Public Concern and the Financial Markets during the COVID-19 outbreak. *SSRN*.

Das, S., & Padhy, S. (2012). Support Vector Machines for Prediction of Futures. *Computer Applications (0975 – 8887)*.

Deng, S., Mitsubuchi, T., Shioda, K., & Shimada. (2011). Combining technical analysis with sentiment analysis for stock price prediction. In Depend- able, autonomic and secure computing (DASC). *IEEE ninth international conference* (pp. 800-807). IEEE.

Devi, U., D.Sundar2, & Alli. (2013). An Effective Time Series Analysis for Stock Trend Prediction Using ARIMA Model for Nifty Midcap-50. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*.

Eisner, B., Rocktaschel, T., Augenstein, I., Bosnjak, M., & Riedel, S. (2016). Learning Emoji Representations from their Description. *4th International Workshop on Natural Language Processing for Social Media.*

Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives*, 157-168.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory And Empirical Work. *The Journal of Finance*, 383–417.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance 68 (3)*, 1267–1300.

Herwartz, H. (2017). Stock return prediction under GARCH — An empirical assessment. *International Journal of Forecasting 33*, 569–580.

Kachare, A. D., Kharde, R., & Dongare, A. (2012). Introduction to Artificial Neural Network. *International Journal of Engineering and Innovative Technology*, 189-194.

Kirange, D., & Deshmukh, R. (2016). Sentiment Analysis Of News Headlines For Stock Price Prediction. *OMPUSOFT: An International Journal of Advanced Computer Technology, 5(3*.

Lin, P.-C., & Chen, J.-S. (2008). A genetic-based hybrid approach to corporate failure prediction. *International Journal of Electronic Finance*, 1-14.

Liu, H. (2018). Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network. Ontario.

Mamaysky, H. (2020). Financial Markets and News about the Coronavirus. *SSRN*, 3565597.

Mittal, A., & Goel, A. (2012). Stock Prediction Using Twitter Sentiment Analysis. *Standford University, CS229*.

Odom, M., & Sharda, R. (1990). A Neural Network Model for Bankruptcy Prediction. *IJCNN International Joint Conference on Neural Networks.*

Oliveira a, N., Cortez a, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems With Applications*, 125-144.

Pant, D., & Neupane, P. (2018). Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. *2018 IEEE 3rd International Conference on Computing, Communication and Security.* IEEE.

Radinsky, K., Davidovich, S., & Markovitch, S. (2012). Learning Causality for News Events Prediction. *Proceedings of the 21st international conference on World Wide Web*, (pp. 909-918).

Rao, T., & Srivastava, S. (2012). Analyzing Stock MArket Movements Using Twitter Sentiment Analysis.

Rounaghia, M., & Zadehb, F. (2016). Investigation of market efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and yearly Forecasting of Time Series Stock Returns using ARMA model. *Physica A: Statistical Mechanics and its Applications*, 10-21.

Ruiz, E., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). Correlating Financial Time Series with Micro-Blogging Activity. *Proceedings of the fifth ACM international conference on Web search and data mining*, (pp. 513–522).

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development 3:3*, 210–229.

Schoneburg, E. (1990). Stock price prediction using neural networks: A project report. *Neurocomputing 2*, 17-27.

Selvin, S., R, V., Menon, V. K., K.P, S., & Gopalakrishnan, E. (2017). Stock Price Prediction Using LSTM, RNN and CNN-sliding Window Model. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1643-1647). Udupi: IEEE.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of. *The Journal of financial studies*, 7-26.

Tang, H., Chiu, K.-C., & Xu, L. (2003). Finite Mixture of ARMA-GARCH Model for Stock Price Prediction. *Proceedings of 3rd International Workshop on Computational Intelligence in Economics & Finance(CIEF'2003)*, (pp. 1112-1119). North Carolina.

Vargas, M., de Lima, B., & Evsukoff, A. (2017). Deep learning for stock market prediction from financial news articles. *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA).* Annecy, France: IEEE.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research 27*, 1131–1152.

Appendix A

	Ticker	LRMAPE1	RFMAPE2
0	FB	17.52	16.33
1	AAPL	25.88	25.75
2	AMZN	16.95	17.87
3	NFLX	12.08	12.00
4	MSFT	12.50	12.15
5	GOOGL	18.71	18.35
6	SPY	14.45	14.12
7	DAL	14.91	17.35
8	XLE	6.31	10.49
9	XLB	19.22	19.00
10	XLY	17.80	17.64
11	XLC	16.67	16.63
12	XLF	13.34	12.84
13	XLRE	4.19	4.76
14	XLV	9.55	9.48
15	XLK	18.28	17.24
16	XLI	17.87	16.50
17	XLU	5.26	5.67
18	XLP	9.44	9.49

Model 1 results showing Linear Regression Mean Absolute Percentage Error (MAPE) and Random Forest MAPE.

	Ticker	LRMAPE1	RFMAPE2
0	FB	17.52	16.62
1	AAPL	25.88	24.35
2	AMZN	16.95	17.70
3	NFLX	12.08	12.75
4	MSFT	12.50	12.39
5	GOOGL	18.71	18.27
6	SPY	14.45	14.38
7	DAL	14.91	19.54
8	XLE	6.31	10.93
9	XLB	19.22	19.30
10	XLY	17.80	17.22
11	XLC	16.67	16.01
12	XLF	13.34	12.90
13	XLRE	4.19	4.70
14	XLV	9.55	9.20
15	XLK	18.28	17.94
16	XLI	17.87	16.96
17	XLU	5.26	5.51
18	XLP	9.44	9.66

Model 4 results showing Linear Regression Mean Absolute Percentage Error (MAPE) and Random Forest MAPE.