FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Top CO2 emission firms stock market exchange rate prediction using Twitter

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Besjon Murturi
Matrikelnummer 01607383

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Wien, 1. März 2021

Besjon Murturi

Wolfdieter Merkl

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Top CO2 emission firms stock market exchange rate prediction using Twitter

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Besjon Murturi
Registration Number 01607383

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Mag. Dr. Wolfdieter Merkl

Vienna, 1st March, 2021

Besjon Murturi

Wolfdieter Merkl

Erklärung zur Verfassung der Arbeit

Besjon Murturi
Randhartingergasse 14-16, 1100 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. März 2021

Besjon Murturi

Danksagung

Zunächst möchte ich Professor Dieter Merkl für seine Verfügbarkeit, ständige Unterstützung und Ermutigung bei meiner Arbeit an dieser Arbeit danken.

Mein besonderer Dank gilt meiner Familie für ihre kontinuierliche Unterstützung in allen Aspekten während meiner Studienjahre und während der Arbeit an dieser Arbeit.

Abschließend möchte ich der Technischen Universität Wien dafür danken, dass sie mir die Möglichkeit bietet, hierher zu kommen und zu studieren.

Acknowledgements

At first, I would like to thank professor Dieter Merkl for his availability, constant support and encouragement during my work on this thesis.

Special thanks goes to my family for their continuous support in every aspect throughout my years of study, and during the work on this thesis.

Finally, I would like to thank the Technical University of Vienna for offering me the opportunity to come and study here.

Kurzfassung

Die weit verbreitete Nutzung des Internets hat es mehr Menschen ermöglicht, ihre Meinung zu verschiedenen Themen zu äußern. Twitter ist eine der am häufigsten genutzten Plattformen zur Meinungsäußerung und öffentlichen Diskussion und generiert eine Vielzahl von Daten, mit denen die Börsenbewegungen für verschiedene Bereiche vorhergesagt werden können. In dieser Arbeit befassen wir uns mit den Methoden, mit denen der Börsenkurs der führenden CO2-Emissionsunternehmen mithilfe einer Stimmungsanalyse anhand der von Twitter gesammelten Daten vorhergesagt werden kann.

Wir analysieren Daten im Zeitraum von zwei Monaten von Juli 2020 bis August 2020. Die gesammelten Tweets haben einen Stimmungswert für jeden Tweet, der mit Valence Aware Dictionary for Sentiment Reasoning (VADER) berechnet wird. Da es sich bei dem Börsenkurs um Zeitreihendaten handelt, berechnen wir die Stimmung über den Zeitraum, d. H. Über eine Zeitspanne von einer Minute, um die Stimmung als Zeitreihenprädiktor zu verwenden. Darüber hinaus berechnen wir die zeitversetzte Stimmung innerhalb der Zeit. Diese Methode richtet die Stimmung im Zeitverlauf an den stark korrelierten Schlusswert der Aktie aus, was bedeutet, dass die Gesamtstimmung innerhalb einer bestimmten Zeit dazu führt, dass die Aktienkurse infolge des Schlusskurses für die gegebene Zeit schwanken. Wir summieren die Stimmung der Tweets über den gegebenen Zeitrahmen (stündlich) als Maß für das gesamte Stimmungsmaß innerhalb der Zeit, da es sowohl die Häufigkeit als auch die Summe der Tweets darstellt. Dieser kumulative Wert wird dann als beabsichtigter Indikator verwendet. In Anbetracht der Tatsache, dass gleitende Durchschnitte im Allgemeinen als Indikator für den Aktienhandel verwendet werden, berechnen wir gleitende Durchschnitts- und exponentielle gleitende Durchschnitte für die Stimmung innerhalb der Zeitwerte, die aus den Tweets berechnet wurden, die sich auf die Hashtags bezüglich Ölproduzenten und Klimawandel beziehen. Der gleitende Stimmungsdurchschnitt wird für die Zeitspanne von 120 Minuten berechnet, und der exponentielle gleitende Durchschnitt wird für eine Zeitspanne von 120 Minuten berechnet, die die beste Korrelation mit dem engen Aktienkurs aufweist. Dies impliziert, dass gleitende Durchschnittswerte der Stimmung innerhalb der Zeit von Tweets vor zwei Stunden mit dem aktuellen Aktienkurs korrelieren. Am Ende nehmen die Regressionsmodelle Support Vector Regression (SVR), Bayesian Ridge Regression (BRR) und Kernel Ridge Regression (KRR) die Prädiktoren auf, die als gleitende Durchschnittswerte, exponentielle gleitende Durchschnittswerte berechnet und nahe verschoben wurden, um die Aktienkurse vorherzusagen.

Unsere Ergebnisse in dieser Arbeit zeigen, dass die Stimmungsanalyse von Tweets von Twitter verwendet werden kann, um den Wechselkurs der Top-CO2-Emissionsunternehmen erfolgreich vorherzusagen.

Abstract

The widespread usage of internet has made it possible for more people to be able to express their opinions about different topics. Twitter being one of the most used platforms for expressing opinions and public discussions, generates a vast amount of data which can be used to predict the stock market movements for different areas. In this work, we tackle the methods that can be used to predict top CO2 emission firms stock market exchange rate with the help of sentiment analysis performed on data gathered from Twitter.

We analyze data in the course of two months period, from July 2020 until August 2020. The collected tweets have sentiment score for each Tweet which is calculated using VADER. Since stock market price is a time-series data, in-order to use sentiment as time-series predictor we compute sentiment across the time period i.e over the time span of a minute. Furthermore, we proceed to compute time-shifted sentiment within time. This method aligns the sentiment over time to highly correlated close value of the stock implying that overall sentiment within a time causes the stock prices to fluctuate as the result of closing price for the given time. We sum the sentiment of the tweets across the given time-frame (hourly) as measure of the overall sentiment measure within the time as it represents both frequency and sum of the tweets. This cumulative value is then taken as the intended indicator. Considering that moving averages are generally used as stock trading indicator, we compute moving average and exponential moving averages for sentiment within the time values computed from the tweets related to the hashtags regarding oil producers and climate change. The sentiment moving average is computed for the span of 120 minutes, as well as the exponential moving average is computed in span of 120 minutes which has best correlation with close price of stocks. This implies that moving averages of sentiment within time of tweets before two hours correlate to the current stock price. In the end, the regression models SVR, BRR and KRR ingest the predictors computed as moving averages, exponential moving averages and shifted close in order to predict the stock prices.

Our findings in this thesis, demonstrate that sentiment analysis performed on tweets gathered from Twitter can be used to successfully predict the top CO2 emission firms exchange rate.

Contents

CHAPTER 1

Introduction

1.1 Problem statement

Stock market has been an area that appealed to researchers for a long time and a lot of studies and analyses about investing in these markets were conducted [BC87], [Gor62], [MM58]. Prices in the stock market are generally induced from new information and tend to have patterns. Quite a few people have tried to derive patterns in the way which stock markets behave and respond to outer stimuli [MG12]. One of the investors goals was to maximize the profits and to minimize the losses. For this purpose, they were based on news, which due to the lack of Internet, needed a long time to be proven [RS$^+$12]. Nowadays, the investors still rely on news about the market. However, the time needed to prove these news, with the emerge of the Internet is largely reduced.

In order to forecast the stock market fluctuations, analysts use a set of different techniques, which include historical data analysis, cycle analysis, trend analysis and a combination between them [BC87]. Additionally, the interaction of the stock market and the real economy is vital for the different channels through which financial markets push economic growth [PM18]. In this way, macroeconomic factors can be used in the stock market predicting process [CRR86].

Everyday activities in today's world are being documented and discussed on social media and can influence stock markets [BMZ11]. Twitter is a social media network where microblogging is being used as a way of communication, where thoughts and feelings are shared in short messages, and is currently one of the most popular social networks by the number of active users[1]. Having such a high number of active users and such a high reach, Twitter is being used not only from the general population but also from people representing different profiles like politicians, actors, singers, etc. in order to share their

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed March, 2020

personal opinions or information to their followers, raising awareness for different topics, such as climate change.

Based on data gathered since 1880, $\sim$ 57 % rise of CO2 in atmosphere, $\sim$ 42-52 % of the rise in surface temperature, and $\sim$ 26-32 % global sea level rise can be directly linked to 90 major industrial carbon producers. Furthermore, one third of all CO2 emissions can be directly associated to the top 20 fossil fuel companies, which constantly exploited the world's oil, gas and coal reserves. Out of these 20 companies, 11 are being state-owned and 9 are multinational firms. These companies are driving the climate crisis, which threatens the future of humanity and have continued to grow their operations even though the awareness of their industry's disastrous impact on the planet. The list of these top 20 companies is composed from Saudi Aramco, Chevron, ExxonMobil, BP, Gazprom, National Iranian Oil Co, Royal Dutch Shell, Pemex, Petróleos de Venezuela, Coal India, ConocoPhillips, British Coal Corporation, Peabody Energy, PetroChina, Total France, Kuwait Petroleum Corp, Abu Dhabi National Oil Co, Sonatrach, CONSOL Energy, BHP Billiton [EBD+17].

Changes in the levels of CO2 in the atmosphere were always of interest for a lot of researchers [EBD+17], [Hee14], [SDRH13]. Based on the research that has been done, dilemmas have been raised about the impact of emissions from coal, oil and gas produced by fossil fuel companies and the damage that these companies are doing to nature with their industry's devastating operations. As described by Taylor and Watts "The great tragedy of the climate crisis is that seven and a half billion people must pay the price – in the form of a degraded planet – so that a couple of dozen polluting interests can continue to make record profits. It is a great moral failing of our political system that we have allowed this to happen." and as Heede said, cited by Taylor and Watts: "These companies and their products are substantially responsible for the climate emergency, have collectively delayed national and global action for decades, and can no longer hide behind the smokescreen that consumers are the responsible parties."[2]. Not only researchers are worried about this topic but also people from different social spheres are worried and engaged in this field. Twitter being one of the most used social media platforms for public discussion, it is also a center of attention for the climate change and global warming topic. Hence, people use Twitter to raise debates and express their opinions regarding the existence of a climate change crisis, whether they are pro or against. Consequently, people's sentiment is now very important, as it can drive the investment decisions of the investors and directly impacting the stock exchange rate of these top CO2 emission firms.

The target of this work is to predict stock prices of the top CO2 emission firms by leveraging the correlation between the sentiment of tweets gathered from Twitter and stock prices, taking them as time-series data.

Research questions:

[2] https://www.theguardian.com/environment/2019/oct/09/revealed-20-firms-third-carbon-emissions, accessed March, 2020

- What is an appropriate approach to predict stock prices of the top CO2 emission firms, based on using the sentiment of tweets gathered from Twitter?

- Is it possible to predict increase in stock prices of the top CO2 emission firms, based on negative cumulative sentiment expressed in tweets?

- Correspondingly, is it possible to predict decrease in stock prices of the top CO2 emission firms, based on positive cumulative sentiment expressed in tweets?

1.2 Aim of the Work

The objective of this research is to find out an appropriate approach to predict stock prices of the top CO2 emission firms based on using the sentiment of tweets gathered from Twitter, and whether it is possible to predict an increase or decrease in the stock exchange rate of these firms. Different supervised machine learning algorithms that fall under the category of regression, will be used in order to find the model that fits the best in predicting stock prices of the top CO2 emission firms.

Due to Twitter being currently one of the most popular social networks by the number of active users[3] and its microblogging function [JSFT07], it has become a platform for sentiment mining. Sentiments expressed on Twitter serve as a reference for people in general when it comes to their decision and assumption making process. Studies made from behavioral economics have shown that emotions can greatly affect individual behavior, this also applies when it comes to the stock market movements. Therefore, indicates that sentiment analysis is a successful tool when it comes to predicting the stock market and the stock market indicators, such as the Dow Jones Industrial Average index (DJIA) [BMZ11]. Studies made from other researchers have shown the same pattern where high correlation between stock prices and Twitter sentiment has been found [MG12], [RS$^+$12], [PP10].

In this work, by analyzing tweets that refer to climate change and global warming, and by using the right supervised machine learning algorithm from the regression category, it is expected to show that sentiment analysis can be used to successfully predict the top CO2 emission firms exchange rate. When the prices of the top CO2 emission firms are increasing, there are a lot of negative tweets regarding the problem of climate change and global warming. Meanwhile, when the prices are decreasing, there are a lot of positive tweets about the work being done regarding climate change and global warming. This would mean that overall negative sentiment will indicate that the prices of the top CO2 emission firms are increasing, while positive sentiment will indicate a decrease in the exchange rate.

[3] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed March, 2020

1.3 Methodological Approach

In order to answer the research question whether we can use sentiment analysis of tweets regarding climate change to successfully predict the stock market movements of the top CO2 emission firms, different approaches will be used. In the first part, in the theoretical phase an explanation for how social media works, how it influences the economy, and how these influences can be used, will be provided. As Arias, Arratia and Xuriguera [AAX14] elaborate: "The dramatic rise in the use of social network platforms such as Facebook or Twitter has resulted in the availability of vast and growing user-contributed repositories of data". Based on these data a lot of conclusions and analyses can be made. Sentiment analysis techniques will be considered, and it will be attempted to study different methods in order to discover the one that suggests the best results with short messages such as tweets. In the second part, the practical one, focus will be put on gathering all the necessary data for a period of two months. Furthermore, tweets that contain the hashtags of #climatechange, #globalwarming, #carbonproducers, #endfossilfuel, #CO2, #CH4, #surfacetemperaturerise #sealevelrise, #exxon, #bp_oil, #chevron will be collected. These tweets will be used for sentiment analysis, in order to predict the rises and falls of stock prices of the top CO2 emission firms. Text-Blob[4] toolkit will be used in this research to process tweets, which is a set of tools specialized in natural language processing with support for VADER sentiment analysis. Yahoo finance[5] will be used to collect actual stock prices of these top CO2 emission firms. Supervised machine learning algorithms like: Bayesian Ridge Regression, Support Vector Regression, Kernel Ridge Regression will be tested in order to find the model that fits the best in predicting stock prices of the top CO2 emission firms. Predictions made regarding the stock prices of the top CO2 emission firms will be compared with the real stock prices of these firms, collected from yahoo finance. Moreover, we will evaluate how close our predictions about the stock prices of these top CO2 emission firms, gained from our model are with the real stock prices of these firms.

1.4 State of the art

Everyday activities in today's world are being documented and discussed on social media, no matter the spectrum they belong to. Everyone has the freedom to express his opinion on these platforms for a particular event, service, product, etc. and these opinions are being used as references from other people in their decision and assumption making process. That being sad, this clearly shows how important are people's opinions on social networks [TBP11]. Twitter is one of the social media networks where these kinds of opinions are shared, and it is currently one of the most popular social networks by the number of active users[6]. Therefore, due to the large number of participants and the large

[4] https://textblob.readthedocs.io/en/dev/, accessed March, 2020

[5] https://finance.yahoo.com/, accessed March, 2020

[6] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed March, 2020

number of tweets, Twitter can be used to analyze public opinion [JSFT07].

Messages shared on Twitter are called tweets. Tweets are up to 280-character messages which can be shared from Twitter's web or mobile application. Tweets contain hashtags to determine the topic of the tweet and the @ character in order to tag a user in a tweet. Having such a high number of active users and such a high reach, Twitter is being used not only from the general population but also from people representing different profiles like politicians, actors, singers, etc. in order to share their personal opinions or information to their followers. Furthermore, businesses are also using this powerful opportunity to reach their clients through Twitter, where from their brand profiles they inform the customers for the latest news about the products that they have, or services that they offer and also gather feedback and give response through comments to their users [MB19]. Additionally, Twitter's importance lies beyond its microblogging function. Many studies have been made showing that it is possible to predict stock market indicators such as DJIA (Dow Jones Industrial Average), NASDAQ-100, SP 500 500 as well as the stocks exchange rate of different companies through sentiment analysis of tweets, where it has been found a high correlation between stock prices and Twitter's sentiments [MG12], [RS+12], [BMZ11]. In order to evaluate public mood directly from social media and to assess if a message sentiment is positive or negative, there exists a variety of sentiment analysis algorithms. Text-blob[7] toolkit being one of the algorithms, is a set of tools specialized in natural language processing.

A lot of research has been done regarding the emissions of carbon dioxide CO2 and methane CH4 on the atmosphere, the rise in surface temperate, and the global sea level rise [EBD+17], [Hee14], [SDRH13]. Based on this research the planet is getting warmer, the glaciers in Antarctica and Greenland are melting at an extreme rate[8]. Many species have started to disappear because of the environmental change where their conditions for life are not being met anymore [TCG+04]. On the other hand, there are people who don't support the climate change theory and the warming of the earth is considered from them as a cycle which the planet goes by periodically [CH13]. All these differences in opinions lay the ground for a big debate regarding climate change and global warming on social media networks. In our research we analyze the social network Twitter and perform sentiment analysis in tweets about above mentioned topics.

1.5 Structure of the work

This master's thesis is split up as defined into the chapters below.

Chapter 2: Theoretical Background. In this chapter the theory about stock markets, top CO2 emission companies stock markets, twitter and sentiment analysis is explained.

Chapter 3: Literature Review. This chapter covers the observations gathered from

[7] https://textblob.readthedocs.io/en/dev/, accessed March, 2020

[8] https://climate.nasa.gov/evidence/, accessed March, 2020

books, scholarly articles, and other sources related to using Twitter to analyse public opinion, Twitter sentiment analysis, stock market prediction using Twitter sentiment analysis, top oil producers stock market analysis and the impact of top oil producers in climate change and global warming.

Chapter 4: Methodology. This section describes the system that we have setup from data finding, data collection and trained models. We highlight the data pre-processing and VADER sentiment analysis model leading to feature engineering of highly correlated predictors. Finally, we establish the relevant Machine Learning (ML) model for regressive forecasting of the stock prices.

Chapter 5: Implementation. This section describes implementation of the Stock Prediction for Oil Corporations based on the Tweet data. It describes the data pre-processing which includes data collection, sentiment analysis of tweets within time and the moving averages. Furthermore, it depicts the training data extraction and the regression models used for prediction.

Chapter 6: Results. In detailed description of the gained results will be shown, followed up by visual representation of the achieved outcome.

Chapter 7: Conclusion. The final chapter contains the conclusion of the work done. Limitations of the study will be discussed together with further work possibilities.

CHAPTER 2

Theoretical Background

2.1 Stock Markets

Stock markets are defined as a conglomeration of markets and exchanges where regular selling, buying, and issuance of shares owned by public companies occurs. A country can have a single stock market at a single specific location. At the same time, a country can have numerous stock trading locations or venues that permit trading transactions in stocks or other securities forms. A share is a potion very crucial in stock exchange markets. A share is a section of ownership or a proportion entitled to a specific person or group in a business. Shares represent a divided company or business entity ownership. For example, if one share is examined, it is regarded as a proportion representing partial ownership of a business entity or a company under examination. Stocks are either sold or bought at an exchange market. Stock is thus a tradeable asset. As a tradeable asset, stock trading is governed by different regulations that vary depending on the market and the country of operation of the trade. Subject to irregularities concerning a country's various laws in managing the stock trading is regarded as a breach in the trading stock protocols in the exchange markets. As a result, a country might enact several legislative measures to govern its stock and, at the same time, ensure that no rule is breached in the process of trading stocks. The main reasons for adopting different rules and laws by different nations are to prevent any attempt of money laundering in trading stocks and prevent fraud and malpractices that may arise in due course of trading stocks.

Stock exchanges are a representation of places where bonds and securities can be purchased and sold. Moreover, other financial securities apart from bonds can also be bought and sold at stock exchanges. Traders or stockbrokers in these markets perform the trade. For companies to participate in the stock exchange as a trade, they must be listed on the formal exchanges. If a company is not listed in the formal exchanges but wants to participate in stock trading, the alternative trading stock is used. The alternative is called over the counter trading. Stock exchanges are entirely a subset of their respective stock markets. That is, all the trading that is done on the stock exchange is regarded as a proportional representation of the respective stock markets. Numerous stock exchanges have stood out to exemplary stock trading performance, showing signs of continuing stock trading. As aforementioned, different nations have their stock exchange markets. Globally, several markets have stood high in comparison with other stock markets around the world. Some of the best performing stock markets include the New York Stock Exchange, the Japan Exchange Group, Deutsche Börse, the London Stock Exchange Group, NASDAQ, Hong Kong Stock Exchange, the Toronto Stock Exchange, and Bombay Stock Exchange, among other top performing exchange markets.

Various company shares are traded on a daily basis in the stock market. As a result, people who sell or buy stocks tend to decide to make the best tactics to use while trading stocks. They examine multiple ways and methods they can employ in trading stocks at the exchange markets. The tradeable assets all possess values, whether they are real assets or those assets that are financial. It is the value of the assets being traded that is of much value. The management of an asset in stock markets is thus crucial. Though it is crucial, it is not the first priority for a stock trader. The top priority for every stock trader is the various factors that drive an asset's values and the factors that make an asset volatile in the market [Mel05]. Therefore, it is critical for a trader to equip themselves with different information regarding trading the different assets, to sell or buy in the stock market. For example, if a trader intends to practice a publicity asset trade, they will not use the same information to trade a real estate asset. It means that the information on trading a publicity asset significantly differs from the information for trading a real estate asset. The same phenomenon will be applied to all assets under trade. A piece of specific information will be required for a specific tradeable asset, not a piece of general information needed to trade all assets. The reason for different information for different assets cannot be overstated beyond the values of the assets being traded and the factors that determine their volatility [SFY19]. As a result, knowing how to analyze the market is very critical for any stock exchange trader. Furthermore, they must equip themselves with the knowledge to determine the actual or the true value of an asset they intend to trade. The different approaches used to determine an asset's actual value in the stock market cannot be fully adopted by everyone in the trading platform. However, it seems that all the stock traders and buyers share a common belief that they cannot depend on the belief that all the assets can be bought at a higher price by a potential customer in the future.

A significant cause of the rise or the fall of the price on assets is the availability of economic news worldwide and in areas that practice the stock exchange. The piece of assets is also immensely influenced by external forces [KM05]. Several other parameters would influence the price of an asset in a stock market. Like in any business, there are risks in trading stock in the exchange market [Yin66]. In economics, risks are regarded as a return in the investment made by a trader that is different from the return the trader expected after trading the assets. Risks are classified as either downside risk or upside risk because risks can go either way in a trade. In an upside risk, the return realized by the trader can be higher than was initially expected to be, meaning that the trader would realize tremendous profits higher than what they expected. In this type of risk, the outcome is always a good one. However, in downside risk, a trader can discover that they have realized a much lower return than they expected after conducting the trade. The return can be positive, that is, a profit is realized though it does not match the expected profit by the asset trader, and negative return, which means that a trader may end up incurring losses in the business at last. This type of risk spells a piece of bad news for prospective investors in the market. When an investor buys an asset, there is an expected return they want to realize after selling the asset. However, their expected return may sometimes, or in most scenarios, differ from the actual return. It is the difference in the expected return and the actual return that defines business risks [Mat16].

2.2 Top CO__2 emission companies stock markets

Several nations have had a significant effect on the quantity of CO__2 emissions in the atmosphere. The emissions have caused a severe reaction to the climate, as they have inflicted change on the climatic conditions, mostly to the worse. Though there are several proofs that CO__2 emissions to the atmosphere are equally harmful, the companies have done little with respect to the emissions. The governments of these nations have also played a part in contributing to the company emissions, as most of them have stood in awe looking at the progress the companies have made financially in the recent past instead of holding the companies responsible for their effects on climate change. Climate change has posed dangerous effects in society, affecting not only the surroundings but also human lives. America is one of the countries that have been impacted heavily by the climate changes due to substantial gas emissions by the various companies in the USA. If not controlled in the near future, the emissions can result in even worse results to society and the ecosystem in general. As a result of these dangerous emissions worldwide, it is vital to examine the various companies that have contributed significantly to the emissions of CO__2 in the atmosphere. The companies are mostly fuel-producing factories that strive to realize profits through their activities to produce oil for use in society. The test that remains for these companies is their compliance with the several guidelines that have been issued in their country of operation to cater for the control of climatic changes that may put the lives of the people at risk. The ideal measure to take remains complying with the scientific regulations on a sustainable climate by limiting to minimal amounts the emissions of CO__2 into the environment.

Among the top leading companies that have contributed to heavy amounts of CO__2 in the atmosphere in order of their volume of productions are Saudi Aramco, Chevron, Gazprom, ExxonMobil, National Iranian Oil Co., BP, Royal Dutch Shell, Coal India, Pemex, the Petroleos de Venezuela, the Petro China, Peabody Energy, the Abu Dhabi National Oil Co., the Conoca Philips, the Iraq National Oil Co., the Kuwait Petroleum Corp., Total SA, Sonatrach, BHP Billiton, and Petrobas, amongst many other companies that produce colossal amounts of carbon emissions into the atmosphere[1]. These companies have played a significant role in the concentrations of carbon in the environments in which they operate. Some have gone ahead to affect even the surrounding areas and inflict adverse outcomes to those regions. It is a work in progress to ensure that these companies comply with the required standards, both on legislative requirements and a scientific basis, to counter the effects of carbon in the atmosphere and reduce the effects of climate change on the economy and the people. These companies have contributed to a climate crisis that has left many parts of the world lavishing in the pains of the losses they have made. Biodiversity has been affected. The natural habitat of animals has been destroyed. In some areas, it has been entirely demolished to render the wildlife no habitat, causing their migrations. To the worse, and one of many reasons these companies have to review their approach to climate change, is due to humans ending up paying the price. With pollution, the interest of people in activity quickly shifts. They tend to focus on other places and other activities.

As indicated, among the companies that play a significant role in producing harmful emissions, the global polluters' list uses the company reported annual production of natural gas, coal, and oil. They then calculate the quantities of methane and carbon in the fuels produced by these companies emitted to the environment from extraction to the final use through the companies' supply chain. While some companies, as mentioned earlier, have taken responsibility for their emissions into the atmosphere, several other companies have maintained that they do not directly influence emission of methane and carbon to the atmosphere. They maintain that the emission should be blamed immensely on the oil consumers [TW19]. Interestingly, one of the significant oil-producing companies in America intends to be free from emitting carbon shortly. The company, Occidental Petroleum, is one of the USA's aggressive oil drillers and has contributed to colossal volumes of oil in the country. The company aims to remain carbon neutral[2] to help the government in its efforts to combat the effects of carbon and reduce the harms of climate change to society. The company leads in shale oil production in America and sits at position five by market valuation. It has started experimenting in various ways that can be used to capture the greenhouse emissions that are released by the company. The company intends to use a carbon capture technology to reduce carbon compounds by company into the atmosphere. It is the largest company in the United States to declare such a war against carbon emission, though it will face intense fights to accomplish this. This is a measure that must be employed by various companies that produce oil in America and the world at large to counter the effects of climate change on society. The goal will face several constraints but is attainable. The effects of climate emissions are worse than have been imagined or equated. It is upon the various divides to join hands in drafting various ways and techniques that will be used to reduce the effects of climate change on the people. These companies' regulation also needs to be made harsher to

[1] https://www.statista.com/statistics/1070949/worldwide-emission-from-oil-production-by-country/, accessed January, 2021

[2] https://edition.cnn.com/2019/03/20/investing/occidental-carbon-neutral-oil-shale/index.html, accessed January, 2021

make the companies start innovative ways to come up with measures to counter the widespread carbon emission[3]. As much as part of the carbon emission is linked to the final consumers, it is vital that the governments where these companies are situated to clarify the measures they have put in place to ensure the consumers are protected from carbon emissions. One of the applied strategies is shifting from fossil fuels in favor of a clean energy source, but it might take a few years to be fully accomplished.

2.3 Twitter

Micro-blogging is an emerging form of blogging that has taken over the world. It is a kind of blogging where a post consists of less than two hundred characters. The phenomenon has been adopted with the new on the go lifestyle that has taken over the world. The phenomenon allows the users to post small or more concise posts from different gadgets such as a mobile phone and a computer, and from any location in the world. Since the the users have less time for engagements, posts are made relatively shorter, in order to spend less time sending the messages [JSFT07]. The frequency of updates is also high as the users send the feedback more often due to the less time taken and the concise nature of the written texts. Therefore, the micro-bloggers would post more frequently than the usual bloggers that we see in other forms of blogging. The essential topics shared on micro-blogging are the daily happenings that comprise information sharing and daily activities. There are various micro-blogging platforms used in the present society, whose influence has impacted a more significant proportion of society. One of the common platforms that are in use is Twitter. Statistics show that Twitter is the eleventh most visited site[4] in the world. People prefer to use Twitter due to the ease of use, and as previously mentioned, one can send concise texts via Twitter and hold a conversation with a large group of people from different parts of the world. The replies and feedback on Twitter are almost immediate and enhance a continues post or conversation between the parties involved [Mur18]. Twitter is used for several purposes that include passing out a piece of information, communicating a vital message to the society, educating the public on some vital factors, for promotional services as seen by various advertising companies, for campaigns as adopted by political users, and for entertainment as employed by artists, amongst many other uses.

Twitter has had a tremendous influence on modern communication. It has been adopted worldwide. The platform is advantageous because it allows the users to spread messages about the occurring events quickly. Using twitter even spread news faster than even newsrooms. The platform has experienced fast growth since its launch in July 2006. The latest financial report from Twitter indicates that globally, it has over three hundred and thirty million monthly active users. The short messages sent by Twitter users are called tweets. On Twitter, the users can send tweets to each other, comprising the most two

[3] https://www.visualcapitalist.com/companies-carbon-emissions/, accessed January, 2021

[4] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, accessed March, 2020

hundred and eighty characters. Before September 2017, Twitter users were limited to using tweets that would only carry up to one hundred and forty characters. Twitter's move was to encourage users to continue using Twitter and enhance information sharing among users. It is a move that was subject to millions of Twitter users' discomfort, and the company decided to double the characters in order to ease the discomfort and increase the potential for influence of the tweets. People send tweets to each other to discuss vital factors that concern them and share their thoughts on various subjects and describe their current status with other people. To be actively involved in Twitter, and to always see tweets from other Twitter users, they could be following those Twitter users or being followed by them. The ability to gain more followers is a phenomenon that increases one's traffic on platforms like Twitter, which increases the engagement one has with their followers once they publish or send a tweet. As a result, it is easy for Twitter users to follow each other on Twitter or be followed. However, it is not always a reciprocated activity to be followed when you follow another Twitter user because the users can choose whom they intend to follow or whom to follow them. If a user follows another Twitter user, the follower will receive all the tweets sent or published by the Twitter user they have followed.

Besides the follow option and the tweet option, three other things make the use of Twitter even more enjoyable. One of the three things is called the hashtag. A hashtag is represented by a symbol of '#' and is essential in connecting various tweets that describe the same topic. The same topics are indicated by the words or phrases that bear the # symbol before them [WHMW11]. When a user feels interested in a specific topic, they can place a hashtag before the topic and find all the tweets concerning the topic they are interested in. The use of # has allowed users to follow their desired topics of interest quickly. The hashtag's adoption was first seen on Twitter, though due to its use and efficiency, it has been quickly adopted by other social media networks and platforms [Sma11]. The second property that makes the use of Twitter even more impressive is the availability of the retweet button. The users can share the tweets of other users that they follow by retweeting. It is the most common way that news and messages are spread out on Twitter. The third option used by Twitter users is the option to mention someone in a tweet. It is applied by inserting the '@' sign before the username of the other Twitter users. Twitter first developed the option to name another person on a tweet when there was still no private message option. It was the only way that a message intended for a specific audience could reach them directly. Twitter users all have profiles, which describe brief information about themselves, and the specific user can choose if they want their profile to be private or public [CMP].

2.4 Sentiment Analysis

Interest for natural language processing has existed for quite some time. In 1954, in New York, the Georgetown experiment was conducted. The experiment was a collaboration between the IBM and Georgetown University. The experiment was aimed at demonstrating the Russian-English translation. The machine translation system for the two languages was put under investigation by the two parties that came together for this experiment. There were up to six grammar rules for the Georgetown Experiment. Moreover, up to two hundred and fifty words were used in this experiment. However, the expected translation in the experiment was unsuccessful at last [Gar68]. Since then, the activity has attracted a lot of public interest, with many trying to work on it and examining the technique behind it. Furthermore, the experiment set a high expectation and dictated whatever must be done in the field of language translation to realize its success and how it can be developed and improved for the future. Sentiment analysis is defined as a process that is used to computationally identify, analyze, and then categorize opinions and ideas that have been addressed in a text or message on a communication platform. The aim of sentiment analysis in a text or a message is to determine if the opinion or idea or the contribution of a person in a conversation is neutral, negative, or positive. Due to the emergence of computers and the internet, there was a need to get in touch with communication and accurately understand the people's intended meaning in a conversation. The structured language of a computer was also a point worth giving attention to. The understanding of natural language as used in a structured computer language was a problem way back, and there were concerns about this. Thus, it was credible to derive a way in which people could easily understand coded natural language on a computer.

The art of natural language processing is a combination of various aspects such as computational linguistics, computer language, and artificial language. The various areas involved in NLP are crucial in understanding natural language and methods in which the human language could be processed or could be analyzed by computer devices [Cho20]. At the initial stages, the NLP was made in such a way that it would only operate by following a specific set of rules in the systems, for example, the rules of grammar [ID10], but all these were later altered. The alteration came about because following the initial set of rules for grammatical orientations in the system was very complicated than it was expected to be. Its use proved a nightmare for the users, who did not meet their targets with this mode. As a result, new modes were later on introduced to replace the initial design. In the new modes, the statistical methods and machine learning methods were introduced to replace natural language processing (NLP). The hand-produced rules for natural language could not match the new methods that were designed, for example, the machine learning methods. The latter's algorithms were focusing on the most common cases and thus were better than the NLP. Moreover, it was automatic, and it also had a more immense volume of input data as compared to the NLP, therefore enabling their precise nature, which was advantageous. In the present day, and with assistance from NLP, it is easy to determine which section of a communication or a sentence is crucial. Consequently, it has been made possible to sentiment the chosen section of a sentence or a communication. Some of the roles performed by the natural language processing systems include determining or identify semantic relationships in a sentence, parsing, stemming, and lemmatization.

Ideally, the logic behind natural language processing is to break the various parts of communication into bits and consequently understand the meanings brought out by each section of a statement or a sentence. Furthermore, it enables the users and those involved in the communication or the conversation to understand the relationship between the various parts of a sentence that have been stemmed or lemmatized. Currently, the uses of natural language processing is found in document summarizing and machine translation. It can also be found in text speech and speech to text conversations. Moreover, it is found in content categorization and sentiment analysis. As aforementioned, a person can analyze other people's opinions and sentiments concerning various services, events, different topics, organizations, and individuals through sentiment analysis. sentiment analysis is also referred to as subjectivity analysis, opinion mining, opinion extraction, or sentiment mining. The various terms all refer to a person's ability to critically analyze the parts of a text or speech in a conversation and provide what they translate or understand as the meaning or the message portrayed by the text's conveyor or the speech. Based on the various ways people may express their ideas and opinions about a particular subject in question or a conversation, they end up evaluating whether it is negative or positive and the extent of its negativity or positivity. They may also evaluate the neutrality of the specified part of speech or a text conveyed [NOMC11]. Social networks have had a direct influence on sentiment analysis in the current day. The influence is because social networks contain a dense communication database that has facilitated continuous communication among people in a group or in general, thus earning the attention to critically analyze a section of a speech or a text in a conversation. Different fields, such as the science fields, highly depend on people's ideas and opinions. They use these opinions to see how their scientific knowledge and discoveries have impacted society at large. Therefore, sentiment analysis is an essential part of day-to-day activities, especially for research purposes. They help identify the various areas that may need attention and to rectify where possible [Fel13].

CHAPTER 3

Literature Review

3.1 Using Twitter to Analyze Public Opinion

In the past many years, for the researchers to obtain information on how the public is generally doing, they had to conduct research by gathering information from random samples representing a population. In the 20th century, analysis of a population could be done via several means such as polling and conducting a survey of a section or the whole population. In the present day, controlled by the spread of information via social networks such as Twitter, people share a wide array of information via these social platforms. Whatever affects a person is shared on a chosen social platform like Instagram or Twitter. Information shared is almost the same as those that were collected via census and survey. People use social media and the Internet to express their thoughts and opinions, and ideas about a specific topic like diseases affecting a community, poverty in some geographical places, political dynamics concerning a region or a country, holding conferences about the progress their organizations and companies are making form different dynamics, such as development agendas, and financially. Consequently, there is a lot of public data that social media communications have made available for the public, making the information easily accessible and facilitated [NOBH+15]. Analyzing the information found in the social platforms has several advantages, such as a substantial mass presence in the time of such analysis. It is relatively cheaper as compared to the manual survey and polling tactics that were previously used by those in the early 20th century and before them, and which they fully relied upon. Furthermore, analyzing the data obtained or contained in the social platforms has proven to be more comfortable. It as well saves on time as the period needed to conduct the analysis is relatively lower than would have been taken in traditional surveys or polls [SRDBC15]. One disadvantage that comes with this mode of data analysis is that a researcher is limited to analyzing only the information that people who use these social platforms have decided to share concerning a specific topic. In the polls and traditional surveys, a researcher could easily employ or divulge other issues that the respondents might not have wanted to talk about relating to a topic previously. Among the various social platforms used to analyze people's opinions today, Twitter is one of those platforms that have stood out to be amid the best used to do such analysis.

Sentiment analysis is crucial in the automatic measurement of emotions in an online opinion, idea, or text, thus enabling online researchers to better their online data analysis modes. There has been continuous development and improvements in various algorithms used to automatically detect the neutrality, negativity, and positivity aspects presented in an online text. The developments have made it easier for social media platforms, and especially Twitter, in this scenario to successfully analyze the various aspects of an online opinion to evaluate the parts that depict a negative or a positive aspect related to a shared conversation between the Twitter users on a specific topic. There was a need to manage twitter to incorporate these milestone developments in their systems. Many people complained of too much negativity in their shared opinions concerning a particular topic [GHB09]. Sentiment analysis has been a significant improvement that has been welcomed by various divides using Twitter as an arena for their conversations regarding various aspects of life that affect them, such as politics, science, etc.

Sentiment detection of an idea or an opinion being aired on a discussion on Twitter can correlate with the topic if it is under investigation. For example, a negative text about racism can directly correlate to the aspects of racism that might have been detected or discussed in an opinion or a conversation. When a person attracts online users' attention on Twitter, especially by complaining of being subjected to racism or racially assaulted, they will, in most cases, receive messages of support and encouragement. Many people will rebuke this heinous act and offer opinions like 'racism has no place in the present world,' 'say no to racism,' ' racism is inhuman,' and 'all people are created equal irrespective of the color of their skins or region,' among many other supportive messages. These kinds of texts on Twitter on a subject like racism cannot be classified as negative sentiments. However, through sentiment analysis, they will be accredited as positive messages of encouragement to racism victims. In an adverse scenario, for example, in the same case of racism, a person complaining to have been racially abused may receive texts that are not a show of support to them, and which in many ways, may be contributing to the abuse of the person than help in healing. Others can even go ahead to racially abuse them by using different words that can be interpreted as racial slang. In such circumstances, the sentiment analysis will reveal that these conversations contain negativity in them. If necessary, through a legal process, corrective measures might be taken against the people who did so, depending on the country. Other negative sentiments might not be too serious like the case of racism, for example, in the case of a football match, a Twitter user may say they disagreed with the decisions made by the match officials, and that even though may look like a negative comment against the decisions taken by the match officials, are a fitting room of communication and does not require any sanctions. Therefore, the above examples only indicate the extent and importance of using sentiment analysis on Twitter and how much it has helped improve the online conversation. The users of Twitter also vary, the audience originates from various nations globally, and this has made Twitter have an enormous figure of texts or posts every day. Thus, sentiment analysis is an effective approach that must be continually used even in the future to build an even sizeable online presence, where individuals can freely air their views without any abuse. Sentiment analysis is also vital in knowing how good or positive people talk about an idea or a project by an organization, especially if it directly influence them, such as from NGO's or the government [KWM11].

3.2 Stock market prediction using Twitter sentiment analysis

The opinions or sentiments that are made by an investor has a direct impact on the market. It can be either a positive impact or a negative impact on the market. However, the question is not how the sentiment will influence the current market, but rather how that sentiment made by the investor can be measured. The suggestion of measuring sentiment analysis had developed in the first half of the 20th century. However, there have been tremendous developments in the recent past that have involved research on the various ways that can be used to measure sentiments of investors in the stock market. Using various research on the subject, two main ways have been so far identified that can be used to measure sentiments in the stock market. In the first method, sentiments are measured from surveys.

In contrast, in the second method, sentiments are measured from the correlation of the objective variable that had a direct relation with the sentiments made by the investor. However, both modes of measurement have various constraints and limitations that come with their use, and as a result, none of them can be efficiently used to determine the measure of sentiment. However, various ways have been used on Twitter to examine the direction that a stock market is taking, or even the direction people perceive it has taken. The opinions here are crucial to make a substantive conclusion regarding the stock market and the progress that the investors expect in the years to come.

Researchers have tried their best and conducted several tests that have indicated that it is indeed possible to determine the direction that a stock market will be taking in the next few weeks, months, or years to come. In the past few years, significant progress has been made on sentiment tracking techniques that have facilitated predictable trading dynamics in the stock markets. Bollen et al. conducted one such successful researches that proved it is possible for a stock market direction to be easily predicted by the sentiment analysis. Their research argues that public opinion, as was expressed in the public tweets that they examined, had a positive or a direct correlation with the direction that the stock market was going to take in the future [BMZ11]. The opinions were also vital in determining how the specified stock market would perform in the future or how it was already performing. For example, the suggestions by this study indicated that if the people or the public, through their tweets, said mostly that they are receiving low-security services or bonds with unreasonable terms from a specific stock market, the correlation in this scenario meant that the stock market was likely to be performing poorly as the business was not attracting more investors. If the comments through the tweets were positive, such as expressing satisfaction with the services rendered by a stock market, it was more likely, or in reality, their fact is that the stock market that was being discussed was exemplary performing better and would most likely attract even more investors soon, which means that it will have a positive growth[1].

In the past, people mainly depended on news to predict the stock market, and on the belief that many carried. Nowadays, people still rely on news and rumors to predict the stock market. However, the news and rumors are spread much faster through the use of internet and technology. Twitter being one of the platforms that are used for fast spread of news, rumors, and the ability to extract public opinion, has had a significant impact on the stock market [RS$^+$12].

3.3 Top oil producers stock market analysis

Like many other companies in the world, the oil producers too always have a problem determining the direction that the business would take in the future. The suspense creates much anxiety for the top oil producers as they sometimes may not be able to figure out the direct price their businesses would hold in the next month, week, year, or so. Even as much as the top oil producers have tried to figure out the best path the business would take, they are affected by unprecedented natural occurrences that deter their plans the profits they intended to make at last. For example, the OPEC countries had set an agreement among themselves to reduce the quantity of oil sales among themselves before the Corona Virus outbreak. Later on, the outbreak would hamper their plans of storing the oil reserves for better prices, and following a request by Russia, they heeded the call and started selling big volumes of oil. The problem is that these unprecedented natural calamities affect the projected price, and in most cases, they sell at relatively significantly lower prices. Stock markets are essential aspects that must be keenly examined by these oil-producing countries and companies, so they may make a solution that would ensure they always know the direction that the market would take. The oil-producing companies have mostly hired professional individuals, who are economics and financial intellectuals, who successfully, in most cases, predict the future of the market of the oil they produce. The individuals can use complex computerized calculations in determining the prices of oil in the near future or, at times, revert to fundamental ideas about oil price fluctuations and the causes of these fluctuations.

One of the basic ideas that these companies use to predict the price is the people's descriptions or prospected customers. These descriptions describe the supply and demand chain. If the demand, which is most important in predicting the price value here, is low, then the oil-producing companies would not increase their oil supply. It is understandably clear that demand is a significant influence of the price of oil in any market. For a stock market, too, low oil demand would mean that the oil prices and its products at the stock

[1] https://towardsdatascience.com/stock-prediction-using-twitter-e432b35e14bd, accessed January, 2021

market will be generally low. Therefore, demand is one of the markets attributes that the oil-producing companies use to predict oil prices in the future or at the stock markets. Demand means a high need for oil; people need to use oil and its products. It also means that there is an insufficient oil reserve already existing among the users. With higher demand, it inevitable that if more oil is not provided or supplied, the customers will run short of the oil. So here are two options that the oil-producing countries choose to raise the price of their oils. High demand generally means that they will have to increase its oil supply to satisfy the need in the market [SWZ09].

3.4 The impact of top oil producers in climate change and global warming

Top oil producers have contributed immensely to the increased destruction of the ozone layer, thus contributing to climate change and global warming. Arguably, oil is the least or the last element that man needs to survive. Interestingly, it has been in very high demand over the past few years than many other crucial elements. The oil demand has consequently increased the chances of climate change. The top oil-producing companies drill oil in their natural reserves and even burn them. The burning of oil releases enormous masses of gas carbon dioxide into the natural ecosystem. When colossal volumes of carbon dioxide are released into the atmosphere, they contribute to global warming. It has even been warned that to prevent further destruction, and the oil producers cannot afford to burn even a third of the existing oil reserves worldwide. More burning has increased the quantity of carbon dioxide gas in the atmosphere, contributing to greenhouse effects in the areas they are burned and the surrounding areas[2]. Top oil producers have also contributed to the production of methane into the atmosphere. Like carbon dioxide, methane has presented very harmful and dangerous effects to the climate, threatening the biodiversity of nature and its elements. Methane is even more dangerous than carbon dioxide. The greenhouse effect produced by methane is almost eighty times more potent than carbon dioxide in over a period of twenty years, and it is estimated to be responsible for over twenty-five percent of the contributors of global warming, as per the data by the United Nations Environment Program. The increased oil production in the past two centuries has doubled, which means that there has been a direct effect from the top oil producers on climate change and global warming. The top oil producers have also dumped huge masses of deposits of oil and their products in the sea and other water bodies, which has affected the aquatic ecosystem. Many nations have taken action against the top oil producers, ensuring they use every possible mechanism to limit climate change effects. Climate change has become a significant threat in recent years, and many governments are striving to ensure that they curb the dangers that may be presented by the condition.

[2] https://www.greenpeace.org/usa/global-warming/issues/oil/, accessed January, 2021

Pollution is a significant area of concern by the top oil producers [KMPH09]. It is a factor that must be put into consideration before releasing methane and carbon dioxide into the natural ecosystem. The top oil-producing companies cannot efficiently manage both the chemical and the physical aspects of their productions, and consequently release an uncontrolled amount into the atmosphere. Burning these fuels must be limited if the oil producers must act against climate change and global warming. It is a collective effort between the top oil producers and the government, as well as the people, to ensure that they work collaboratively to ensure there is a limited chance of global warming in the future. What is essential is not the number of hazardous gases in the atmosphere, but which measures all the stakeholders involved adopt to ensure a reduced climate change and global warming.

CHAPTER 4

Methodology

The section describes the system that we have setup from data finding, data collection and trained models. We highlight the data pre-processing and VADER sentiment analysis model leading to feature engineering of highly correlated predictors. Finally, we establish the relevant Machine Learning (ML) model for regressive forecasting of the stock prices.

4.1 System Components

4.2 Data Collection

The data was collected from two different sources first Twitter using Twitter API to collect tweets with highlighted hash tags such as #climatechange, #globalwarming, #carbonproducers, #endfossilfuel, #CO2rise, #sealevelrise, #exxon, #bp_oil, #chevron. The purpose of the tweets collected is to amass a considerable amount of time-scaled tweets which is representative of worldwide opinion on intersecting topics related to climate change and oil producers and carbon producers in the world. The other aspect of data is the stock prices of large conglomerates which are the world's largest carbon producers. These stocks data were collected using Yahoo Finance API.

Tweepy API

API Description Tweepy is a MIT Licensed python library to call Twitter API to extract tweets from twitter. There are different endpoints to extract both old and real-time tweets from Twitter. Tweepy helps access information, tweets, on the hashtags we have enlisted related to carbon producers and carbon emissions. We use "search" API endpoint in Tweepy to collect tweets.

Tweet ID	Tweet Text
1279200392673460225	Make The Climate A Priority Again,' Says Germany's Student Activist Neubauer https://t.co/IWJmEhdf6f #ClimateChange #GlobalWarming #ActOnClimate #GreenNewDeal #ClimateAction #ClimateChangeIsReal #EndFossilFuelAddiction #StopTheGreed #ProtectOurWater #FridaysForFuture #Fam46
1279200148611125250	Climate change threat to tropical plants
1278181449938563072	Global temperature anomalies since 1900 The best time to avoid run-away #ClimateChange was 20 year ago The second best...
1279476620324147200	415.23 parts per million (ppm) #CO2 in the atmosphere July 3, 2020 Up from 413.43 ppm a year ago

Table 4.1: "Sample of Tweet Texts Downloaded from Twitter API"

Data Collection We collected the data using tweepy cursor between the designated time. This will help us collect useful full sized tweets and save them.

Sample Data The sample tweets we got from our tweets collection are shown in Table 4.1. The sample tweets are reflective of people's opinion towards the current climate change, fossil fuel consumption and CO2 rise. The tweets, upon manual inspection, seems to be either informative facts, criticism of fossil fuel usage or producers or praise of actions taken towards mitigation of climate change by different entities.

Yahoo Finance API

API Description Yahoo Finance API is a public API which requires no authentication to get finance data but is rate limited, meaning the number of requests should be delayed to prevent Internet Protocol (IP) Address being banned. A ticket module gets the data from the New York Stock Exchange (NYSE) data from 9:30 am to 4:30 pm.

Data Collection Data is acquired from a ticker module to access the ticker data. The ticker is the frequency (int time) to obtain the finance data. Each data point should be within 2000 requests per IP or 48000 requests per day. The Data can be collected for different oil producers using their Stock ID such as EOX for Exxon, BP for BP Oil etc, which are few of the leading Oil Producers in the world. The sample data for Exxon using Yahoo Finance API, Python Library yFinance, is shown in Table 4.2.

Index	Open	High	Low	Close	Adj Close	Volume	Date
1	89.279999	89.709999	89.059998	89.565002	89.565002	149066	2020-07-01 13:30:00
2	89.610001	89.820000	89.519997	89.610001	89.610001	24202	2020-07-01 13:31:00
3	89.592499	89.739998	89.410004	89.690002	89.690002	19647	2020-07-01 13:32:00
4	89.665001	89.809998	89.570000	89.690102	89.690102	33122	2020-07-01 13:33:00
5	89.779999	89.979897	89.730003	89.955002	89.955002	32379	2020-07-01 13:34:00

Table 4.2: Sample Data of Stock Prices from New York Stock Exchange for Exxon retrieved by Yahoo Finance API

4.3 Data Clean Up

The collected twwets are text and numerical data which requires extensive feature engineering and data pre-processing to make them suitable for model building. The text data tweets were pre-processed using TextBlob Library in Python to clean non text characters, removing common stopwords without any semantic meanings and stemming, which means removing any higher order of inflations in verbs to their root meaning (Verb-1). Similarly the data collection from Yahoo Finance API were plotted and checked for consistency. Upon inspection we found that the prices were enlisted only from time 9:30 AM to 4:30 PM based on New York Stock Exchange (NYSE) opening and closing time. The data were collected based on time order and merged using Pandas DataFrame using TimeSeries data on DateTime Index.

Tweets Pre-processing The full tweets text obtained from twitter are cleared for non text characters such as whitespaces and punctuation. The stop words as listed in NLTK corpus[1] are removed from the sentences. URLs are removed from the full texts and retweets which begins from RT are removed as well. We are focusing on obtaining a clean version of the original tweets. Using TextBlob we used stemming and lemmatization within the sentences to transfer each token/word into its root form.

VADER Sentiment Analysis of Tweets VADER[2] (Valence Aware Dictionary for Sentiment Reasoning) is a sentiment analysis model that takes into account sentiment also called polarity and also emotions. It uses lexical features of sentiment and emotions to understand polar opinions and complex semantics. The sentiment is accumulated based on minutes and summed up. Each minute is highlighted by total sentiment collected within the given minute.

Stock Price Pre-processing The stock prices from Yahoo Finance have stock prices available for each day from 9:30 am to 4 :30 pm in Easter Time Zone. The first step in

[1] https://www.nltk.org/api/nltk.corpus.html, accessed July, 2020

[2] https://pypi.org/project/vaderSentiment/, accessed July, 2020

pre-processing includes conversion to Stock Prices into UTC timezone. Using Pandas data-frame library in Python we synchronize the time between tweets and stock prices using pytz library in Python.

```
tweets['Datetime'] = pd.to_datetime(tweets['created_at']).
tz_convert("UTC")
exxon["date_rd"] = (exxon.index.tz_convert("UTC")).
strftime('%Y-%m-%d %H:%M:%S')
```

4.4 Statistical Models

Predictors Correlation Predictor in our model is only sentiment as we are inspecting the correlation between the sentiment of Tweets and Stock prices. Hence we compute Pearson Co-efficient between the predictors i.e. sum of compounded polarity of the tweets given by VADER sentiment analysis which forms our main predictor. The correlation of between the stock prices data (Open, Close, High, Low and Volume) helps us find the appropriate response variable with higher correlation to the predictor.

Moving Averages of Sentiment Moving averages are statistical method is generally used in stock trading to calculate the influence of previous values of a data-point. Generally these data-points are Open, Close, High, Low or Volume in Stock Prices. However, we are using sum total of sentiment score over a time period as data point to calculate the backward-looking indicator to predict the stock prices. We inspect the co-relation between the different types of moving averages and select the predictor which has highest correlation with Stock Prices as shown in Figure 4.1. The two types of indicators are:

Moving Average Sentiment : Simple Moving Average is the sum of sentiment for k period divided by k.

$$M.A = \frac{S_n + S_{n-1} + ... + S_{n-k}}{k}$$

where,

S_n = Delayed Sentiment Sum for n^{th} Stock Price period

k = time period for moving average calculation in minutes

Exponential Moving Average : Exponential Moving average is exponentially weighted moving average of sentiment for k period divided by k.

$$E.M.A_n = S_n + \frac{2}{n+1} + E.M.A_{n-1}\frac{1-n}{n+1}$$

where, S_n = Delayed Sentiment Sum for n^{th} Stock Price period

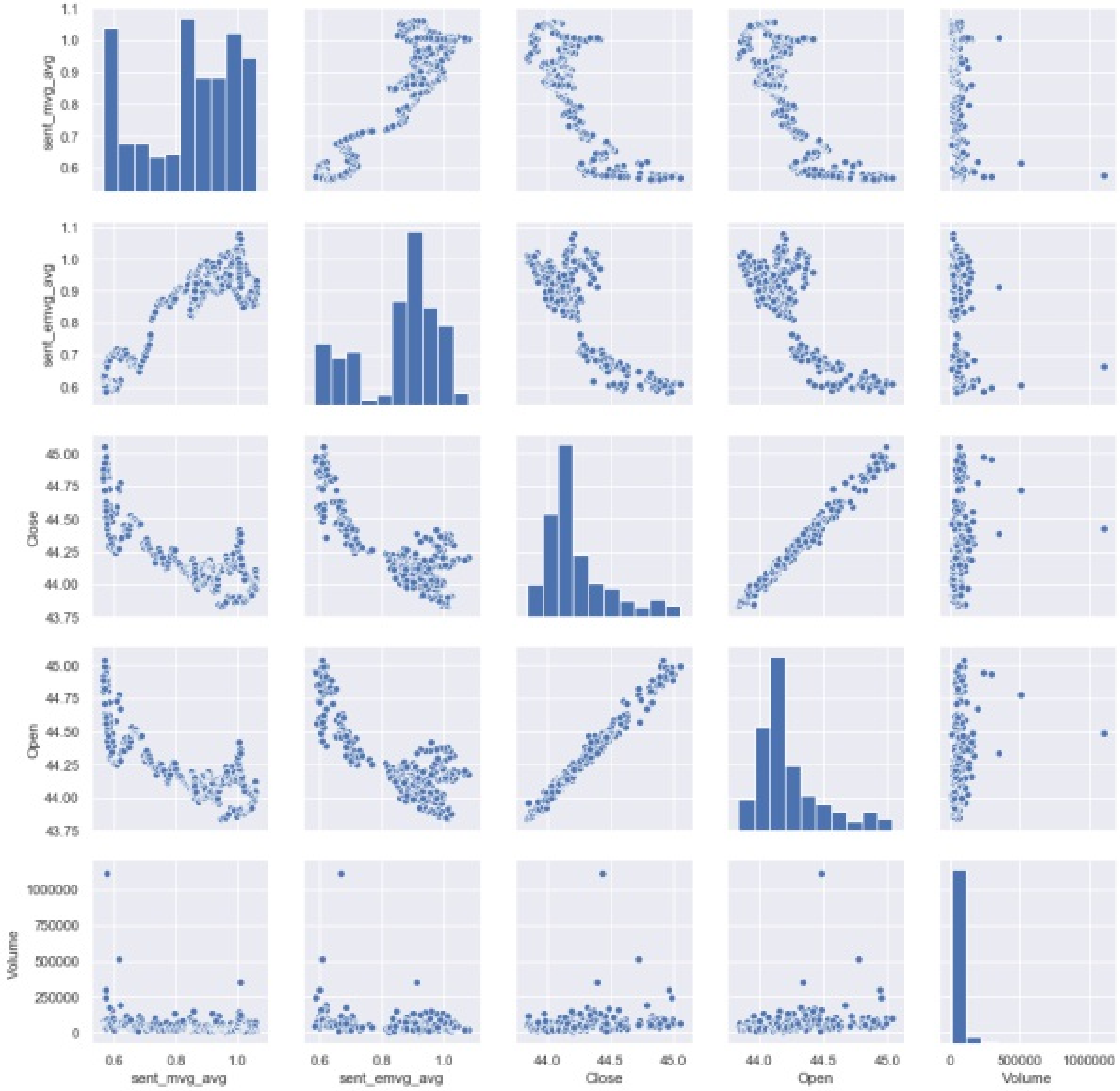

Figure 4.1: A Parplot to highlight Correlation between predictors

$E.M.A_{n-1}$ = Exponential Moving Average of Sentiment Sum $n - 1^{th}$ Stock Price Period

k = time period for moving average calculation in minutes

As shown in Figure 4.1, Both Moving averages are negatively correlated with the stock prices. We use Pearson's correlation co-efficient to find which of the predictors are highly correlated for a given k such that $k = 5, 10, 20, 30, 60, 120$ in minutes. We only opt Close values because we are trying to predict the Closing stock prices at the end of every minute. From the Table 4.3, we can clearly see that for 120 minutes of EMA computed

Predictors	Close	k-th Period
Sentiment Score	0.07768	10
S.M.A	0.13242	10
E.M.A	0.21458	10
Sentiment Score	0.1229	20
S.M.A	0.30736	20
E.M.A	0.3272	20
Sentiment Score	0.1229	20
S.M.A	0.30736	20
E.M.A	0.3272	20
Sentiment Score	0.2457	120
S.M.A	0.6873	120
E.M.A	0.7039	120

Table 4.3: Pearson Correlation Coefficient for Predictors and Close

the sentiment sum has more than 0.70 Correlation. Thus we choose K as 120 and use Sentiment E.M.A as predictor to predict sentiment scores.

Regression Model Building These sentiment predictors M.A and E.M.A are then used as predictors to predict Close stock prices. This is supervised Machine Learning approach where we have the predictors and target values. Target values are the stock prices which are predicted on the basis of highly correlated predictors. This method is called Regression. We use Regression models for prediction of the outcome variable. The predictor is then fed into a regression model for prediction of Stock Prices. The Regression functions used are:

Bayesian Ridge Regression BRR is used to predict the outcome based on probabilistic regression with prior probability of ω, which is computed by Gaussian function. The probabilistic regression model is given by:

$$P(\omega|\lambda) = \mathcal{N}(\omega|0, \lambda^{-1}I_p)$$

where,

λ, ω are priors of a Gamma Function

This is a Normal Distribution representation where mean is given by prior $\omega|0$ and variance is $\lambda^{-1}I_p$. The parameters are estimated by Maximised Log-Likelihood Function given as

$$\underset{\beta}{\arg\min} \ \tfrac{1}{\sigma^2}||y - X\beta||_2^2 + \tfrac{1}{\tau^2}||\beta||_2^2$$

where,

y is the Close Stock Price

X is the vector containing sentiment

σ^2 is the variance closer estimated as the sample variable

$beta$ is the regression co-efficient estimated by Bayesian Approximation

τ is ridge regression penalty

Support Vector Regression SVR is a non parametric method of estimating an outcome variable from etimators using a kernel function. This method works by creating support vectors and by mapping predictor X in higher dimensional space. Support Vector Machines are based on class of hyperplanes.

$(w \cdot x) + b, w \subset \mathcal{R}^N, b \subset \mathcal{R}$ .... (a)

The optimal hyperplanes defined by Equation a above predict the values based on corresponding decision function.

$f(x) = sign(w \cdot x) + b$

Support Vector Models are mathematically simple for analysis and interpretation. Even though a non-linear kernel allows us to map the dataset into higher dimensional space, it can be interpreted as a linear model in the said higher dimensional space. It allows heuristic methods, as linear or non-linear methods to achieve different models for both classification and regression tasks. Support Vector methods can also be used, for this reason, as time-series model using non-linear kernels like Radial Basis Function (RBF).

$k(x, y) = exp(\frac{-x||x-y||^2}{s\sigma^2})$

Support Vector Method support multiple non-linear kernels such as quadratic, polynomial, sigmoid and a custom kernels which are a form of L2-regularised solution to a linear model which can be represented in higher dimensional space.

Kernel Ridge Regression The KRR like the SVR uses a non-linear linear kernel to solve a larger dimensional features. The Kernels can be linear as well, however the using a linear kernel requires more computational time and the convergence is sub-optimal for Univariate regression problems. It is a non parameteric regression method which uses squared error as loss function with l2 regularization. This method uses grid search which is faster that Support Vectors. However, Kernel Ridge Regression tends to be sparse and less efficient on larger dimensional dataset. The Kernel Ridge regression minimises the following Squared cost

$\mathbf{C}(w) = \frac{1}{2}\Sigma(y_i - (w)^T x_i)^2$

Using L2 Normalization, we take the first derivative of the equation in-order to find the function to be minimised as shown below:

$\underset{i}{\Sigma}(y_i - w^T x_i)x_i = \lambda w \xrightarrow{w}= \{\lambda I + \underset{i}{\Sigma} x_i x^T\}^{-1}\{\underset{j}{\Sigma} y_j x_j\}$

In Kernel Ridge Regression, the data with the feature vector are replace with ϕ_i which represent the features in higher dimensions such that $x_i \to \phi_i = \phi(x_i)$

$$y = w^T \phi(x) = y(\phi^T \phi + \lambda I_n)$$

Now projecting the solution in to w dimension we get predicted y,

$$y = y(K + \lambda I_n)^{-1} \mathcal{K}(x)$$

Kernel Ridge Regression provides edge over Support Vector Machines. In Support Vector Models sparseness is created due to constraints set by hyper-planes are unequal. As a result those dimensional representations are inconsequential which leads to wrong interpretation. There is no such effect in Kernel Ridge Regression.

Model Training and Validation According to CRISP-DM, after understanding the problem and curating the data we train the model for Bayesian Ridge Regression, Support Vector Regression and Kernel Rigde Regression with default parameters using Scikit-Learn[3] Library. The model is trained into a 3:1 split for training and testing dataset. The Test dataset, also known as validation dataset, is used to measure the performance of the model. The model performance for Regression model is computed by Mean Squared Error and R-Squared. Mean Squared Error computes the variance and baisness in the model. This gives the measure of variance between predicted values and true values.

$$M.S.E = Bais^2 + Variance$$

R-Squared Error gives measure of proportion of the variance explained by the regression model for the test dataset using predicted and ground truth. The value of R-Squared ranges from 0 to 1. 1 meaning 100% of the variance is explained by the model which is ideal case for a model and R-squared value of 0 means that none of the variance withing the response is explained by the model with given predictors.

Hence, we opt for the best model using Model training and we select the best model. Then we tune the model parameters to obtain the best performance of the model using Cross Validation. We use K-Fold Cross validation of the model to obtain best parameters for the regression model and train the model using those parameters to predict stock prices.

[3] https://scikit-learn.org/stable/supervised$_learning.html, accessedMarch,$ 2020

CHAPTER 5

Implementation

5.1 System Components

This section describes implementation of the Stock Prediction for Oil Corporations based on the Tweet data. The system flow diagram is shown in Figure: 5.1. The system comprises of following components and sub components:

1. Data Pre-Processing
   - Data Collection
   - Sentiment Analysis of Tweets within Time
   - Moving Averages
2. Training Data Extraction
3. Regression Models
   - Model Cross Validation
   - Best Fit Model

5.2 Data Pre-Processing

Data Pre-processing involves cleaning the raw format of the data into a ordered set that can be used by machine learning models to perform predictions. The source of data for our systems are Tweepy API and Yahoo Finance API which serve data in JSON (JavaScript Object Notation) format. We collect these dataset, clean out the redundancy and missing values using Pandas dataframe library in Python and prepare the dataset.

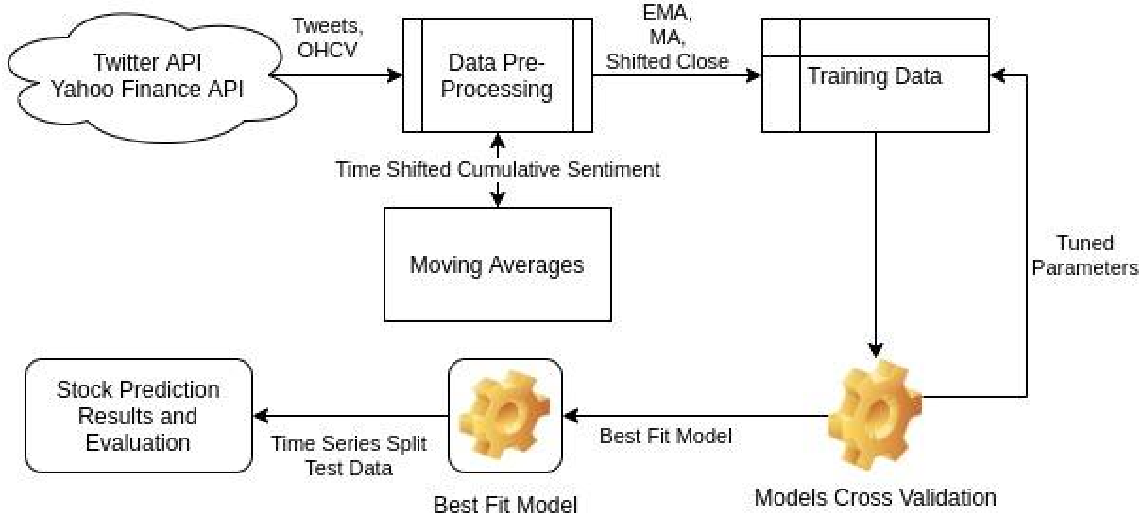

Figure 5.1: A System Flow Diagram of the System

Pandas Dataframe allows us to orient JSON data into database-like tabular structure which supports Text, Categorical and Numerical dataset. The data obtained from API endpoints such as tweets and stocks data can be tallied by juxtaposing the dataset based on Date-time indices available for both dataset. For further use, we compute tweet sentiment, moving averages and aligning columns using time-shifted data which gives best correlations.

5.2.1 Data Collection

As mentioned before data is collected using Twitter Search API. Twitter Search API facilitates collection of tweets globally that facilitates us to filter the tweets based in “query” argument. We enlist the hashtags associated with the carbon producers and popular hashtags associated with climate change, carbon emission and environment change such as #climatechange, #globalwarming, #carbonproducers, #endfossilfuel, #CO2, #CH4, #surfacetemperaturerise #sealevelrise, #exxon, #bp_oil, #chevron. These hashtags enabled us to collect opinions in the twitter feed which are defaulted to English. The API requires authentication using OAuth2 authentication which requires us to validate API access using access and secret tokens provided to specific twitter user from Twitter Developers account, Twitter Dev for short. In total we have collected 936000 tweets out of which, 230000 have been selected for use. The reason why only 230000 have been selected is because the stock market working hours are from 9 to 4 and the tweets that impact the stock market prices, from our finds are usually tweets of up to 2 hours before. So, we selected tweets only from 7am until 4pm, excluding the weekends as well. The snippet below shows us how tweets are collected in the code base.

Code:

```
tweets = tweepy.Cursor(api_handle.search, q=query,
         tweet_mode='extended', lang="en", show_user=True,
         since='2020-07-01', until='2020-08-30')
```

The API handle is obtained after user authentication after providing the said User Access Tokens. We can provide the data rage from *since* and *until* arguments which are set to start of July to end of August 2020. The extended mode allows us to collect full tweets and avoid the 160 characters restricted incomplete tweets, in-order to get correct sentiment polarity.

The stocks data is collected using Yahoo Finance API which provides data of stocks and financial news on companies in NYSE (New York Stock Exchange). The data we collect are OHLCV (Open High Low Close and Volume). The following snippet extracts OHLCV data from Yahoo Finance using API. Here the stock market code, Ticker Code, is the main argument for acquiring any stock data from the respective company, for example, Exxon has XOM as their ticker code in NYSE. The next arguments are the historical data from starting date to end date which will be start of July to end of August and over the interval of 1 minute. This collects all the OHLCV data from Yahoo finance API on Exxon for every minute.

Code:

```
import yfinance as yf
exxon = yf.download('XOM',start='2020-07-01',end='2020-08-31',
         interval='1m')
```

After collection of data, tweets and stock data needs to be organized and formatted. This ingests the high volume of data acquired from the Cloud APIs and then indexes them to prepare training data. We needed to implement this pipeline as tweets are raw text and to predict sentiment, find correlation between dependent variables and finally compute moving exponential averages hinges upon correct sentiment prediction for which the tweets text much me well formatted. The Stock Prices from New York Stock Exchange data were published only during work hours which required us to align the tweets data and stock price data for the same time zone.

5.2.2 Sentiment Analysis of Tweets within Time

The collected tweets have sentiment score for each Tweet. Since stock market price is a time-series data, in-order to use sentiment as time-series predictor we compute sentiment across the time period i.e over the time span of a minute. This function plugs the total sentiment across one minute time frame using Pandas Data-frame. This function generates the input predictor for Regressive Machine Learning model. The code snippet shows the implementation of computing Sentiment Analysis of Tweets with Time:

Code:

```
from textblob import TextBlob
blob = TextBlob(tweet_text)
sentiment = blob.sentiment.polarity
```

The method uses Textblob for sentiment analysis prediction. Textblob is an open-source Python library for Natural Language processing for text dataset. The blob type object loads the tweet text which allows user to access various NLP functionalities such as Sentiment, Parts of Speech Tagging, Noun Phrases, Dependency Trees etc. The sentiment of the tweet is computed using polarity score of the text given by VADER (Valence Aware Dictionary for Sentiment Reasoning) sentiment analysis model which is available in Textblob library. The polarity score for a given text (tweet) ranges from -1 to 1 where -1 represents extremely negative sentiment, closer to zero implies neutral sentiment and 1 represents extremely positive sentiment.

Furthermore, we proceed to compute time-shifted sentiment within time. This method aligns the sentiment over time to highly correlated Close value of the stock implying that overall sentiment within a time causes the stock prices to fluctuate as the result of closing price for the given time. We sum the sentiment of the tweets across the given time-frame (hourly) as measure of the overall sentiment measure within the time as it represents both frequency and sum of the tweets. This cumulative value is then taken as the intended indicator, similar to OHLCV Stock data price for a ticker.

Code:

```
tweets['date_rd'] = (tweets.index.round('H')).tz_convert("UTC")
                    .strftime('%Y-%m-%d %H:%M:%S')
exxon["date_rd"] = exxon.index.tz_convert("UTC"))
                    .strftime('%Y-%m-%d %H:%M:%S')
combined_data = exxon.merge(tweets,"inner",on=["date_rd"])
```

This snippet of Python code shows us how the dates are converted to UTC format. Then by the process of inner-join on the date-time format we create one-to-one relation between stock data and respective tweets within that time index. This allows us to perform time-wise aggregation of sentiment and stock prices for further computation.

5.2.3 Moving Averages

Moving averages are generally used as Stock Trading indicator. As discussed, we compute moving average and exponential moving averages for sentiment within the time values computed from the tweets related to the hashtags regarding oil producers and climate change. The codes below compute moving average and exponential moving average using rolling window function in Pandas library in Python.

Code:

```
sentiment_min["sent_mvg_avg"] = sentiment_min['sentiment']
    .rolling(window='120',
    win_type='triang',
    min_periods=3).mean()
sentiment_min["sent_emvg_avg"] = sentiment_min['sentiment']
    .ewm(span='120',
    adjust=False).mean()
```

The moving average is computed as rolling value based on minute-wise OHLCV data collected from Yahoo Finance API. The sentiment moving average is computed for the span of 120 minutes, as well as the exponential moving average is computed in span of 120 minutes which has best correlation with Close Price of stocks, as shown by correlation table. This implies that Moving averages of sentiment within time of tweets before two hours correlate to the current stock price.

5.2.4 Training Data Extraction

Dataset are saved as Comma Seperated Values (CSV) files for all of the tweet dataset containing the enlisted hashtags with its sentiment values and Financial data i.e Stock Prices. The purpose of this process is to extract data-values from saved CSVs. Pandas, a Python dataframe library, loads huge files and allows for seamless processing of the dataset. The library creates the reference points for chunks of data without taking up too much of system memory. Pandas data-frame allows CSV files to perform SQL-like operations and perform mathematical computations and Graphical visualisation. A shifted Close value is also appended as the column onto the dataset, during pre-processing, onto the training dataset which is shifted by 60 minutes which gives us the one hour before Closing price as predictor which is then fed into time-series regressive model to predict future Closing price.

The Training dataset is saved and then passed into computational models. The loaded dataset are split into TimeSeries split, which is a functional way of creating Training and Testing data for evaluation of the regression models under Model Cross Validation.

5.3 Regression Models

This model ingests the values computed as Moving Averages, Exponential Moving Averages and Shifted Close predictor to predict the stock prices. This unit contains regression models using training dataset and new dataset that are fetched from the system. This model predicts the stock prices for next minute based on the current dataset i.e. sum sentiment of tweets delayed by K-period as discovered in Section 4.2.

5.3.1 Model Cross Validation

The Model Cross Validation method firstly computes a model from our current set of Regression algorithms namely Support Vector Regression, Bayesian Ridge Regression and Kernel Ridge Regression that provides the best fit for the current training dataset at hand. The model with optimum performance, measured in R-squared and Mean Squared Errors is then used for Model Cross Validation to asses the parameters for tuning. This process also enables the identification of the best algorithm suited for the current dataset. The following code shows Exxon stock prediction Model Cross validation Method

Code:

```
model1 = linear_model.BayesianRidge()
model2 = SVR(kernel='rbf', gamma=10)
model3 = KernelRidge(kernel='rbf', alpha=0.1, gamma=1)
def score_plot(model,dfs,title):
    model.fit(dfs[["sent_emvg_avg","sent_mvg_avg"]],dfs["close"])
    score = model.score(dfs[["sent_emvg_avg","sent_mvg_avg"]][-20:],
             dfs["close"][-20:])
    print ("Score",score)
    predicted = model.predict(dfs[["sent_emvg_avg","sent_mvg_avg"]])
    print ("R Squared",r2_score(dfs.close,predicted))

model2title = {
    "Bayesian Ridge Regression":model1,
    "Support Vector Regression": model2,
    "Kernel Ridge Regression":model3
}


date_range = " July 1 to Aug 31, 2020"
for title,model in model2title.items():
    score_plot(model,train_data,title+date_range)
```

The codes above define Bayesian Ridge Regression, Support Vector Regression (SVR) and Kernel Ridge regression with Radial Basis (RBF) kernel with respective alpha and gamma value for the Kernel Function being used. The score_plot function fits the data with relevant predictors which are moving averages and predicts the model fit scores and R-squared score in-order to ascertain the best model for the current Oil Producer data. Based on the output from the score_plot function, we ascertain the suited algorithm and proceed to Time-Series Wise Cross Validation of the selected algorithm.

Code:

```
from sklearn.model_selection import TimeSeriesSplit
splits = TimeSeriesSplit(n_splits='6')
for train_index, test_index in splits.split(X):
    train_model(train,y_train,test,y_test)
```

The above code snippet is a summary of how TimeSeriesSplit is used to perform 6-Fold Cross Validation of the dataset. The value of n_splits is determined by iterative execution of this process. The split data is then trained using the best appointed model defined under train_model function which serves the computed models and their performance measured is compared to find the Best Fit Model along with the parameters for training the best fitted model.

5.3.2 Best Fit Model

This process selects the model with highest R-squared value and lowest mean squared error from the Model Cross Validation process. The selected model is deemed to be the best performing mode which is trained on inclusion of the only relevant data for the training dataset. This means only predictors such as Moving Averages and Shifted Close values which are influential in prediction of best stock prices are determined. This model incorporates the opted best performing model ranked by K-Fold Cross validation across opted Regression model. This model is saved as the best of the Regression Models which is then loaded in system memory. This process can acquire new tweets from Tweepy API for next hour. This Best fit model can be used in prediction of next hour closing stock price, after computation of Sentiment Moving Averages for current set of relevant tweets available.

CHAPTER 6

Results and Discussion

A time-series prediction is mapping of current and previous observation. A multi-step head propagation of time-series data propagates errors further ahead which causes high margin or error. For this reason we have used a correlation parameter for sentiment as corrective dependent variable to predict the stock prices.

6.1 Regression Results

6.1.1 Exxon:

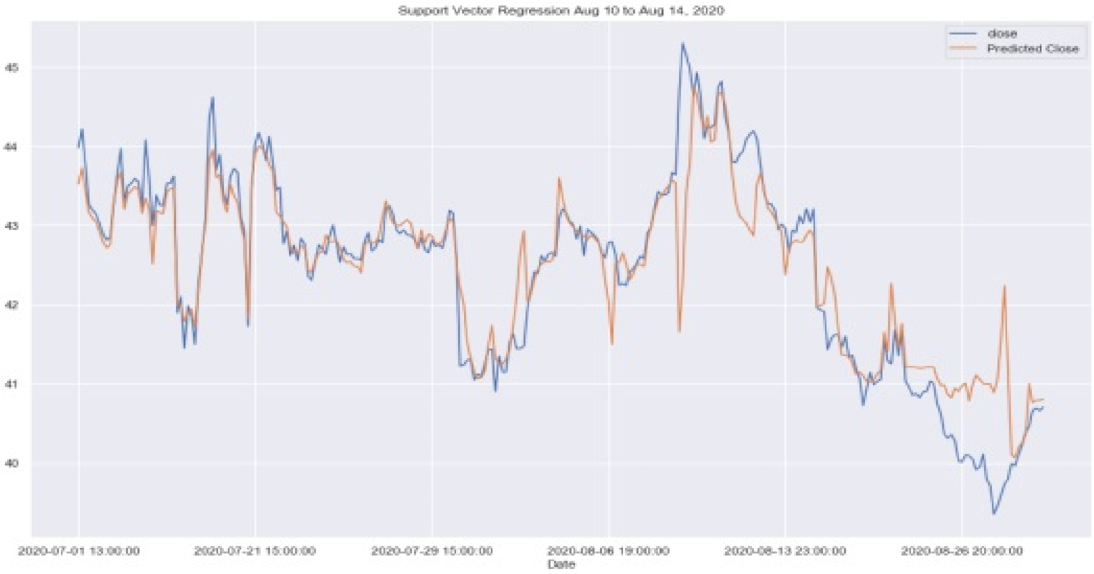

Figure 6.1: SVR performance with Exxon data

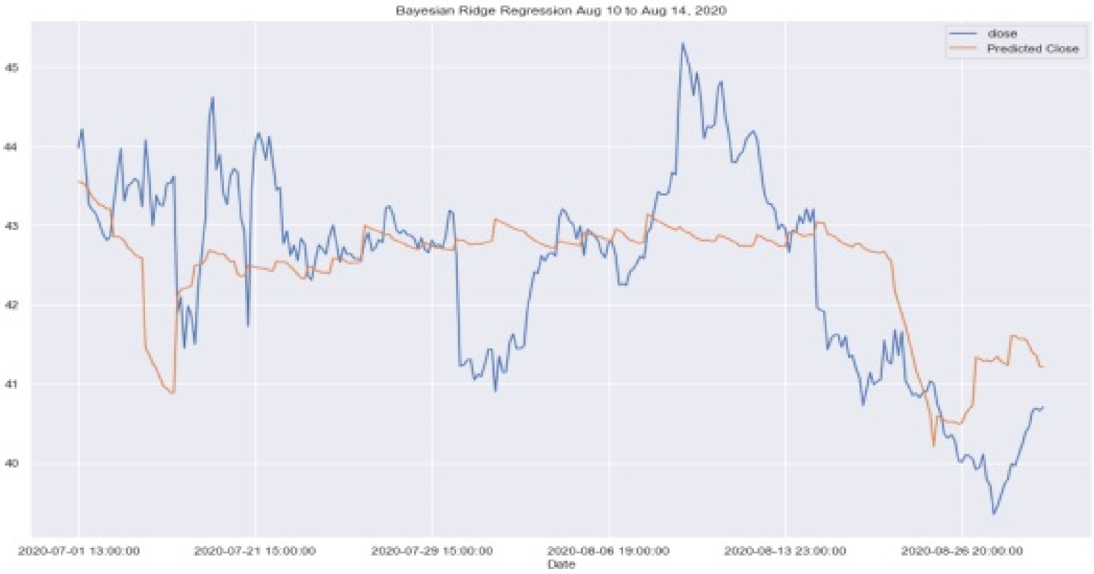

Figure 6.2: Bayesian ridge regression performance with Exxon data

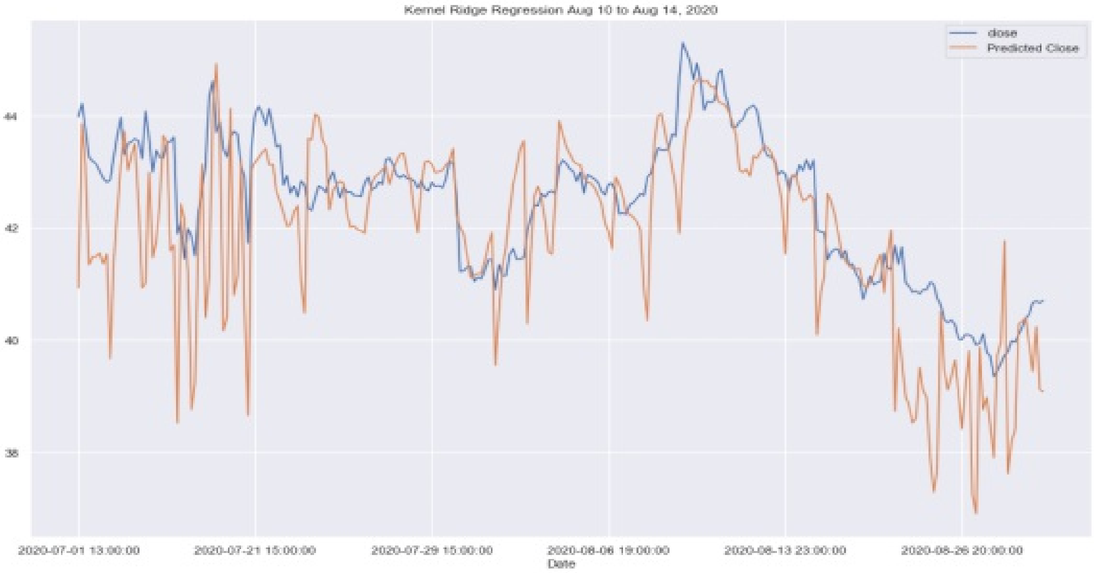

Figure 6.3: Kernel ridge regression performance with Exxon data

The Line Graphs in Figure 6.1, 6.2 and Figure 6.3 show us the Close Price and Predicted close price on the dataset using Support Vector Method, Bayes Ridge Regression Method and Kernel Ridge Regression models. The line plots clearly show that Support Vector Regression model with Radial Basis Kernel with gamma parameter as 1 and alpha parameter as 0.1 is outperforming the other two algorithms.

After adopting the best performing algorithm i.e. Support Vector Regression model with radial basis function as kernel function, parameters gamma = 1 and alpha = 0.1, we fit the model using 6- Fold Time-Series split Cross validation. The Line Plots in Figure 6.4, 6.5 and Figure 6.6 we can see that for the first graph in Figure 6.4 when the model is trained with 40 rows of recent data containing Sentiment Moving Averages, Sentiment Exponential Moving averages and Shifted Close price, the predicted price denoted by green line is flat which shows that there is no variance measured in predicted price. As is confirmed by the measure of R-squared for the same split which is -0.204461. The Mean Absolute Error (M.A.E) is highest which implies that the predicted value has highest deviation from true value. Mean Absolute Log Error (M.A.L.E) shows us how varied the errors are. In summary, the best model has highest R-squared value and low M.A.E and M.A.L.E. The second graph in Figure 6.4, shows that the model has close approximations for older stock prices in test data, Table 6.1 shows that out of 118 recent observations, with training size as 79 and testing size as 39 observations, the model has second highest R-squared value at 0.585571 with low M.A.E = 0.295586 and low M.A.L.E. = 0.000118. This implies that the range of the training dataset impacts accuracy of the model to predict unseen test data.

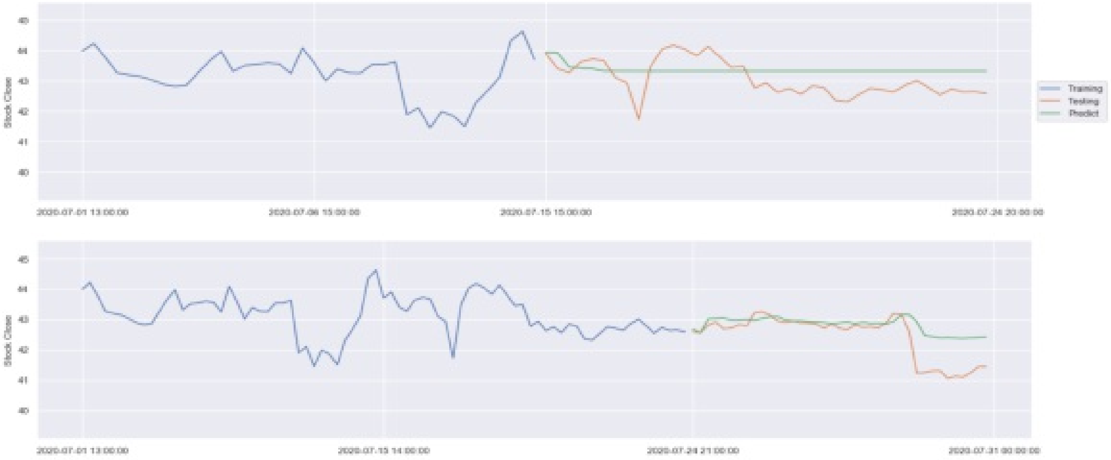

Figure 6.4: Exxon results part 1 out of 3

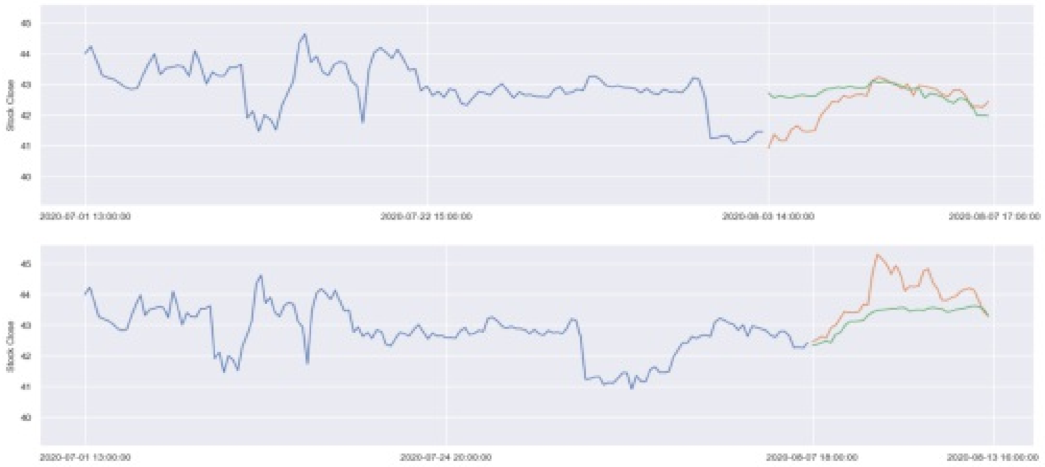

Figure 6.5: Exxon results part 2 out of 3

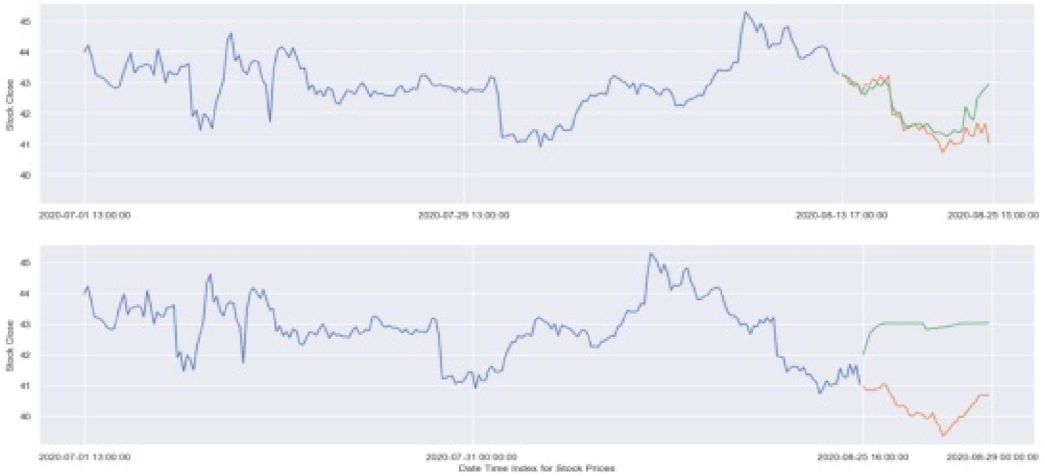

Figure 6.6: Exxon results part 3 out of 3

Index	Observations	Training-Size	Testing-Size	R-Squared	M.A.E	M.A.L.E
1	79.0	40.0	39.0	-0.204461	0.562073	0.000211
2	118.0	79.0	39.0	0.585571	0.295586	0.000118
3	157.0	118.0	39.0	0.299077	0.352551	0.000152
4	196.0	157.0	39.0	0.355480	0.438428	0.000173
5	235.0	196.0	39.0	0.682144	0.234102	0.000119
6	274.0	235.0	39.0	-32.51616	2.601228	0.003902

Table 6.1: Performance Measures Table for Time-Series Split Model for Support Vector Regression, Exxon Stock Prediction, M.A.E = Mean Absolute Error, M.A.L.E = Mean Absolute Log Error

From the Table 6.1, we can see that from the K-Fold Cross Validation of the Support Vector Regression model in Time Series Split data, the split with lowest error rates and highest R-squared is at 235 observations with 196 rows of training data and 39 rows of test data which gives R-squared as 0.682144, Mean Absolute Error as 0.234102 and Mean Absolute Log Error as 0.000119. Hence, for our case study, it is apt to conclude that best model fit for predicting Exxon Stock prices using Twitter Sentiment data from July 01, 2020 to August 31, 2020 is Support Vector Regression with Radial Basis Kernel, parameters gamma = 1.0 and alpha = 0.1, with 235 observations from which 196 rows are used for training and 39 rows for testing.

6.1.2 Chevron

From the line graphs, 6.7, 6.8 and 6.9, show us the close price and predicted close price on the dataset using Support Vector Method, Bayes Ridge Regression Method and Kernel Ridge Regression models. The line plots clearly shows that Support Vector Regression model with Radial Basis Kernel with gamma parameter as 1 and alpha parameter equals to 0.1 has better prediction compared to other machine learning models.

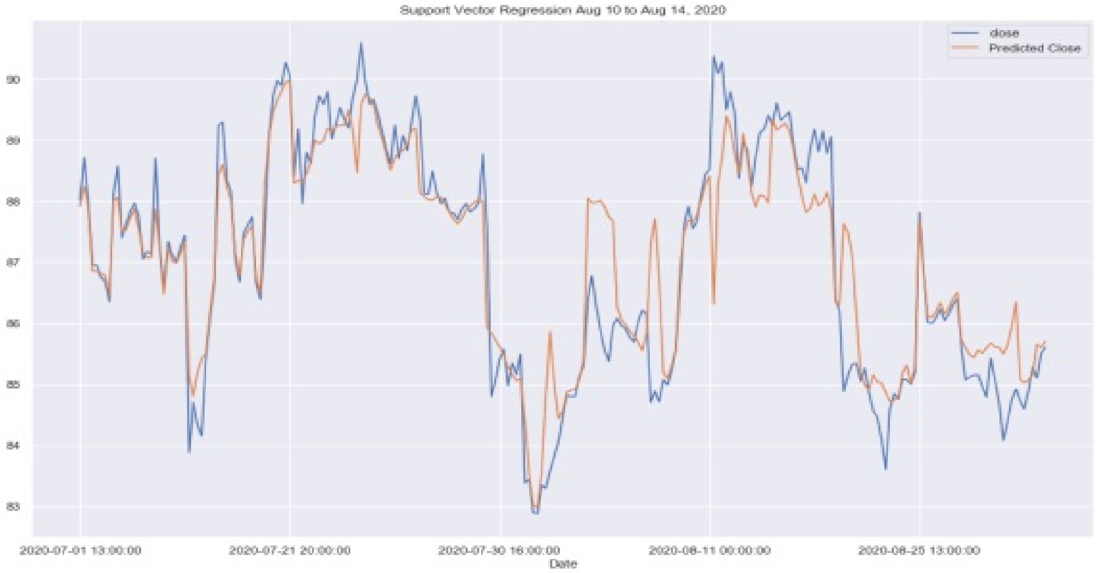

Figure 6.7: SVR performance with Chevron data

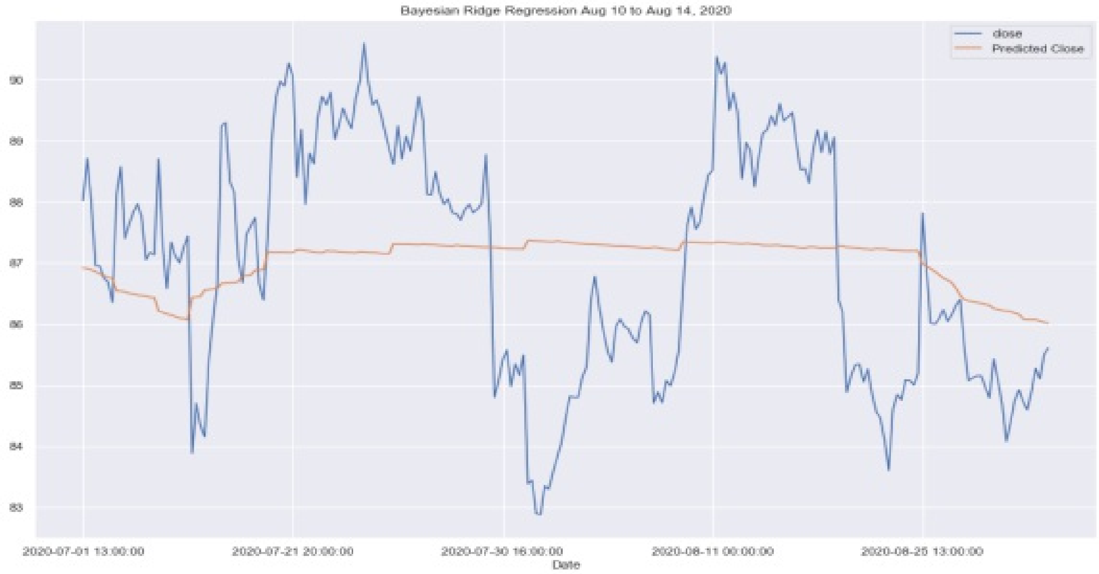

Figure 6.8: Bayesian ridge regression performance with Chevron data

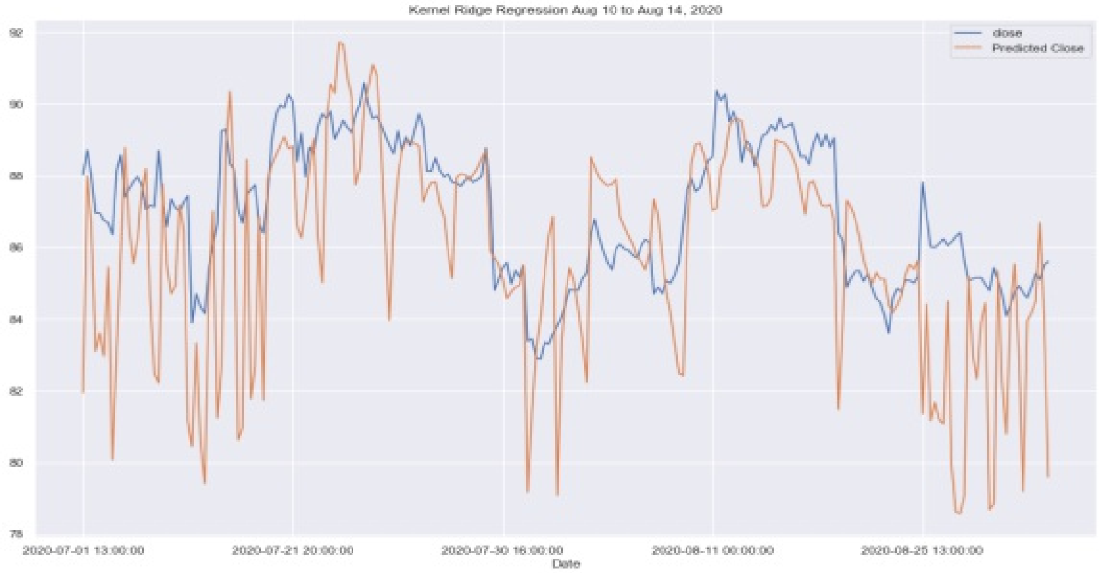

Figure 6.9: Kernel ridge regression performance with Chevron data

After adopting the best performing algorithm i.e. Support Vector Regression model with radial basis function as kernel function, parameters gamma = 1 and alpha = 0.1, we fit the model using 6- Fold TimeSeries split Cross validation as done above. The Line Plots in Figure 6.10, 6.11 and Figure 6.12 we can see that for the first graph in Figure 6.10 when the model is trained with 33 rows of recent data containing Sentiment Moving Averages, Sentiment Exponential Moving averages and Shifted Close price, the predicted price denoted by green line is flat which shows that there is no variance measured in predicted price. As is confirmed by the measure of R-squared for the same split which is -3.391720. The Mean Absolute Error (M.A.E) is one of the highest which implies that the predicted value has one of the highest deviation from true value. Mean Absolute Log Error (M.A.L.E) shows us how varied the errors are. In summary, the best model has highest R-squared value and low M.A.E and M.A.L.E. The second graph in Figure 6.11, shows that the model has close approximations for older stock prices in test data, Table 6.2 shows that out of 165 recent observations, with training size as 132 and testing size as 39 observations, the model has second highest R-squared value at 0.62408 with low M.A.E = 0.882649 and low M.A.L.E. = 0.000154. This implies that the range of the training dataset impacts accuracy of the model to predict unseen test data.

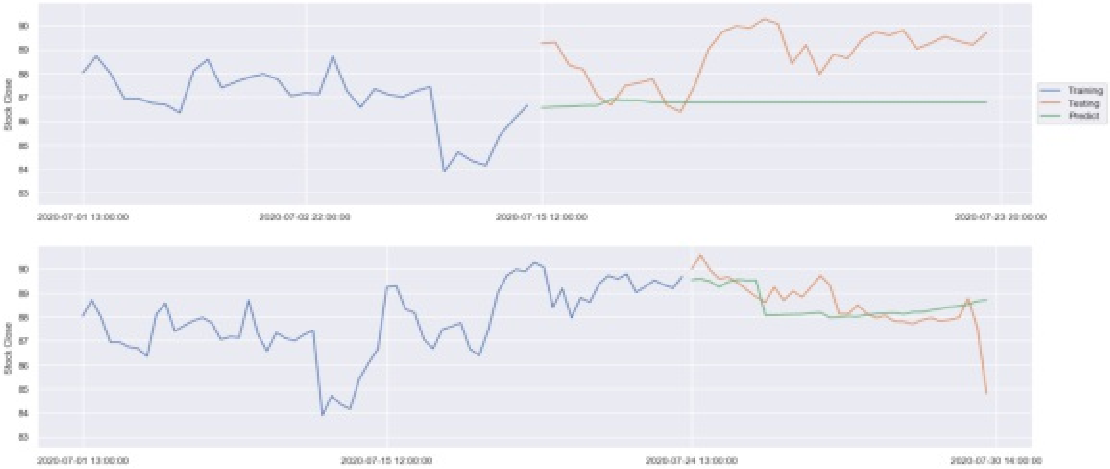

Figure 6.10: Chevron results part 1 out of 3

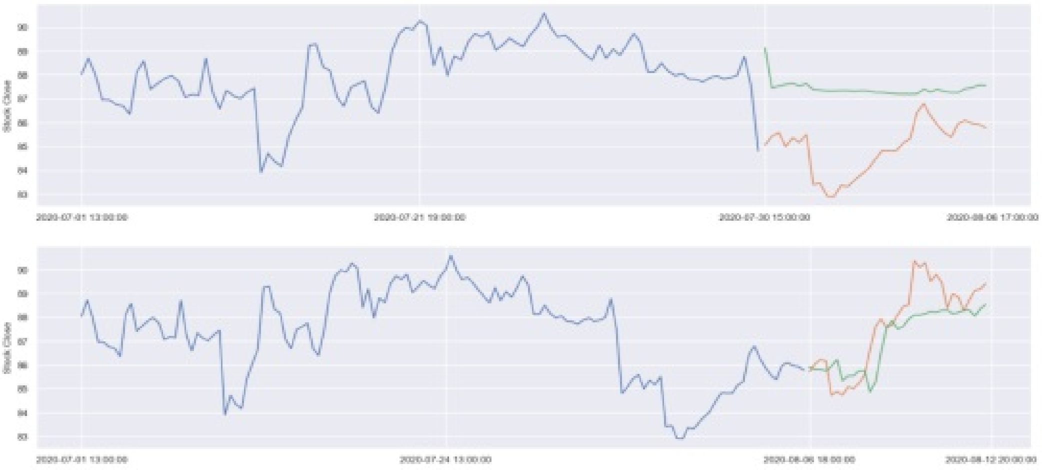

Figure 6.11: Chevron results part 2 out of 3

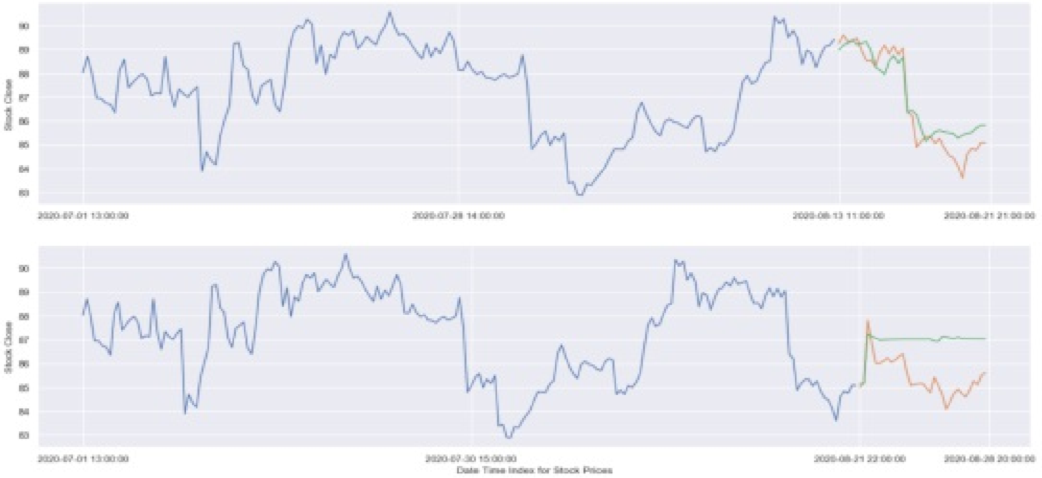

Figure 6.12: Chevron results part 3 out of 3

Index	Observations	Training-Size	Testing-Size	R-Squared	M.A.E	M.A.L.E
1	66.0	33.0	39.0	-3.391720	2.016106	0.000637
2	99.0	66.0	39.0	0.163284	0.647082	0.000114
3	132.0	99.0	39.0	-5.447126	2.479950	0.000973
4	165.0	132.0	39.0	0.624083	0.882649	0.000154
5	198.0	165.0	39.0	0.884001	0.565986	0.000066
6	231.0	198.0	39.0	-4.012119	1.534323	0.000388

Table 6.2: Performance Measures Table for Time-Series Split Model for Support Vector Regression, Chevron Stock Prediction, M.A.E = Mean Absolute Error, M.A.L.E = Mean Absolute Log Error

From the Table 6.2, we can see that from the K-Fold Cross Validation of the Support Vector Regression model in Time Series Split data, the split with lowest error rates and highest R-squared is at 198 observations with 165 rows of training data and 39 rows of test data which gives R-squared as 0.884001, Mean Absolute Error as 0.565986 and Mean Absolute Log Error as 0.000066. Hence, for our case study, it is apt to conclude that best model fit for predicting Chevron Stock prices using Twitter Sentiment data from July 01, 2020 to August 31, 2020 is Support Vector Regression with Radial Basis Kernel, parameters gamma = 1.0 and alpha = 0.1, with 198 observations from which 165 rows are used for training and 39 rows for testing.

6.1.3 BP Oil

From the line graphs, 6.13, 6.14 and 6.15, show us the Close Price and Predicted close price on the dataset using Support Vector Method, Bayes Ridge Regression Method and Kernel Ridge Regression models. The line plots clearly shows that Support Vector Regression model with Radial Basis Kernel with gamma parameter as 1 and alpha parameter equals to 0.1 also has better prediction compared to other machine learning models, similar to Chevron and Exxon results.

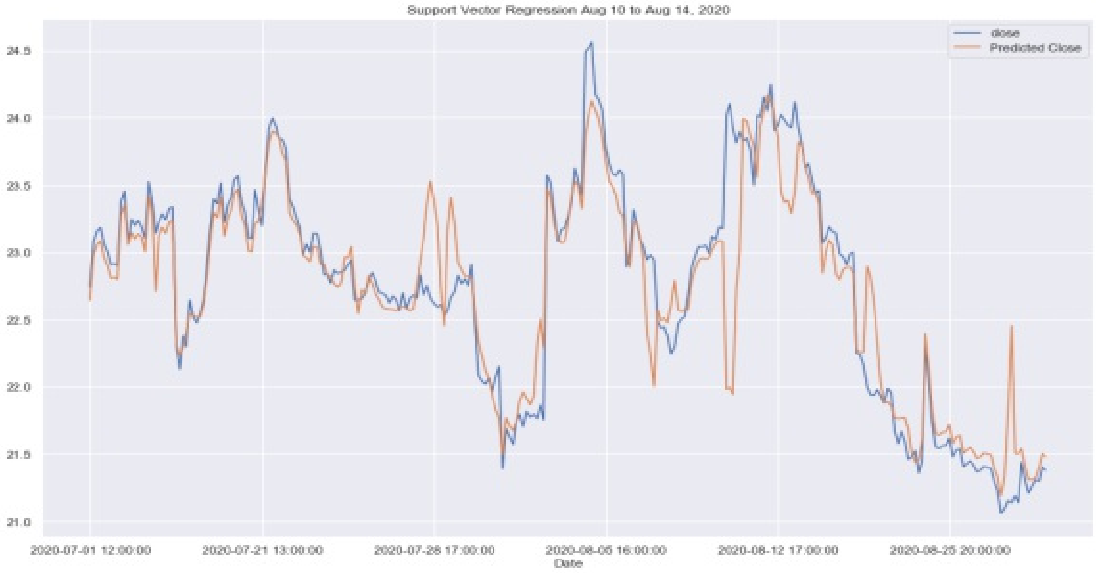

Figure 6.13: SVR performance with BP Oil data

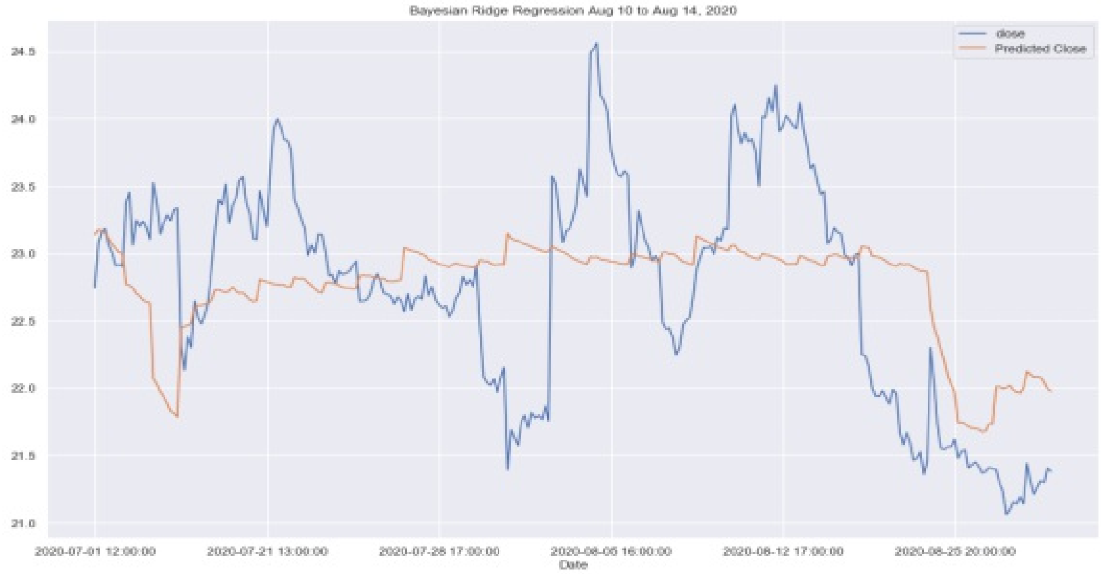

Figure 6.14: Bayesian ridge regression performance with BP Oil data

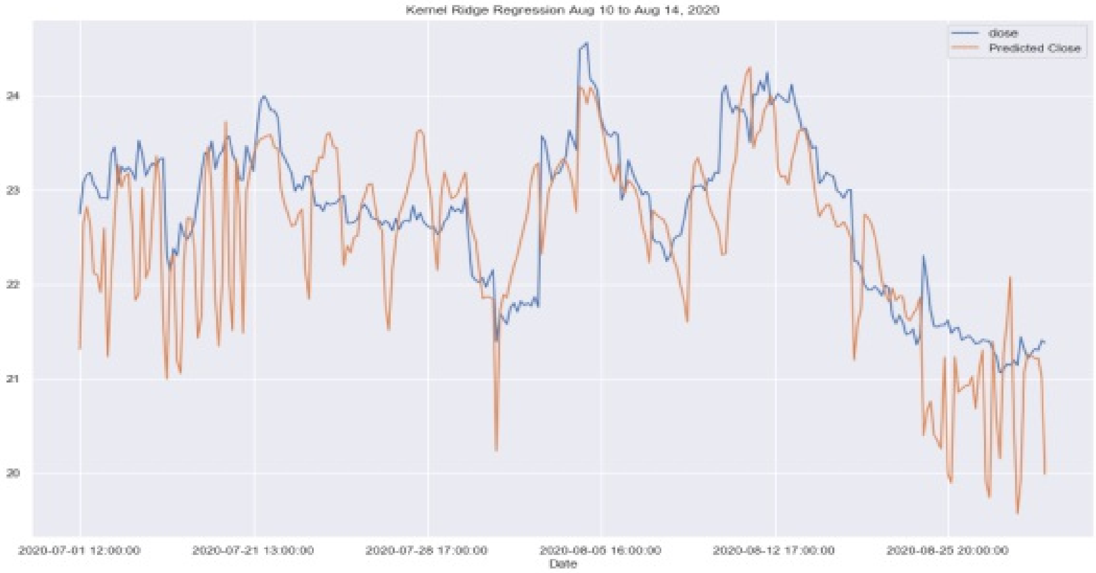

Figure 6.15: Kernel ridge regression performance with BP Oil data

After adopting the best performing algorithm i.e. Support Vector Regression model with radial basis function as kernel function, parameters gamma = 1 and alpha = 0.1, we fit the model using 6- Fold TimeSeries split Cross validation as done above. The Line Plots in Figure 6.16, 6.17 and Figure 6.18 we can see that for the first graph in Figure 6.16 when the model is trained with 45 rows rows of recent data containing Sentiment Moving Averages, Sentiment Exponential Moving averages and Shifted Close price, the predicted price denoted by green line is flat which shows that there is no variance measured in predicted price. As is confirmed by the measure of R-squared for the same split which is -0.001287. The Mean Absolute Error (M.A.E) is one of the highest which implies that the predicted value has one of the highest deviation from true value. Mean Absolute Log Error (M.A.L.E) shows us how varied the errors are. In summary, the best model has highest R-squared value and low M.A.E and M.A.L.E. The second graph in Figure 6.17, shows that the model has close approximations for older stock prices in test data, Table 6.3 shows that out of 201 recent observations, with training size as 162 and testing size as 39 observations, the model has second highest R-squared value at 0.679446 with low M.A.E = 0.258039 and low M.A.L.E. = 0.000214. This implies that the range of the training dataset impacts accuracy of the model to predict unseen test data.

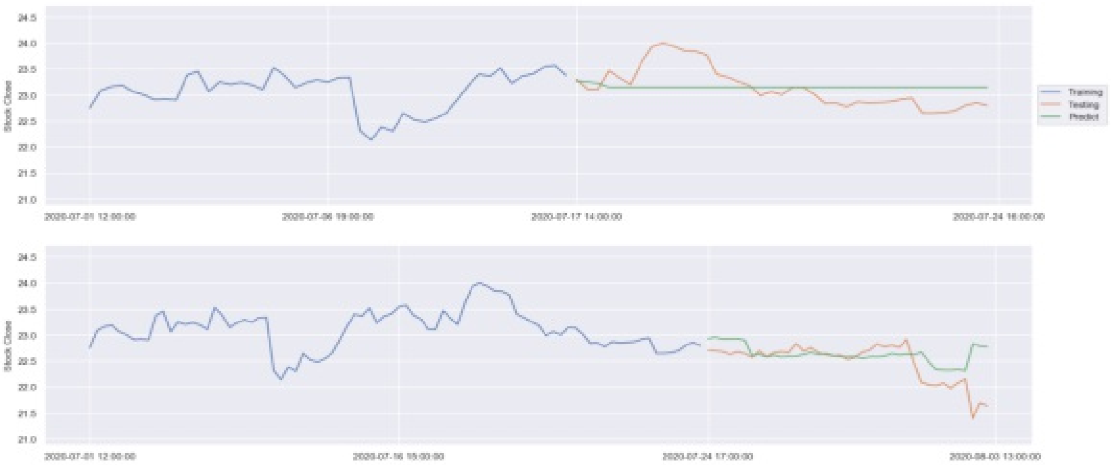

Figure 6.16: BP Oil results part 1 out of 3

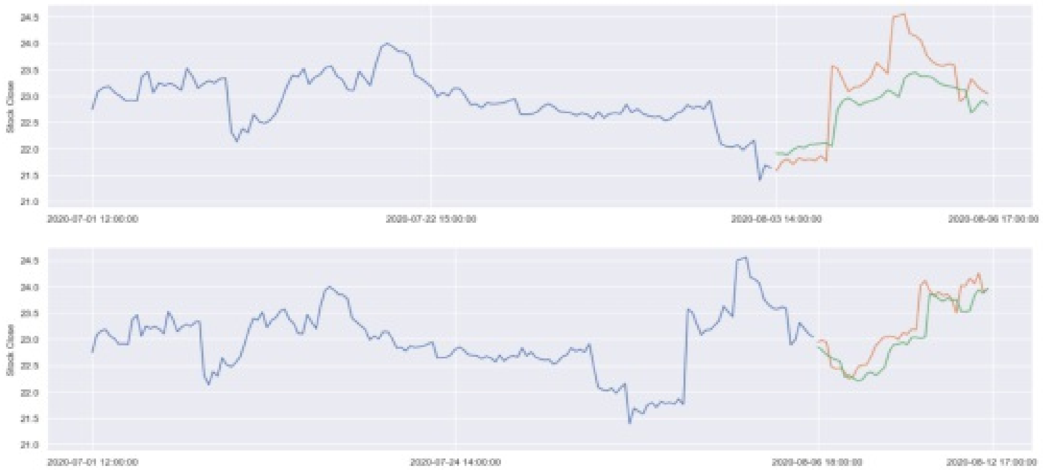

Figure 6.17: BP Oil results part 2 out of 3

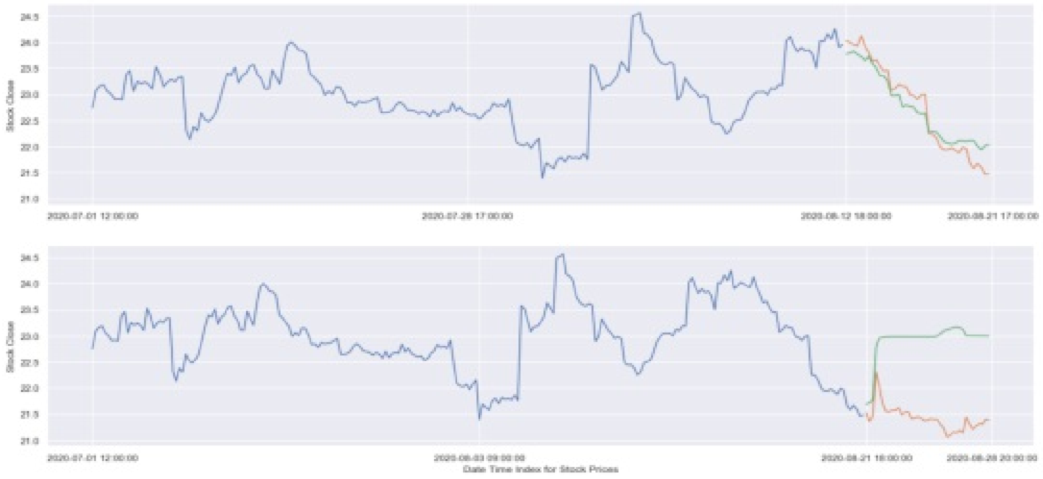

Figure 6.18: BP Oil results part 3 out of 3

Index	Observations	Training-Size	Testing-Size	R-Squared	M.A.E	M.A.L.E
1	84.0	45.0	39.0	-0.001287	0.317618	0.000258
2	123.0	84.0	39.0	-0.201413	0.251879	0.000295
3	162.0	123.0	39.0	0.529281	0.477335	0.000613
4	201.0	162.0	39.0	0.679446	0.258039	0.000214
5	240.0	201.0	39.0	0.902785	0.227979	0.000131
6	279.0	240.0	39.0	-44.19945	1.490381	0.004522

Table 6.3: Performance Measures Table for Time-Series Split Model for Support Vector Regression, BP Oil Stock Prediction, M.A.E = Mean Absolute Error, M.A.L.E = Mean Absolute Log Error

From the Table 6.3, we can see that from the K-Fold Cross Validation of the Support Vector Regression model in Time Series Split data, the split with lowest error rates and highest R-squared is at 240 observations with 201 rows of training data and 39 rows of test data which gives R-squared as 0.902785, Mean Absolute Error as 0.227979 and Mean Absolute Log Error as 0.000131. Hence, for our case study, it is apt to conclude that best model fit for predicting BP Oil Stock prices using Twitter Sentiment data from July 01, 2020 to August 31, 2020 is Support Vector Regression with Radial Basis Kernel, parameters gamma = 1.0 and alpha = 0.1, with 240 observations from which 201 rows are used for training and 39 rows for testing.

CHAPTER 7

Conclusion

Taking in consideration the data gathered since 1880, one third of all CO2 emissions can be directly associated to the top 20 fossil fuel companies, which with their actions they continuously exploited the world's oil, gas and coal reserves. Additionally, opinions on specific topics discussed in today's social media platforms can influence the decision process of investors regarding the stock markets.

This thesis consists of examining if tweets gathered from Twitter can help in predicting top CO2 emission firms stock market exchange rate. We analyze data which was collected for the time span of two months, from the beginning of July 2020 until the end of August 2020. In order to predict top CO2 emission firms stock market exchange rate, we analyzed sentiment of tweets which contained the hashtags #climatechange, #globalwarming, #carbonproducers, #endfossilfuel, #CO2, #CH4, #surfacetemperaturerise #sealevelrise, #exxon, #bp_oil, #chevron.

The use of VADER (Valence Aware Dictionary for Sentiment Reasoning) which is a sentiment analysis model, served us for the purpose of extracting the sentiment from the collected tweets with the corresponding hashtags. The results which we gained, showed us that there is a correlation between Exxon, Chevron and BP Oil stock market prices and sentiment collected from tweets which express public opinion. Respectively, we found out that there is a negative correlation between sentiment indicators and stock prices. Increment in sentiment indicator causes decrease in stock prices, whereas reduction in sentiment indicator causes increase in stock prices. When the prices of the top CO2 emission firms are increasing, there are a lot of negative tweets regarding the problem of climate change and global warming. Meanwhile, when the prices are decreasing, there are a lot of positive tweets about the work being done regarding climate change and global warming. For the purpose of forecasting the prices of Exxon, Chevron and BP Oil, we used three different regression algorithms: SVR, BRR, KRR. Out of these three algorithms the SVR outperformed BRR and KRR and is used as the algorithm of choice for the whole predictions. Furthermore, based on the gained results we can see that recent sentiments increase prediction accuracy in stocks, whereas older sentiments worsen the predictions.

As a conclusion we can say that, it is possible to predict top CO2 emission firms stock market exchange rate by using public opinion attained from Twitter. However, real time stock predictions models are uncertain.

List of Figures

List of Tables

Acronyms

BRR Bayesian Ridge Regression. xi, xiii, 26, 53

KRR Kernel Ridge Regression. xi, xiii, 27, 53

SVR Support Vector Regression. xi, xiii, 27, 53

VADER Valence Aware Dictionary for Sentiment Reasoning. xi, xiii, 4, 6, 23, 32, 53

Bibliography

[AAX14] Marta Arias, Argimiro Arratia, and Ramon Xuriguera. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–24, 2014.

[BC87] Helmut Braun and John S Chandler. Predicting stock market behavior through rule induction: an application of the learning-from-example approach. *Decision Sciences*, 18(3):415–429, 1987.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[CH13] Linda H Connor and Nick Higginbotham. “natural cycles” in lay understandings of climate change. *Global environmental change*, 23(6):1852–1861, 2013.

[Cho20] KR Chowdhary. Natural language processing. In *Fundamentals of artificial intelligence*, pages 603–649. Springer, 2020.

[CMP] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684.

[CRR86] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. *Journal of business*, pages 383–403, 1986.

[EBD$^+$17] Brenda Ekwurzel, James Boneham, MW Dalton, Richard Heede, Roberto J Mera, Myles R Allen, and Peter C Frumhoff. The rise in global atmospheric co 2, surface temperature, and sea level from emissions traced to major carbon producers. *Climatic Change*, 144(4):579–590, 2017.

[Fel13] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[Gar68] Paul L Garvin. *The Georgetown-IBM experiment of 1954: an evaluation in retrospect.* Mouton, 1968.

[GHB09] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis. *Entropy*, 17:252, 2009.

[Gor62] Myron J Gordon. The savings investment and valuation of a corporation. *The Review of Economics and Statistics*, pages 37–51, 1962.

[Hee14] Richard Heede. Tracing anthropogenic carbon dioxide and methane emissions to fossil fuel and cement producers, 1854–2010. *Climatic change*, 122(1):229–241, 2014.

[ID10] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2010.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, 2007.

[KM05] Asim Ijaz Khwaja and Atif Mian. Unchecked intermediaries: Price manipulation in an emerging stock market. *Journal of Financial Economics*, 78(1):203–241, 2005.

[KMPH09] Thomas R Karl, Jerry M Melillo, Thomas C Peterson, and Susan J Hassol. *Global climate change impacts in the United States*. Cambridge University Press, 2009.

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[Mat16] VV Matveev. Monitoring of economic risks of business activities. *Russian Journal of Agricultural and Socio-Economic Sciences*, 55(7), 2016.

[MB19] Adrija Majumdar and Indranil Bose. Do tweets create value? a multiperiod analysis of twitter use and content of tweets for manufacturing firms. *International Journal of Production Economics*, 216:1–11, 2019.

[Mel05] Antonio Mele. Understanding stock market volatility. *London School of Economics Financial*, 2005.

[MG12] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[MM58] Franco Modigliani and Merton H Miller. The cost of capital, corporation finance and the theory of investment. *The American economic review*, 48(3):261–297, 1958.

[Mur18] Dhiraj Murthy. *Twitter*. Polity Press Cambridge, 2018.

[NOBH$^+$15] GJ Nason, F O'Kelly, D Bouchier-Hayes, DM Quinlan, and RP Manecksha. Twitter expands the reach and engagement of a national scientific meeting: the irish society of urology. *Irish Journal of Medical Science (1971-)*, 184(3):685–689, 2015.

[NOMC11] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[PM18] Lei Pan and Vinod Mishra. Stock market development and economic growth: Empirical evidence from china. *Economic Modelling*, 68:661–673, 2018.

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[RS$^+$12] Tushar Rao, Saket Srivastava, et al. Analyzing stock market movements using twitter sentiment analysis. 2012.

[SDRH13] Kirk R Smith, Manish A Desai, Jamesine V Rogers, and Richard A Houghton. Joint co2 and ch4 accountability for global warming. *Proceedings of the National Academy of Sciences*, 110(31):E2865–E2874, 2013.

[SFY19] Zhi Su, Tong Fang, and Libo Yin. Understanding stock market volatility: What is the role of us uncertainty? *The North American Journal of Economics and Finance*, 48:582–590, 2019.

[Sma11] Tamara A Small. What the hashtag? a content analysis of canadian politics on twitter. *Information, communication & society*, 14(6):872–895, 2011.

[SRDBC15] Skipper Seabold, Alex Rutherford, Olivia De Backer, and Andrea Coppola. The pulse of public opinion: using twitter data to analyze public perception of reform in el salvador. *World Bank Policy Research Working Paper*, (7399), 2015.

[SWZ09] Didier Sornette, Ryan Woodard, and Wei-Xing Zhou. The 2006–2008 oil bubble: Evidence of speculation, and prediction. *Physica A: Statistical Mechanics and its Applications*, 388(8):1571–1576, 2009.

[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[TCG$^+$04] Chris D Thomas, Alison Cameron, Rhys E Green, Michel Bakkenes, Linda J Beaumont, Yvonne C Collingham, Barend FN Erasmus, Marinez Ferreira De Siqueira, Alan Grainger, Lee Hannah, et al. Extinction risk from climate change. *Nature*, 427(6970):145–148, 2004.

[TW19] Matthew Taylor and Jonathan Watts. Revealed: the 20 firms behind a third of all carbon emissions. *The Guardian*, 9(10):2019, 2019.

[WHMW11] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714, 2011.

[Yin66] Charles C Ying. Stock market prices and volumes of sales. *Econometrica: Journal of the Econometric Society*, pages 676–685, 1966.