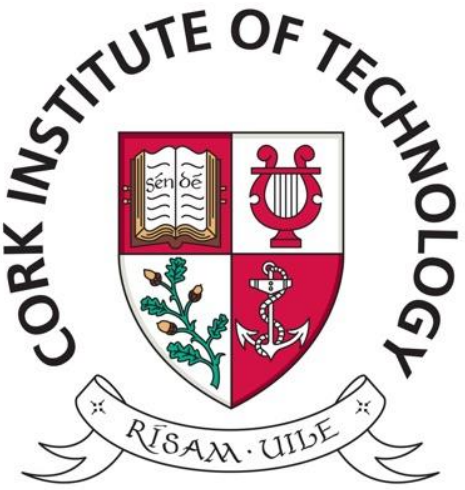
CORK INSTITUTE OF TECHNOLOGY
DEPARTMENT OF MATHEMATICS



DATA SCIENCE & ANALYTICS PROJECT

Sentiment Analysis of Financial Data

Svetlana Ivanov
R00181392

Supervisor :
Dr David Hawe

18 May, 2020

Abstract

With rapid development of Data Science and Artificial Intelligence, there is a significant volume of research being conducted in the field of sentiment analysis and natural language processing (NLP). It is used widely for estimating customer satisfaction during the launch of new products on the market.

In the era of 'Big Data' most of decisions makers want to know what Social Media and Internet users are thinking about their products and services. But the volume of information is so large that it is impossible to read all texts. And often it is not necessary. It is more important to know the sentiment of reviews. This gave a new boost to development of sentiment analysis. Recently it is used in different fields of human activity. Companies are becoming interested now not only in analysing their customers feedbacks or Social Media reviews. They want to understand how their business can be influenced by news that mention related products or services.

Financial specialists very often have to take instant decisions about buying or selling some stocks or assets. Forecasting's for stock and commodities market are done all around the world. Prices on this market are influenced by a large set of factors.

However, there is not a significant volume of literature on the sentiment analysis techniques being applied to financial data. This report investigates the relationships between sentiment from some social media sources, financial sites news, in particular from financial site Bloomberg, and prices for energy sources.

Declarations

This report was written entirely by the author, except where stated otherwise, and has not been submitted for another degree, at Cork Institute of Technology or elsewhere. The source of any material not created by the author has been clearly referenced. The work described in this report was conducted by the author, except where stated otherwise.

Signed: Date:

Acknowledgements

I would like to thank my CIT supervisor Dr David Hawe for his great support, encouragement and guidance through the project. I would also like to thank Dr Patrick Tuite from Brookfield Renewable for the idea of this project.

Contents:

1. Introduction

With the intensive growth of researches in the field of sentiment analysis there is still relatively little literature about researches and techniques of sentiment analysis being applied specifically to the financial data. The scope of the project was to gather data from web sources such as Bloomberg news and Tweeter and to conduct a sentiment analysis on news articles and tweets and find a link between this analysis and price for oil.

"*Large and persistent changes in the real price of oil have an important impact to the welfare of both oil-importing and oil-producing economies. This is why forecasts of the oil prices can be of interest for a wide range of applications. Central banks and private sector forecasters view the price of oil as one of the key variables in generating macroeconomic projections and in assessing macroeconomic risks. The price of oil is important in predicting recessions.* "(Alguist, Killian and Vigfusson, 2011)

Forecasting the price of oil can be very helpful for energy companies in the process of managing their finance and providing energy at a competitor price as well as winning new customers and increasing profits.

With an intense development of Data Science and Artificial Intelligence, it is logical to investigate all factors that can influence stock prices and, in particular price of oil, in order to improve forecasting methods.

This project presents an attempt to investigate the impact of sentiment level in economic news such as Bloomberg or Media Data as tweets, influence dynamics of energy prices, in particular of oil or gas.

"Crude oil prices are determined by global supply and demand. Economic growth is one of the biggest factors affecting petroleum product—and therefore crude oil—demand. Growing economies increase demand for energy in general and especially for transporting goods and materials from producers to consumers. The world's transportation sector is almost totally dependent on petroleum products such as gasoline and diesel fuel. Many countries also rely heavily on petroleum fuels for heating, cooking, or generating electricity. Petroleum products made from crude oil and other hydrocarbon liquids account for about a third of total world energy." (EIA, 2020)

When work on this project was still in process, World Health Organisation (WHO, 2020) announced COVID-19 outbreak a pandemic at 11 March 2020. The pandemic requires a strict lockdown and social distancing and it brought a severe decrease in business activity all around the world. And as result demand for oil in the world dropped down and price for oil followed. Moreover, the Russia-Saudi Arabia oil price war was triggered in March 2020 by Saudi Arabia in response to Russia's refusal to reduce oil production. This economic conflict resulted in a significant drop of price of oil.

In April 2020 price for oil became negative as even the lowest possible production level generated much greater supply than demand, thus oil industry had nowhere to store oil and producers were ready to pay for oil to be taken away and free production storages. (Hansen, 2020)

Both these events introduced a serious bias in this analysis and its results. To reduce the bias at some degree in the keyword list for streaming tweets hashtags were introduced #COVID-19 and #coronavirus that were not planned at beginning of the project.

2 Literature review

2.1 Forecasting price of oil

There are attempts to forecast price of oil on base of futures. "*Futures are financial contracts obligating the buyer to purchase an asset or the seller to sell an asset and have a predetermined future date and price."* (Chen, 2018) Futures look like a simple and transparent tool that can give good forecast and can be easily explained. Such forecasts don't always bring expected result and in the past they leaded to big errors. (Alquist, Killian, Vigfusson, 2011). But if not futures, what else can be helpful in this case?

There is a market hypothesis that the security market is very reactive and reflects the information instantly. "*However, there are many anomalies showing that the exogenous information plays an important role in the stock market. News, for example, as one type of the exogenous information, has been intensively investigated for its influence on the stock market, including the relation between the news and stock prices*" (Chan, 2003)

In his paper Chan set out to determine a link between news and no news about a company or a commodity and how this influences stock price. He built a significant portfolio of news going back from before 1980s and focused his research not on quality of news sources, but on how investors react to public news. His conclusions are: "*In summary, the following patterns stand out: even after making some adjustments no-news stocks experience short term reversal, and news losers show substantial drift*". His conclusions are as follows: "*The results shown above indicate that smaller stocks seem to underreact to bad news. Why might this be the case? Researchers have offered two explanations, which are not necessarily mutually exclusive. The first is that investors simply have differential attitudes to good and bad news. They may consistently underreact to bad news, but have a different response for positive signals.*" He is using different terminology, but he speaks about sentiment analysis here.

It was announced (Schumaker, 2018) that analysis of stock price movement following financial news showed that news in general can influence stock prices and the impact of this influence depends of many factors. "*For example, articles released through WSJ, Reuters – UK Focus, NYT and FT experienced significant positive returns, whereas articles in Barrons,*

MarketWatch, Forbes and Bloomberg experienced significant negative return. Again, here we are speaking not about news in general, but about "bad" or "good" news. "

The role of media in stock price movement has been explored in finance literature. Most of cases were limited in exploration of long-term effect. Many of these studies show that news can cause price changes. How strong is this influence and how precise are forecast based on news? "*One study argues that by the time a story goes to press it is old news and that price adjustments were already made*." (Fang and Peress, 2009).

Not every person is reacting to the news in the same way. There are financial experts who can have a high reaction speed and can buy or sell stocks in minutes after news appeared, but most of investors are more conservative and prefer to wait and not take decisions purely on news.

As a physical commodity, oil prices are predictable to some extent, but their forecast don't always provide accurate results. Oil prices depend by oil fundamentals and in particular by global economic activity. To get an accurate forecast for oil price we have to take in consideration a set of factors as market dynamics, political events, discoveries of new energy technologies, global warming and, as present shows, even pandemics and epidemic situations. And all these events are not always predictable.

"*Recently, oil prices were stagnated or had a very small growth caused by oil market oversupply*. "(Montgomery, Cheryl, Kulachi, 2015).

The attempt to artificially increase price for oil by reducing oil supply was the main conflict between Saudi Arabia and Russia and lead to the event that we witnessed last month.

At 22 Apr 2020 in the price of oil was detected an anomality. Price became negative. We can consider this as an intervention point in the time series.

2.2. Sentiment and Public Information analysis

"*In the last years, Sentiment Analysis has become a hot-trend topic of scientific and market research in the field of Natural Language Processing (NLP) and Machine Learning."*
(Symeonidis, 2018)

A lot of data is generated by the social media website users as Tweeter, Facebook and similar to them and the content of this data plays an essential role in decision-making of many companies. Volume of information is so large that it is impossible to read the whole text that appear in the Internet. Sometimes it is important only to know that feelings and emotions are expressed by this text. And here sentiment analysis makes the task easier by providing the polarity to the text and classifying text into positive and negative classes. Polarity is classified by emotions that a text is generating and can be positive, negative or neutral. Polarity is not limited to every word analysis. It takes a phrase or a sentence to determine the emotion about a subject. For example, word "good" itself seems to represent a positive sentiment, but in the context "Are you calling this a good service?" we understand that behind this sarcasm stands a negative emotion. Polarity is not measured only by positivity or negativity, but also by intensity of this sentiment. Two sentences: "Food in this restaurant is good", "Food in this restaurant is good!!!" have the same words in them, but intensity in the second sentence is definitively higher than in the first one. In the real life we can estimate emotions by tone of voice or by face expression. We have experience how to do it, but often we can estimate it wrong. In the text we can use signs of punctuations, special lexicon, sarcasm, exaggerations or emojis. But how good are these estimations?

"*At moment on the market there are a large list of sentiment analysis tools, that declare that they can analyse everything without any additional effort from user side. We can mention such tools as Mediatoolkit, StarMine, HubSpot's Service Hub, Lexalytics, Talkwalker, Lexalytics and others*." (Fontarella, 2020)

Some of these tools are free to use and others require a subscription or one-time payment. Most of them are oriented to analyse customers feedback and customer satisfaction and are more useful for estimation on customer service quality level than for sentiment analysis of financial news.

The classification of sentiment analysis (Fontarella, 2020) is based on approaches for their development. (Figure 2.2.1)

Sentiment analysers are classified in 2 major groups: using Machine learning approach or lexicon-based approach. The lexicon-based analysis is classified again as lexicon-based approach and corpus-based approach.

For this project was used a sentiment analysis tool called VADER that stands for (Valence Aware Dictionary and sEntiment Reasoner) and it is a lexicon and corpus-based sentiment analysis tool. It was developed specifically to estimate sentiments from social media. VADER has incorporated a combination of a list of lexical features (words) that are labelled according to their semantic orientation as being positive or negative. In the same time there is a rule-based part that analysis words in context with rules in the human speech.

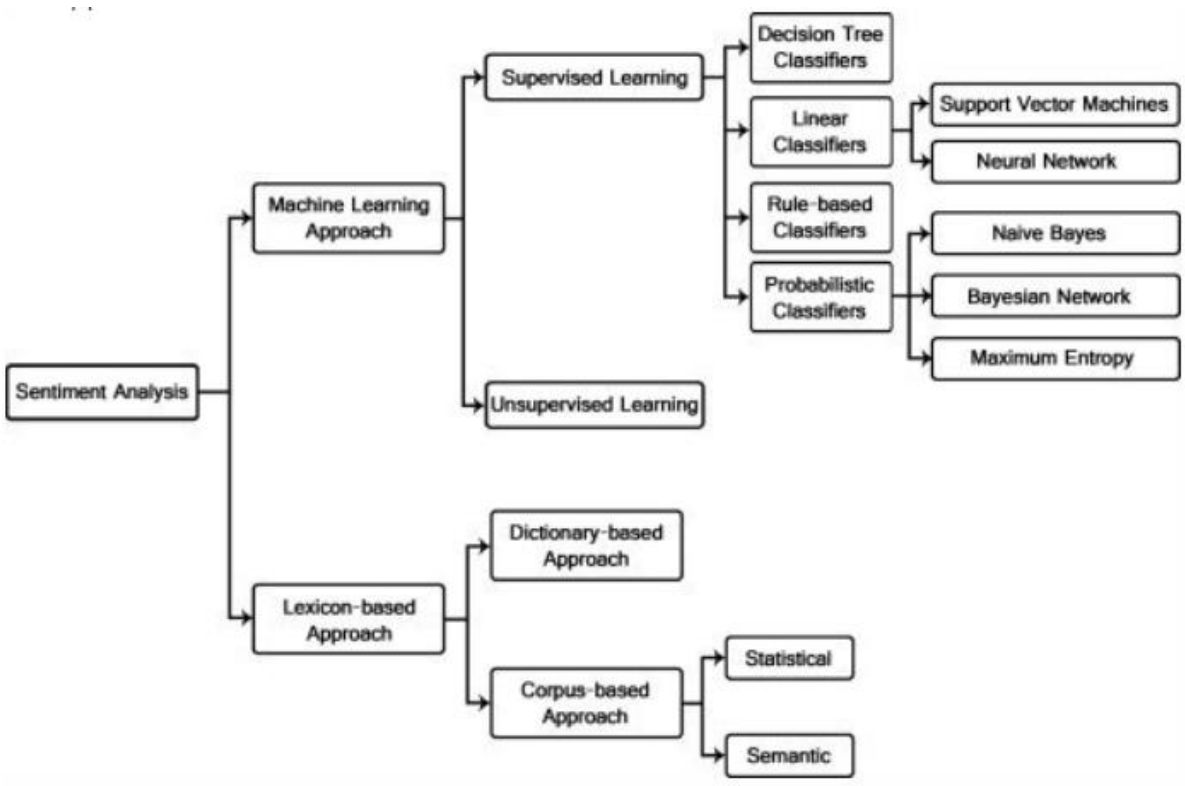


Figure 2.2.1 Sentiment classification techniques. (Fontarella, 2020)

VADER is largely used for analysis of different social media texts, such as news, movie and product reviews, tweets, Facebook comments. VADER is not only labelling words as positive or negative, but also estimating how positive or negative a sentiment is. (Hutto& Gilbert, 2014).

It is a free to use tool. Another big advantage of this tool is that it evaluates not only the text, but also the emojis that are often used in tweets or comments in social media.

VADER uses the polarity_scores method to obtain the polarity indices for the given sentence.

The Positive, Negative and Neutral scores represent the proportion of text that falls in these categories. This means that a sentence can be rated, for example, as 70 % Positive, 25% Neutral and 5% Negative. The sum of these ratings are adding to 1.

The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive). The higher compound scores the more positive is a phrase or a sentence.

2.3 Time Series Models

When we are trying to predict some trends in financial market or prices of oil, for example, we are analysing them from time point of view. The used for this purpose information is recorder with a certain frequency and in the most cases it is recorder for a long period of time.

"*Time series are one of the most common data types encountered in daily life. Financial prices, weather, home energy usage, and even weight are all examples of data that can be collected at regular intervals. Almost every data scientist will encounter time series in their daily work."* (Zinflou, 2018)

The volume of literature for Time series and how to analyse them is enormous. It is good to mention here a book (Montgomery, 2016) that was classified by Tableau as one of 7 great books about Time series Analysis (Tableau, 2020) and is recommended for learning.

A very good source for statistical knowledge that can be used for Time Series analysis it the book "An Introduction to Bootstrap" (Bradley, 1993) where is described the bootstrapping methodology used in time series forecasting.

“Introduction to Time Series Analysis and Forecasting” (Montgomery, 2015) is a hands-on textbook that presents the basics of time series analysis and includes data sets to practice statistical forecasting,”, (Tableau, 2020)

All datasets in this project represent time series: price of oil, news, tweets.

With the price of oil going negative, it is necessary to use consider that we are dealing with an event that occurred first time in the history of prices for oil. We have an anomaly situation that it is aggravated by another one anomaly: coronavirus pandemic. This makes forecasting of situation with the oil a very hard task.

In their book “Anomaly Detection Principles and Algorithms” (Mehrotra, Moham, Huang, 2017) authors mention that some anomalies may have been encountered previously in data. *“However, the classification approach can work only in detecting known problems of specific kind, whereas the greatest damage is caused by unknown problems newly created by ba*d actors...We may not have an explicit model, pattern or rule that describes the anomalies*.”*

2.4 Tweeter Tools

At moment Tweeter has a large list of useful tools for developers (Tweeter, 2020). Some of them are free for use, but most of them come for a price. After my application for Tweeter access was approved, access to a premium account was opened. Getting access to a Tweeter developer API, a list of authentication keys and tokens were generated. This set of credentials is required every time you must pass authentication with each request and are specific to the user that makes the request. For the project was used a developer tool, called OAuth 1.0a. On their site Tweeter indicates which libraries can be used for certain language. For python users there is a special library called Tweepy that contains all necessary information and instructions for developers.

2.5. Bloomberg news

Bloomberg news was one of the main sources of information for this project. The site (Bloomberg Green, 2020) is very popular among financial specialists. News on the site are divided in few departments: “Markets”, “Technologies”, “Politics”, “Pursuits”, “Opinion”, “Business week” and “Green”.

For the project were collected news from department “Green”, that it is further divide in a few sub departments: “Science and Energy”, “Climate Adaptation”, “Finance”, “Politics” and “Culture & Design”.

News from near all “Green” sub departments except “Culture & Design” were collected. In general, in “Culture & Design” were published in this period of time only a few news that were not related with prices for energy and were more discussing fashion and art.

Division of news in Departments doesn’t follow a strict rule. Very often the same news can be found in different departments.

2.6. Mining Bid Data, map-reduce approach

All stages of this project were realised at a home computer with good resources but still not enough for storing and processing that large volume of information that came from Tweeter. Home computer was working 12-14 hours every day for this project. For dataset pre-processing was used .the MapReduce method that it is usually used for Hadoop.

“*Hadoop supports the MapReduce model, which was introduced by Google as a method of solving a class of petascale problems with large clusters of inexpensive machines. The model is based on two distinct steps for an application:*

• *Map: An initial ingestion and transformation step, in which individual input records can be processed in parallel.*

• *Reduce: An aggregation or summarization step, in which all associated records must be processed together by a single entity*.” (Venner, 2009)

Figure 2.6.1 illustrates how the MapReduce model works.

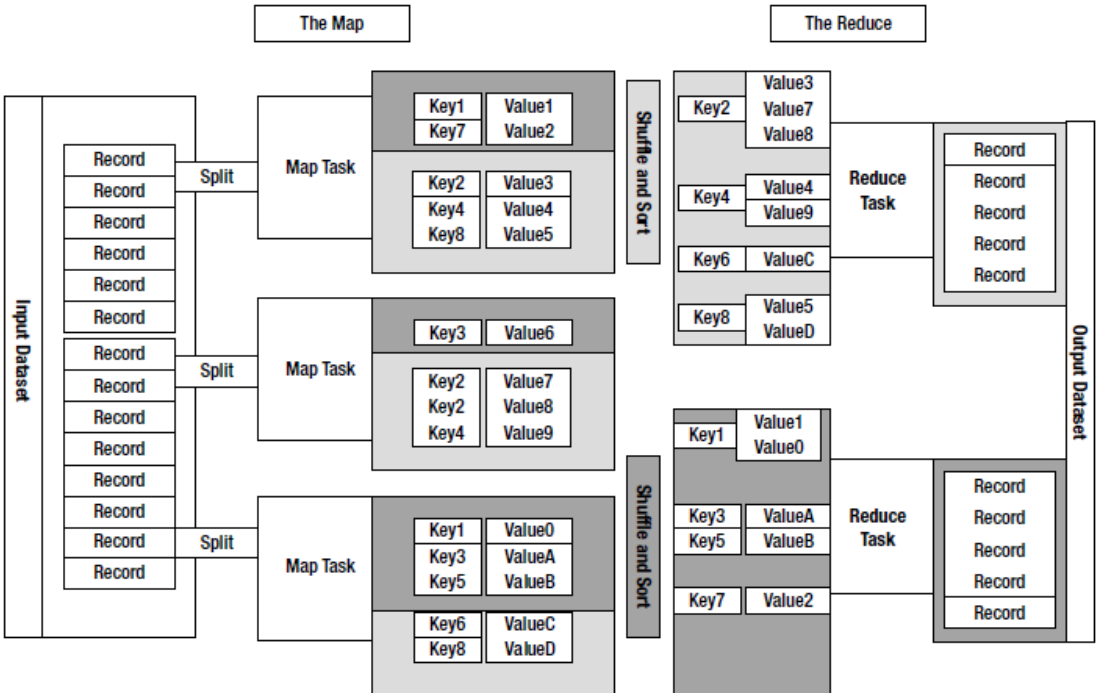


Figure 2.6.1 The MapReduce model (Venner, 2009)

Map phase in the project, when the input is sliced into independent blokes, was done automatically by Tweeter streamer. Streamed tweets were divided in files that stored not more than 20,000 tweets.

The map phase schema for MapReduce method id well represented in (Mendelevitch, Casey and Eadline, 2017) and is presented here in the Figure 2.6.2

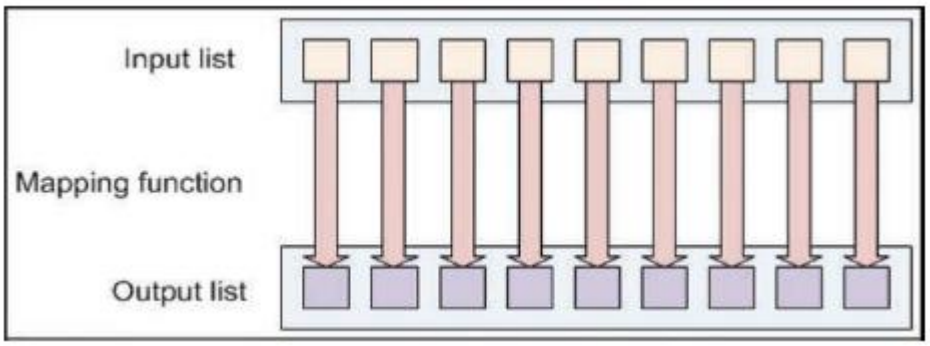


Figure 2.6.2 Map reduce Phase: The input list is sliced into independent blocks. (Mendelevitch, 2017)

The mapping function for Hadoop is performed on each block in parallel. The output list is a collection of key/values pairs.

The Reduce Phase is represented in the Figure 2.6.3

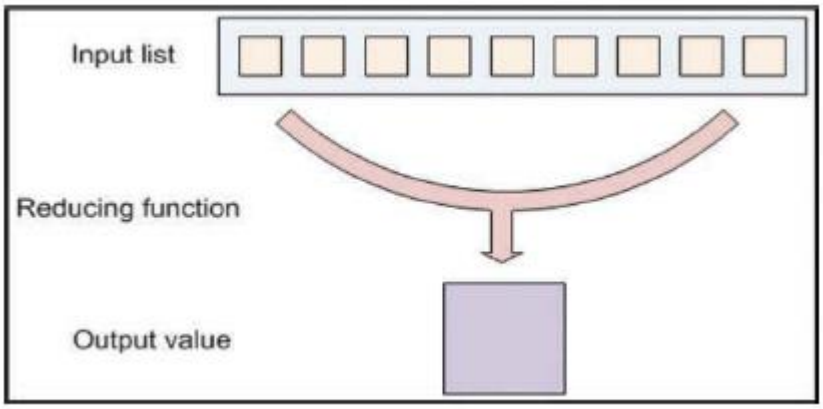


Figure 2.6.3 Reduce Phase of MapReduce method (Mendelevitch, 2017)

For the project files were processed in consecutive order and as key was used date and time when a file was created and as values – compound term of sentiment analysis using VADER, All data was aggregated inside of each file, divided by groups of keywords and stored as a single row into the final file

As result was created a file with only 551 rows that was lately used for analysis.

3 Methodology

3.1 Datasets description

3.1.1 Bloomberg news dataset

Bloomberg News it is known for providing business and economic news to investors. It is operated by Bloomberg LP, a private financial-data services and media company. Bloomberg News is spread through Bloomberg Terminals, Bloomberg Television, Bloomberg Radio, Bloomberg Businessweek, Bloomberg Markets, Bloomberg.com and Bloomberg's mobile platforms. (Bloomberg, 2020)

This company is known as a learning platform for a lot of financial journalists and their news is highly regarded by financial professionals. "*Obligatory for almost anyone with an occupation in finance, the terminals and their software-as-a-service successors offer comprehensive and vital information to 320,000 paying customers around the world*." (McFarlane, 2020)

Bloomberg site requires a paid subscription for access their news. Bloomberg's "Terms of service" (Bloomberg, 2020) states that you cannot automatically scrap news for engineering work without a special permission from the company. The monthly subscription is very expensive ($39.99 per month) and I decided that rights to scrap their news will be something out of reach. I got a discount for a 3 months subscription that gave me the possibility to get access to their news articles and I selected manually extracts of articles from department "Green" where is published news about energy, new renewable technologies, oil and climate adaptation. News are not published on a strict regular base and Bloomberg is not offering access to their archives. Extract from news were stored in an excel file that contains date when news was created first time (some news are updated during few days), title and headlines.

Fields for title and headlines were selected by analogy with other similar projects, such as one done by Chan (Chan, 2003) that was mentioned above and the project "Extract Stock Sentiment from news Headlines", done by DataCamp site (Gonzalez-Vallinas, 2018).

This approach seems very reasonable, as people seldom read all news, they usually quickly scan titles and headlines. If a headline is missing, they might read the first paragraph to see what news is about. If it catches their interest, they are going to read all text.

Dataset collected for this project contains 482 news articles published between 16 February and 5 May 2020.

3.1.2 Tweets dataset

Terms of use of Tweeter Data (Tweeter, 2020) requires an official permission for streaming tweets. I registered with them and got access to an official developer account as a student and to a special developer tool OAuth 1.0a. Such type of account permits a real-time streaming. A small streaming module was active for 12-14 hours a day for all this period. Tweets dataset consists of 9,108,864 tweets from 20 Mar till 5 May 2020. Because of all formalities with registration and because Bloomberg site gives possibility to access data from previous 7-10 days, this data set is shorter in time than news dataset by near 1 month.

Streamed tweets were stored in JSON format.

3.1.3 Oil prices dataset

Oil prices were downloaded from Yahoo Finance site (Yahoo Finance, 2020) in csv format one row for every day for the same period of time as news and Tweets.

3.2 Data collecting and cleaning

3.2.1 Bloomberg news dataset

Bloomberg news were collected manually and contain time, title and headlines. They were stored in an excel file that had 482 extracts from news articles.

News dataset is relatively small in relation to tweets dataset and tweets were filtered in process of collection, as news were used without filtering except that they were published in the same department of Bloomberg news. Theoretically they were sorted by Bloomberg, but could be good to assign them some “hashtags” similar to tweets. After manually analysing news and finding specific words that make one news different from other, was created a dictionary of specific words. Dictionary can be found in Appendix 1.

As there is an inconsistency in the news, when was calculated sentiment of news, it was necessary to aggregate data when there was a few news in the same day and to fill days with missing news with the values from previous day.

In the figure 3.2.1.1 we can see how news are distributed into the keyword’s groups. The sum of all proportions is not 100% as some news were covering few topics and had overlapping keywords. Most news was related with climate adaptation and weather change. As was expected, coronavirus and pandemic were a hot topic of discussions in this time. Oil itself was not on top of news, and it follows by popularity such group of news as about energy in general and renewable sources of energy in particular. News about pandemic completely overshadowed other groups of news including news about critical drop in price of oil.

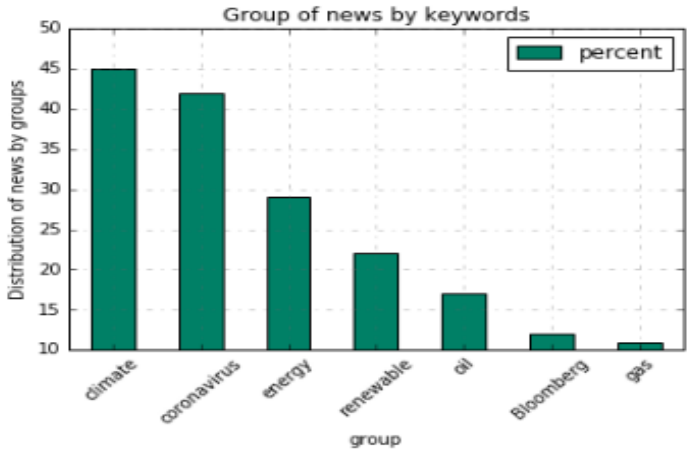


Figure 3.2.1.1 Distribution of news by keywords.

The statistical description of news dataset it is represented in Figure 3.2.1.2

	Sentiment	Renewable	Climate	Gas	Oil	Energy	Bloomberg	Coronavirus
count	80	80	80	80	80	80	80	80
mean	-0.109	-0.007	-0.051	0.008	-0.029	0.003	-0.007	-0.108
std	0.287	0.137	0.229	0.067	0.121	0.158	0.088	0.193
min	-0.863	-0.550	-0.863	-0.173	-0.550	-0.550	-0.373	-0.863
25%	-0.249	-0.019	-0.147	0.000	-0.085	-0.022	0.000	-0.171
50%	-0.130	0.000	0.000	0.000	0.000	0.000	0.000	-0.048
75%	0.061	0.017	0.042	0.002	0.000	0.057	0.000	0.000
max	0.576	0.414	0.576	0.359	0.416	0.534	0.416	0.395

Figure 3.2.1.2 Statistical information about news dataset

Sentiment intensity of news in these 80 days was varying from -0.86 to 0.57. In general, level of sentiment in news was enough low with a mean of -0.109. And what is surprising is that news about oil were not as negative as expected.

News classification in groups it is a subjective process and during news pre-processing was estimated the general sentiment of news as well as of the groups by keywords.

To better understand the emotional level of news in this time, below is a graph where we can see the level of general sentiment in the news and in the first two most popular topics about climate and coronavirus.

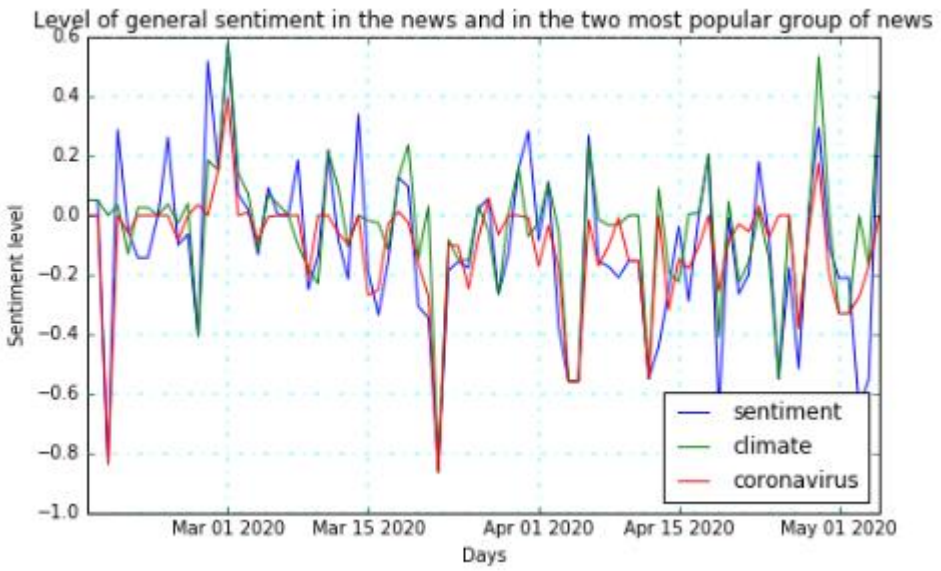


Figure 3.2.1.3 General sentiment in the news and in the 2 leading groups of news: Climate and coronavirus.

The graph (Figure 3.2.1.4) of level of general sentiment of news and of groups of news about oil and gas looks different. Most of days there were not at all news about gas or oil and when there was news, their level was more neutral with only one negative evident peak at end of March when price for oil was negative,

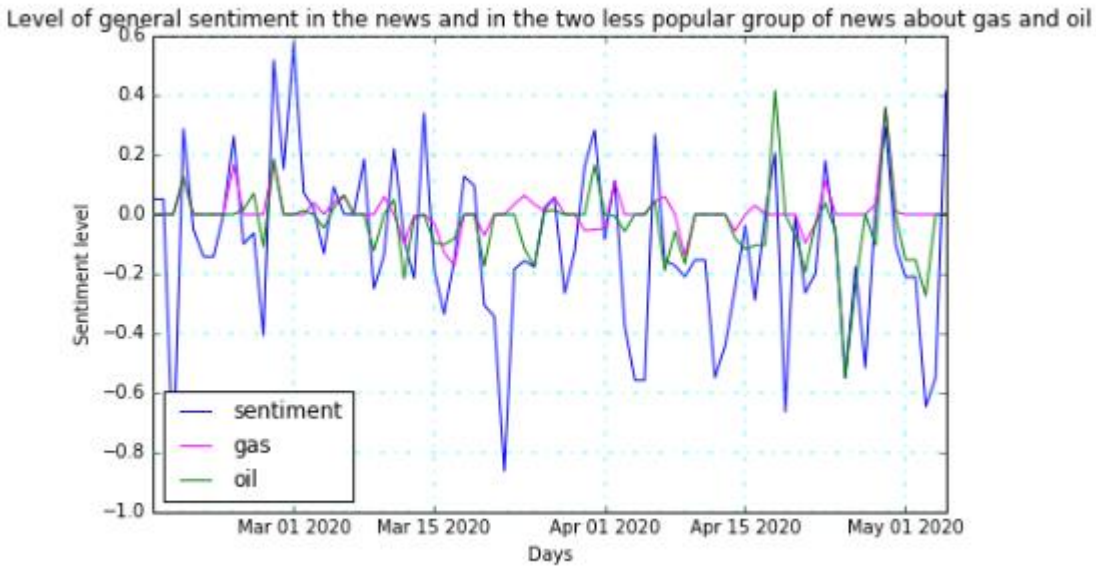


Figure 3.2.1.4 Level of general sentiment in the news and in the group of news about gas and oil.

From the matrix we can see another correlation between group of news about coronavirus and climate. This is due to the news related to the pandemic and its effect on air pollution.

General level of sentiment in the news (Figure 3.2.1.4) shows two peaks of negativity:

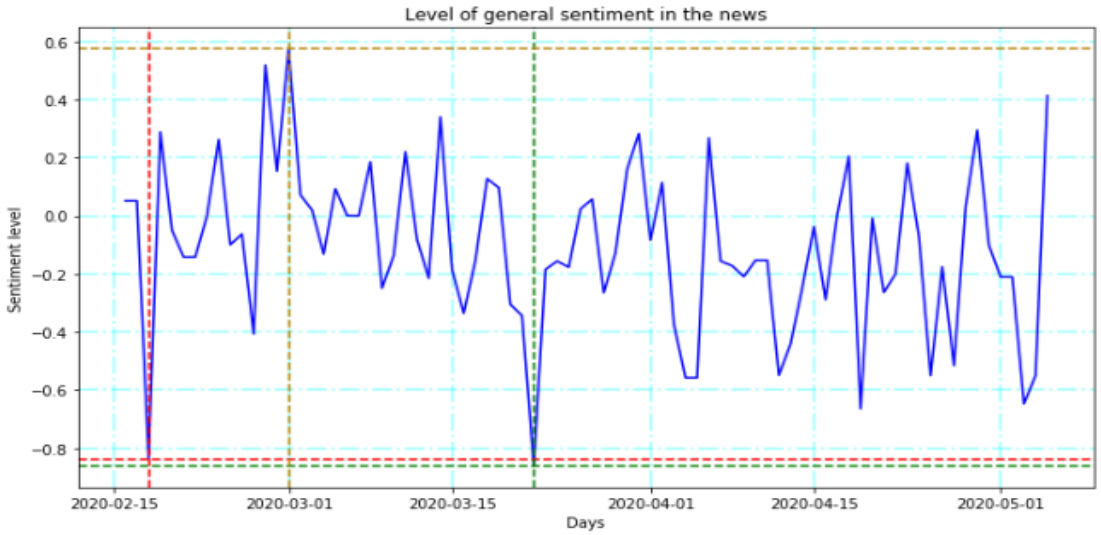


Figure 3.2.1.4 General level of sentiment from Bloomberg news for February – May 2020

Here we can see one positive peak at 1st March and two troughs of negative sentiment in news: at 18 February and at 22 March.

In the middle of February virus in China started to show dangerous proportions and news reflect this situation with a trough of negativity. At the beginning of March lockdown in China started to show signs of improvement and news were enough positive about coronavirus. At 20 March WHO announced coronavirus pandemic. (WHO, 2020) and its consequences are again reflected in the news.

3.2.2 Tweets dataset

Tweets were streamed in real – time mode using tools described in literature review.

A special permission from Tweeter gives access to the streaming. Due to waiting periods for getting access to a developer account and the fact that streaming is done in real-time mode, historical tweets were not available and there is data collected only between 20 March 2020 and 05 May 2020.

Streaming tweets started later than collection of news and filtering keywords were selected to reflect diversity of topics that were discussed in the news at that moment. As tracking keywords for the streamer's filter were used hashtags: "#renewable", #climate", "#gas", "#energy", "#oil", "#Bloomberg", "#COVID-19", "#coronavirus".

The streamed tweets represent a small 1% sample of total tweets in real time. My module was streaming tweets for 12-14 hours per day during all this period of time.

Tweets are streamed automatically and stored in files not more than 20,000 rows per file. In total were created 551 files with the total volume of 66.6 Gb. No file contained tweets from more than one day.

Analysing distribution of keywords in tweets was observed that sentiment in the majority of tweets was estimated equal to null. From 9,093,808 obtained tweets only 2,748,931 tweets (30%) were containing useful information. A loot of tweets contained links to different web pages and videos.

Streaming of tweets started the same time when WHO announced COVID-19 pandemic (WHO, 2020) and such an event attracted all attention of Tweeter users, resulting in about 80 % of tweeters to discuss coronavirus.

Proportion of tweets about oil was, only 3.5%. It looks like a very small sample, but it consists of 98,788 tweets that is considerable larger than number of collected news.

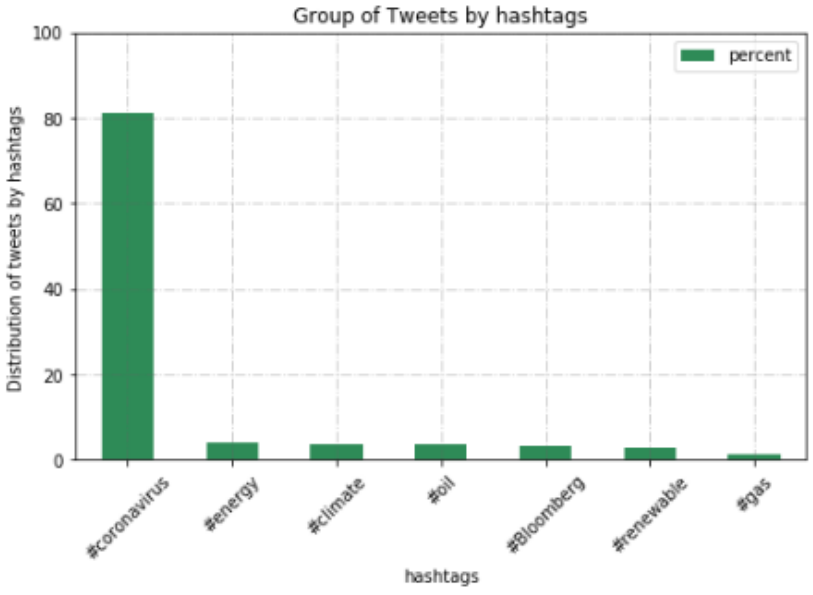


Figure 3.2.2.1 Distribution of tweets by hashtags

After a few days of streaming data, was clear that volume of tweets dataset is going to be enormous and was decided to use an approach similar with MapReduce method. Was created a module for Map Phase that realised the following steps of used approach:

Files from a specific directory were read consecutively. From each tweet was extracted a {key: values} pair that represented date and time as key and tweet's text as value. Sentiment of text from each tweet was estimated using VADER' compound coefficient.

Text was checked to see if it contained any of keywords of interest. (Appendix 2)

If any keyword was found score for general sentiment of the text was also copied into the field of sentiment for certain group of keywords.

After all tweets in the file were analysed, an aggregation process was summarising all fields and results were saved in a CSV format file.

The final file had 551 rows that is the same number as collected files with streamed tweets.

Each row of the file has 17 variables: date and time of the first tweet from the file; sum of all general sentiment estimations of tweets from each file, total number of tweets that were in each file; 7 pairs of sums of sentiment estimations for each hashtag and total number of tweets related with each hashtag.

On the Reduce Phase, the final file was loaded into the computer memory and rows were sorted by day. The time field was reduced to date and an aggregation of all rows of each day was performed. The final dataset had now 46 rows that is equal to number of days when tweets were streamed.

The statistical description of tweets dataset it is represented in Figure 3.2.2.2

	Sentiment	#Bloomberg	#Climate	#coronavirus	#energy	#gas	#oil	#renewable
count	46	46	46	46	46	46	46	46
mean	0.018	0.089	0.099	0.008	0.221	-0.009	-0.104	0.211
std	0.033	0.101	0.101	0.079	0.095	0.118	0.084	0.101
min	-0.041	-0.315	-0.078	-0.228	-0.083	-0.365	-0.364	-0.047
25%	-0.008	0.046	0.036	-0.025	0.164	-0.051	-0.149	0.142
50%	0.019	0.119	0.076	0.008	0.251	-0.005	-0.103	0.236
75%	0.042	0.152	0.151	0.054	0.286	0.060	-0.054	0.278
max	0.097	0.294	0.458	0.186	0.339	0.248	0.063	0.383

Figure 3.2.2.2 Statistical description of tweets dataset.

Estimation of general sentiment in tweets varies from -0.041 till 0.097 with the mean of 0.018. The variation of sentiment in the tweets is smaller than in the news (-0.86 to 0.57 and mean of 0.109) and tweets in general are more positive than the news.

To better understand the emotional level of tweets in this time, below is a graph where we can see the level of general sentiment in the tweets and in the most popular topic of tweets: coronavirus. (Figure 3.2.2.3)

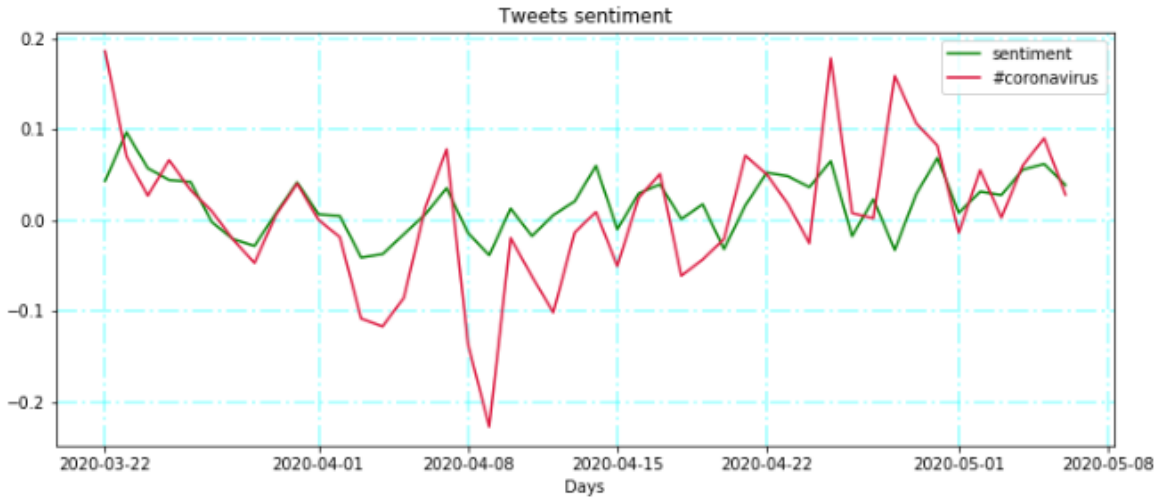


Figure 3.2.2.3. General sentiment level in the tweets vs sentiment level in the #coronavirus group.

Sentiment level in the hashtag group #coronavirus has bigger variation than general sentiment in the tweets.

Figure 3.2.2.4 general sentiment level in the tweets vs sentiment level in the #oil and #gas groups of tweets.

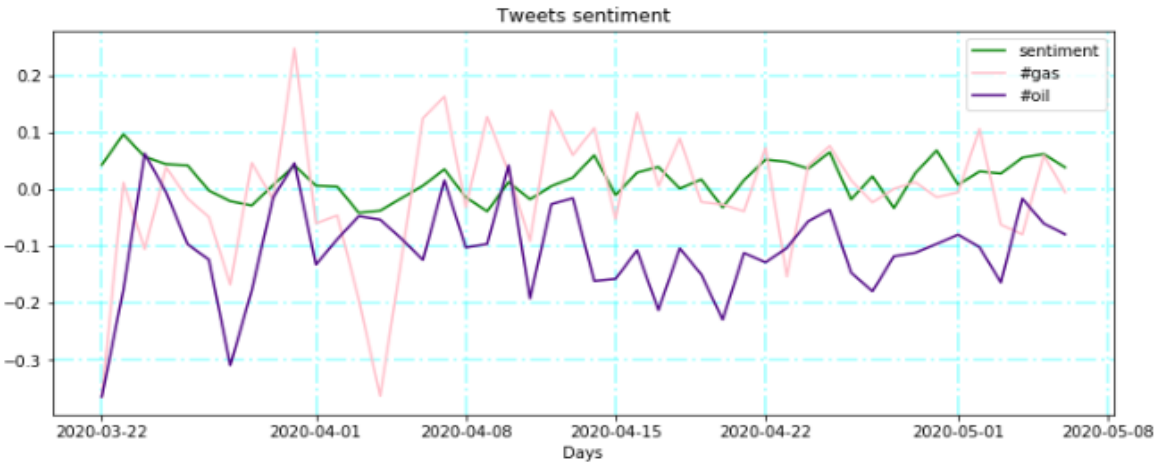


Figure 3.2.2.4 General sentiment level in the tweets vs sentiment level in the #oil and #gas group of tweets.

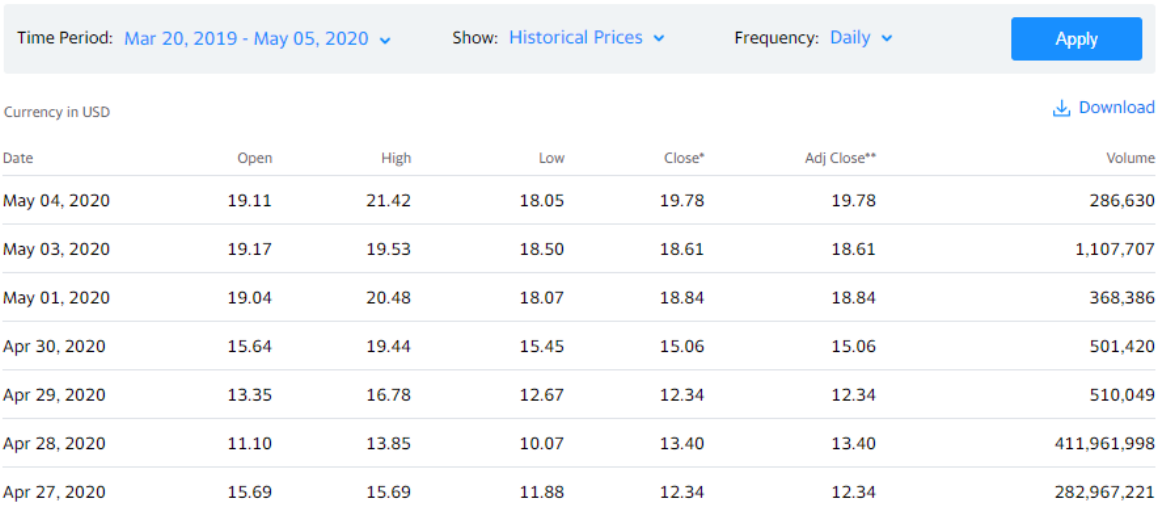
Sentiment level in the #oil group in general was lower that general sentiment level, but didn't show evident troughs around 19 April 2020 when prices of oil became negative. Tweets reaction to oil crises was more neutral than in the news.

3.2.3 Yahoo Finance

Prices for oil were downloaded from https://finance.yahoo.com/ for the same period of time as the news and the tweets.

Information is provided for Sunday to Friday and is missing Saturdays. As a measure for our analysis was used the closing price of the day.

Dataset is sorted in a reversable mode: first are listed most recent days:



Time Period: Mar 20, 2019 - May 05, 2020 Show: Historical Prices Frequency: Daily Apply

Currency in USD Download

Date	Open	High	Low	Close*	Adj Close**	Volume
May 04, 2020	19.11	21.42	18.05	19.78	19.78	286,630
May 03, 2020	19.17	19.53	18.50	18.61	18.61	1,107,707
May 01, 2020	19.04	20.48	18.07	18.84	18.84	368,386
Apr 30, 2020	15.64	19.44	15.45	15.06	15.06	501,420
Apr 29, 2020	13.35	16.78	12.67	12.34	12.34	510,049
Apr 28, 2020	11.10	13.85	10.07	13.40	13.40	411,961,998
Apr 27, 2020	15.69	15.69	11.88	12.34	12.34	282,967,221

Figure 3.2.3.1 Price of oil on Yahoo Finance (Yahoo Finance, 2020)

As we can see, at 20th April price for oil was negative:

Apr 21, 2020	1.40	13.86	0.01	9.06	9.06	1,771,636
Apr 20, 2020	15.27	15.31	-39.44	-2.72	-2.72	72,609,436
Apr 19, 2020	17.56	17.56	16.93	17.04	17.04	343,713

Figure 3.2.3.2 Negative price of oil at 20 Apr 2020

As incredible as it can seem, this was situation in April on the market. In an article from Forbes (Hansen, 2020), we can find explanation what it means:

"The price of one American oil futures contract plunged Monday into the negative for the first time in history, revealing just how badly an already-fragile market has been hit by the coronavirus crisis; as demand hits rock-bottom and storage tanks fill up, companies are now paying traders to take oil off their hands."

The cause of this effect is pandemic lockdown and demand for fuel going rock-bottom. Plus, here is the effect a week-long price war between 2 biggest oil producers: Saudi Arabia and Russia.

Dataset columns are a set of variables indicating price of oil at different time of the day. For this project was used only price for end of the day. It can be found in the column "Close".

The statistical description of dataset for oil prices is shown in the Figure 3.2.3.3

	Close
count	66
mean	29.280
std	13.576
min	-2.720
25%	20.280
50%	25.280
75%	43.748
max	53.780

Figure 3.2.3.3 Statistical information for initial dataset for oil prices.

Prices for oil are between 53.78 and -2.70 with the mean equal to 29.280. This dataset values are situated in another range of values than estimations of sentiment level from News and Tweeter datasets that are in the interval [-1, 1]. Difference in the scales can lead us to difficulties in modelling and to larger error values. This is why in such cases it is recommended to rescale values to the same range. The best way to do this without losing any data properties it is to bring large values to the smaller intervals. Practically we are multiplying all values by one and the same coefficient that is keeping all other properties intact. To solve this issue, oil prices dataset was rescaled to the interval [-1, 1].

For this purpose, was used following formula (Wikipedia, 2020):

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

where a, b are the min-max values.

In our case a = -1 and b = 1 as minimal and maximal values in interval [-1, 1] and final formula is

$$x'' = 2\frac{x - \min x}{\max x - \min x} - 1$$

Value x from formula was replaced by value “Closed” from petrol dataset.

After this dataset for oil price looked as Figure 3.2.3.4:

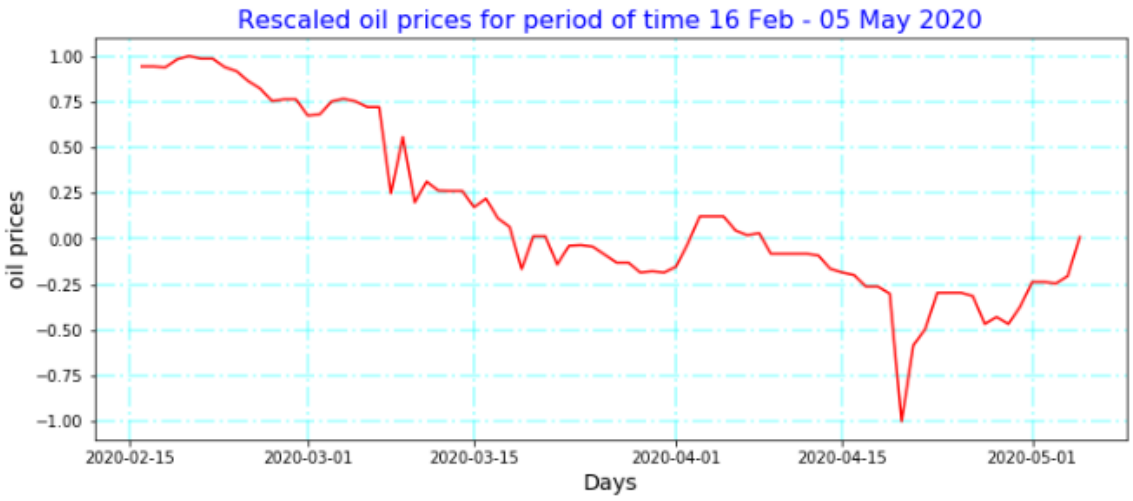


Figure 3.2.3.4 Rescaled oil prices

3.3. Sentiment analysis. VADER.

“*There are basically three types of sentiment classification techniques – Machine Learning, lexicon-based and hybrid*” (Surbhi, Neetu, 2018):

The difference between them is how they are processing information. For this project a hybrid sentiment analyser VADER was used.

A hybrid analysed represent a technique that combines machine learning with lexicon-based approach.

The sentiment Analysis package VADER (Valence Aware Dictionary and aEntiment Reasoner) is one of the best for social media analysis. (Hutto, 2014). VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

According to some sources (Pandley, 2018), VADER has a lot of advantages over traditional methods of Sentiment Analysis, including:

- It works exceedingly well on social media type text, yet readily generalizes to multiple domains
- It doesn't require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon
- It is fast enough to be used online with streaming data, and
- It does not severely suffer from a speed-performance trade-off.

To get most of their ratings, the developers of VADER used Amazon's Mechanical Turk marketplace (Amazon, 2020). This website is a crowdsourcing place where developers can hire humans to do tasks that are performed better by humans than by computers.

Complete details about VADER can be found on their Github Page (GitHub, 2014).

The methods and process approach overview of VADER are as follows (Hutto, 2014):

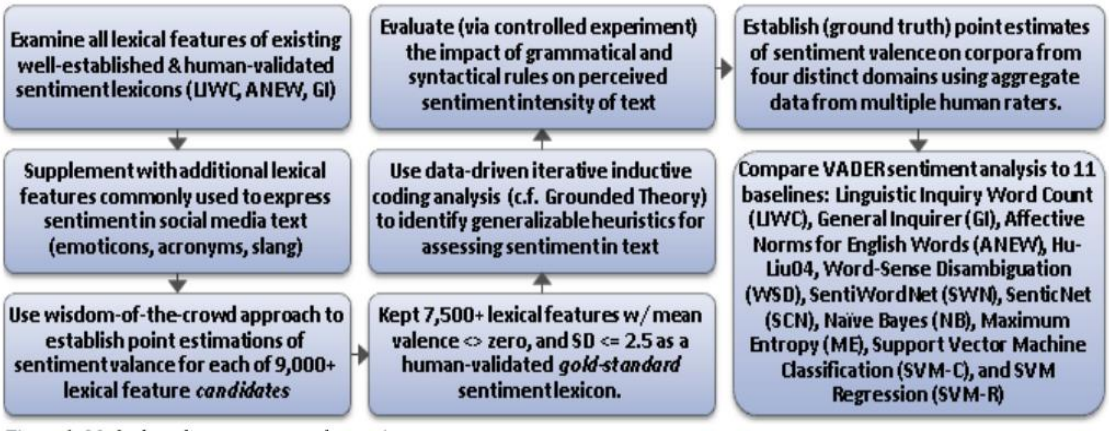


Figure 3.3.1 Methods and process approach of VADER (Hutto, 2014)

VADER performs very well with emojis, slangs, and acronyms in sentences. VADER can easily detect sentiment from emojis and slangs which form an important component of the social media environment at Tweeter or Facebook.

Below are few examples of emojis sentiment estimation:

```
print(sentiment_analyzer_scores('I am 😄 today'))
print(sentiment_analyzer_scores('😊'))
print(sentiment_analyzer_scores('😓'))
print(sentiment_analyzer_scores('☹'))

#Output

I am 😄 today---------------------------- {'neg': 0.0, 'neu': 0.476,
'pos': 0.524, 'compound': 0.6705}

😊---------------------------------------- {'neg': 0.0, 'neu': 0.333,
'pos': 0.667, 'compound': 0.7184}

😓---------------------------------------- {'neg': 0.275, 'neu':
0.268, 'pos': 0.456, 'compound': 0.3291}

☹---------------------------------------- {'neg': 0.706, 'neu':
0.294, 'pos': 0.0, 'compound': -0.34}

💔---------------------------------------- {'neg': 0.0, 'neu': 1.0,
'pos': 0.0, 'compound': 0.0}
```

Figure 3.3.2 How a compound score of emojis is calculated in VADER.(Hutto, 2018)

My impression about VADER after doing this project it is not as high as expected. It was a bit slow in processing tweets with a lot of results calculated as having neutral sentiment level.

3.4 Time series

All 3 datasets analysed for this project represent time series and therefor we can use different methods of time series analysis for our datasets.

In process of tweets streaming values of field “created at” contained date and time till seconds. Bloomberg Green site has time of publishing in the same format as tweets. Yahoo Finance site can provide information with different frequency starting from seconds and finishing with 1 year. It looked logical to try and analyse these 3 datasets with a frequency at least to 1 hour, but there were few moments that influenced decision to analyse them at a frequency of 1 day.

1. Inconsistency of Bloomberg Green news. There were days when we can find few different news published in the same day. Other days didn't have any news.
2. Tweets were streamed from a home computer during 12-14 hours a day, but there was a break in streaming during the night. Beginning and end of streaming period could be different for different days.
3. Sales of oil on Saturdays and some national holidays are closed.

Taking in consideration these limitations, analysis for the project was done with a minimal frequency of 1 day.

"One powerful yet simple method for analysing and predicting periodic data is the additive model. The idea is straightforward: represent a time-series as a combination of patterns at different scales such as daily, weekly, seasonally, and yearly, along with an overall trend. Your energy use might rise in the summer and decrease in the winter, but have an overall decreasing trend as you increase the energy efficiency of your home. An additive model can show us both patterns/trends and make predictions based on these observations."
(Koehrsen. 2018)

Inconsistency in datasets was solved using a Python (Pandas) function resample() used for working with time series.

The function has downsampling and upsampling transformation modes. Downsampling is used when frequency of time series is decreased and upsampling - then we are looking to increase frequency. If it is necessary to downsample a time series, values are aggregated, Upsampling will imputate missing values in accordance with the chosen method.

Bloomberg Green news dataset was aggregated to the mean values when there were a few news articles in the same day and filled with values from previous day for the days when there were no news at all.

Tweets dataset was aggregated to the mean values of each day.

In prices for oil dataset days with missing values got values from previous day.

In case of weekly analysis, all datasets were aggregated to the mean values at every day of the week.

Time series analysis is useful to see dynamic of changes in our datasets and to find a correlation between them.

4. Results

4.1. Sentiment level evaluation of Bloomberg news vs Price of oil.

Daily model of time series for price of oil and time series for Bloomberg news reflects that a few days before the event of significant price of oil drop, the general sentiment level in the news and sentiment level from oil keywords group (Appendix 2) had a few small drops of sentiment level, but in the Figure 4.1.1 we cannot observe significant common patterns between these time series in form of peaks or throughs.

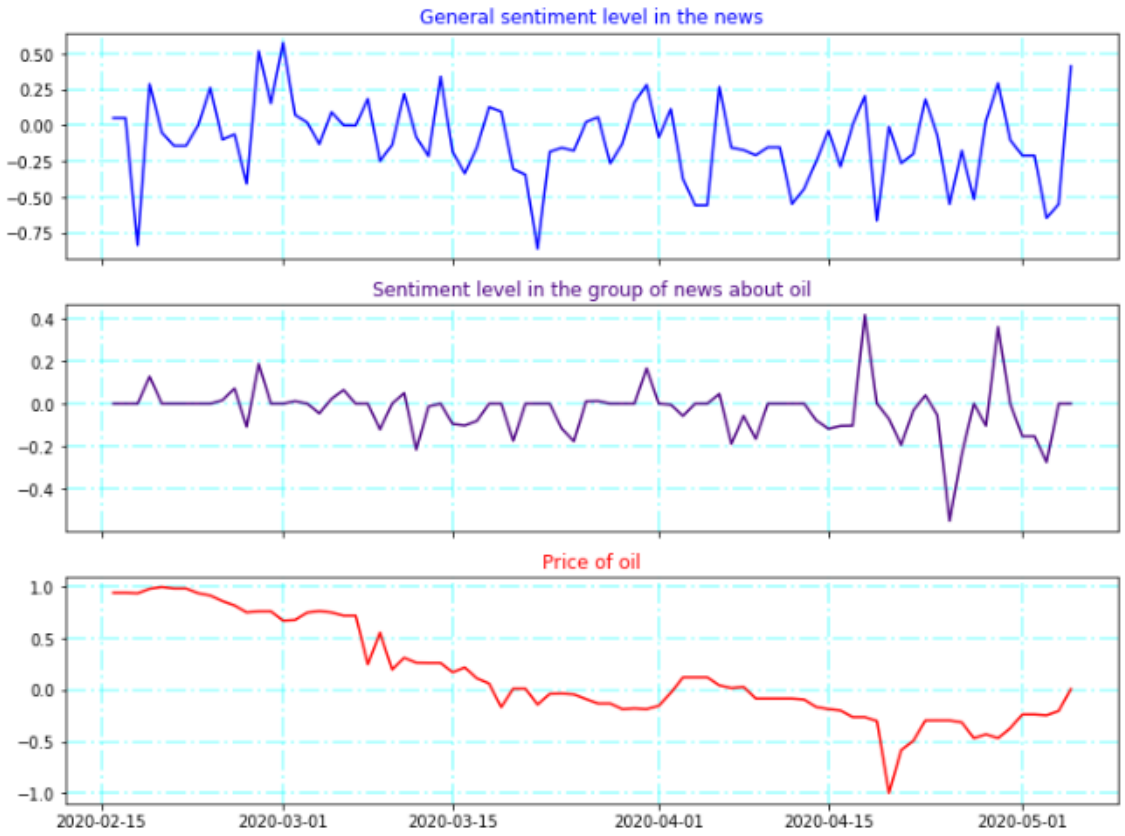


Figure 4.1.1 Plot of time series for Bloomberg news and price for oil with a daily frequency.

Analysing plots we cannot detect any signs in sentiment level in the news that can warn us about coming critical drop in price of oil. There is a through in the general sentiment level that happened in the same day when COVID-19 pandemic was announced and this event's

impact we can observe on the plot and it is not directly related with price of oil. In the group of news about oil there is a small though of negativity, but only after drop in the price of oil took place.

Changing model from daily to weekly frequency can help smooth our patterns, but as it can be observed in the Figure 4.1.2 it doesn't improve the general result essentially.

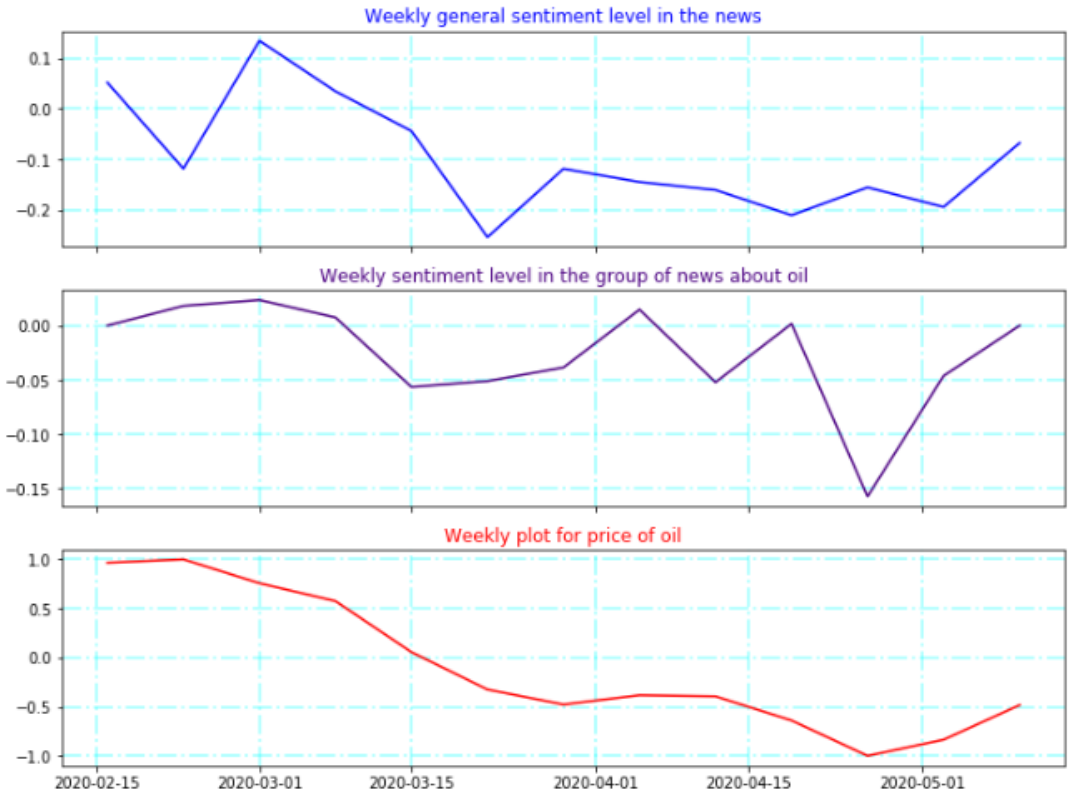


Figure 4.1.2. Plot of time series for Bloomberg news and price of oil with a weekly frequency.

Sentiment in the group of news about oil occurs in the same time as drop of price of oil, but again it is not predicting the event.

More interesting looks general sentiment level in the news that has a through 2 weeks previous to drop in the price of oil. But we have to keep in mind that, as was mentioned above, in that time happened another unique event – beginning of COVID-19 pandemic and most like this event generated the negative response in the news.

Another way of smoothing the through in the price for oil can be rolling average approach or Simple Moving Average (SMA) (Nau, 2020). For the project was used a 4 days time

interval as a median between daily and weekly models. Using it for the news time series and price of oil time series, gives us the effect represented in the Figure 4.1.3.

We can see from the plot representation that the decline in prices for oil started before pandemic and coronavirus outbreak only intensified this process.

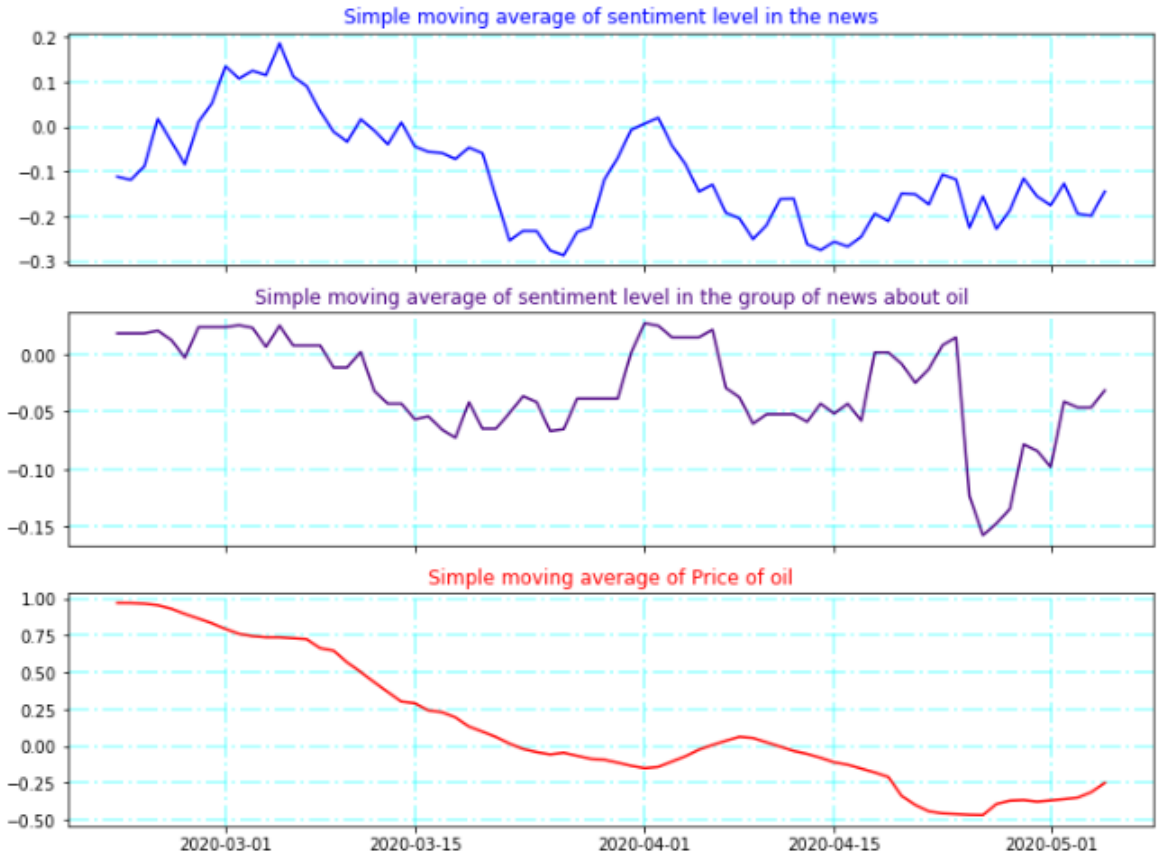


Figure 4.1.3. Plot of time series SMA for Bloomberg news and price of oil with a daily frequency.

However, the general sentiment of the news was high at the beginning of March and went low with coronavirus outbreak. The SMA sentiment in the group of news related with oil was going down for a while and got again high when prices raised a bit.

Calculating correlation between these 3 cases we are getting the result from Table 4.1.1

Correlation table between Price for Oil and sentiment level in the Bloomberg news

Model	General sentiment in the news	Sentiment in the news about oil
Daily frequency	0.2182	0.1956
Weekly frequency	0.7245	0.6139
SMA	0.3842	0.3304

Table 4.1.1. Correlation table between Price for Oil and sentiment level in the Bloomberg news.

Figure 4.1.3 Represent correlation coefficients for positive and negative lags.

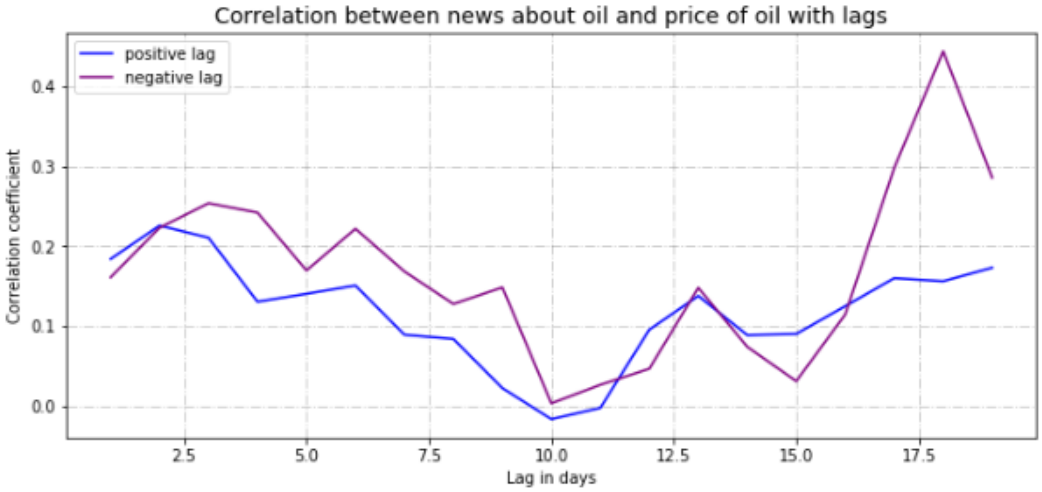


Figure 4.1.3 Correlation coefficients between sentiment level in the news about oil and price for oil with time lags

The best correlation coefficient it in the range of 2-3 days before and after event and 18 days before event, but this is the time of COVID-19 outbreak. Possible that there were already concerns about expected economy stagnation and drop in the oil use. But all of these correlation coefficients are situated in the interval [0, 0.4] and are not significant.

Weekly model seems to look better than the daily and SMA models, but our datasets contain only 13 weeks of data in a period of time with 2 worldwide events that happened for the first time ever in our history: COVID-19 pandemic and negative price of oil. These 2 events had a strong anomaly impact and introduced a strong disturbances in the sentiment level of news and social media and cannot be used as a strong proof that there is or not a

link between sentiment level in the news and price of oil that can be used for the future to predict any changes in the price of oil.

4.2 Tweets sentiment level vs Prices for oil

We didn't find a direct link between sentiment level in the Bloomberg news and Price for oil, but maybe Social Media, in particular, Tweeter reaction was stronger in this period of time and foresaw the situation with Prices of oil?

Plotting again general level of sentiment in tweets and in the group of tweets with hashtag '#oil' next to prices for oil in case of daily model, we can see in Figure 4.2.1 that here situation looks different than in the case of Bloomberg News:

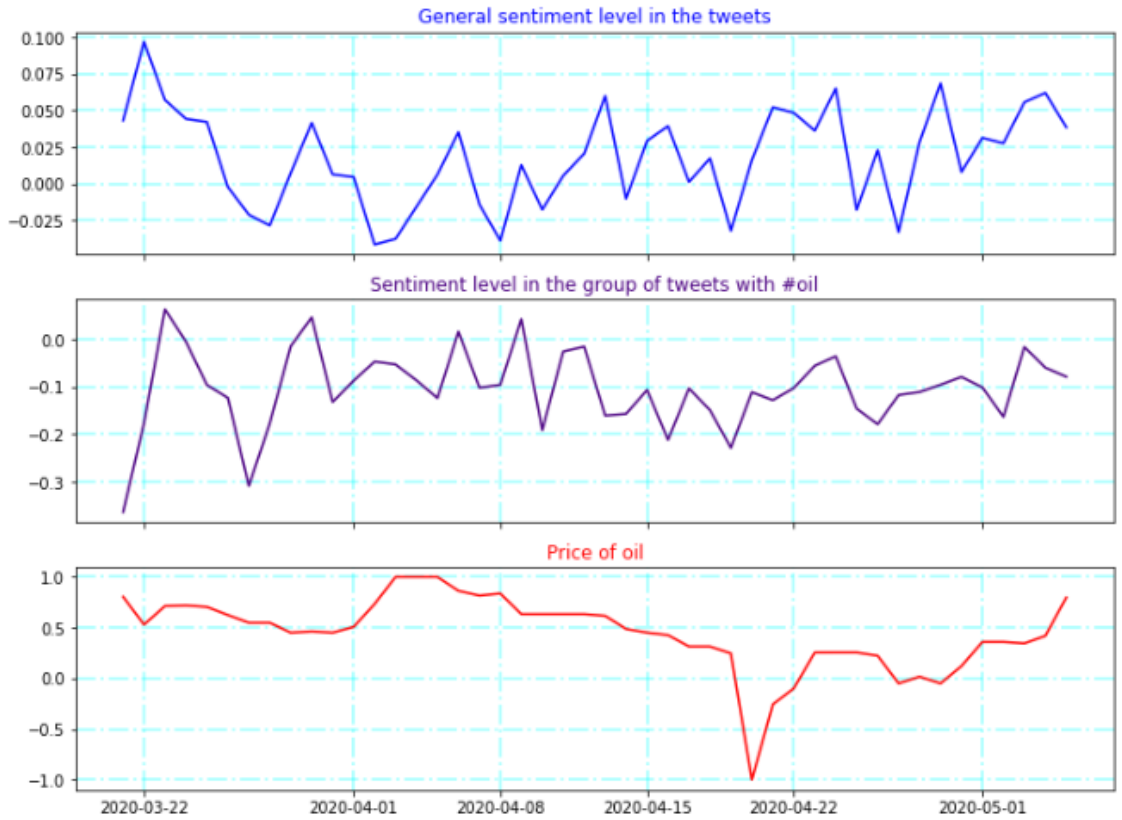


Figure 4.2.1. Plot of time series for sentiment level in the tweets and price for oil with a daily frequency.

Possible that Social Media users were too busy discussing COVID-19 pandemic and drop in the Price of oil was out of their concern. Next few days after drop in the price of oil sentiment level was a bit higher than before. It is understandable that most of Tweeter

users are not investing in oil and lower prices of oil were met with some happiness. People took news about drop in price of oil more like a logical consequence to pandemic lockdown.

Weekly model (Figure 4.2.2.) looks a lot better with a distinctive similarity in the sentiment level of the tweets with #oil. Possible that this hashtag was used by financial specialists or companies involved in the oil trade.

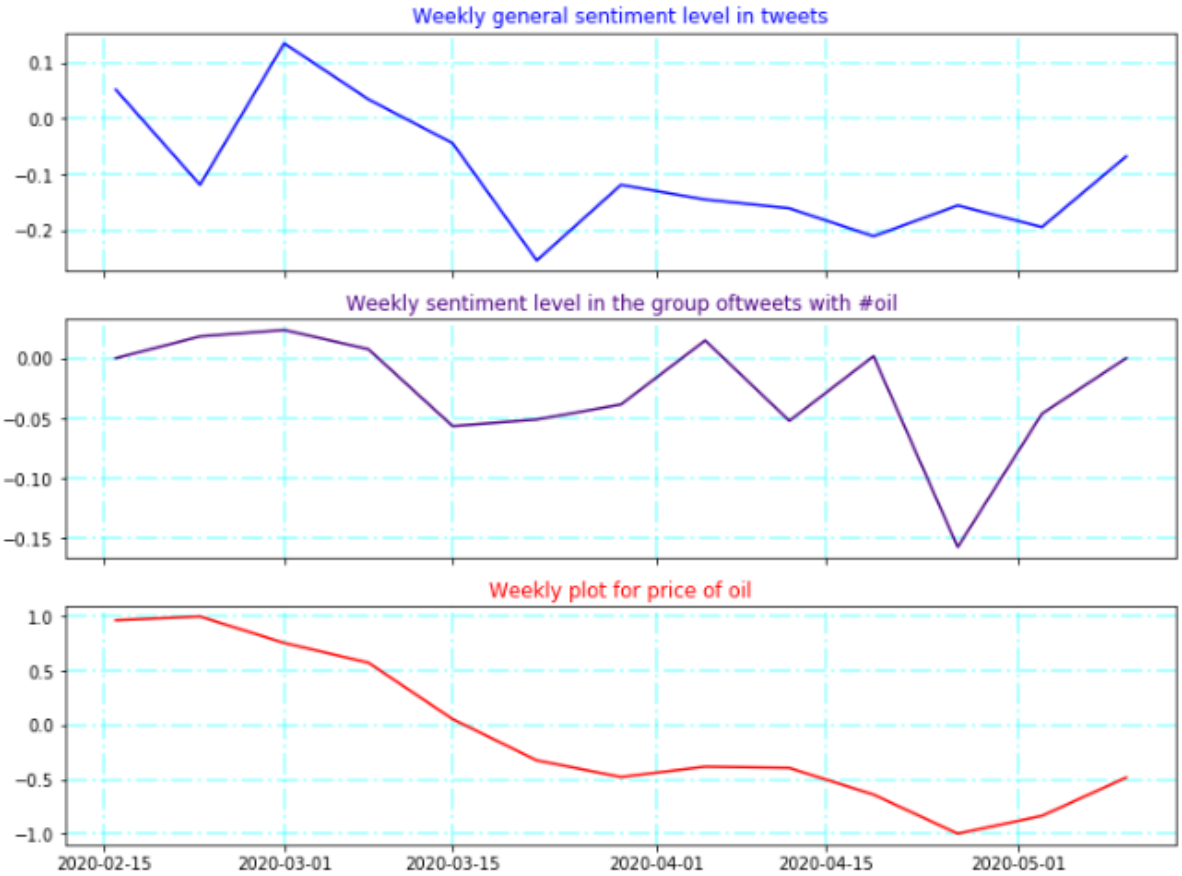


Figure 4.2.2. Plot of time series for sentiment level in the tweets and price for oil with a weekly frequency.

Using simple moving average (SMA) gives us possibility to see the intermediary model between daily and weekly one. For SMA was used again a 4 days period of time as for news analysis. The results of SMA model are presented in the Figure 4.2.3 and similarity in patterns looks better than in daily and weekly models. We can observe even a small 1-2 days lag in sentiment level in the tweets with #oil. Tweeter users started to discuss situation with oil before the drop in price for oil took place.

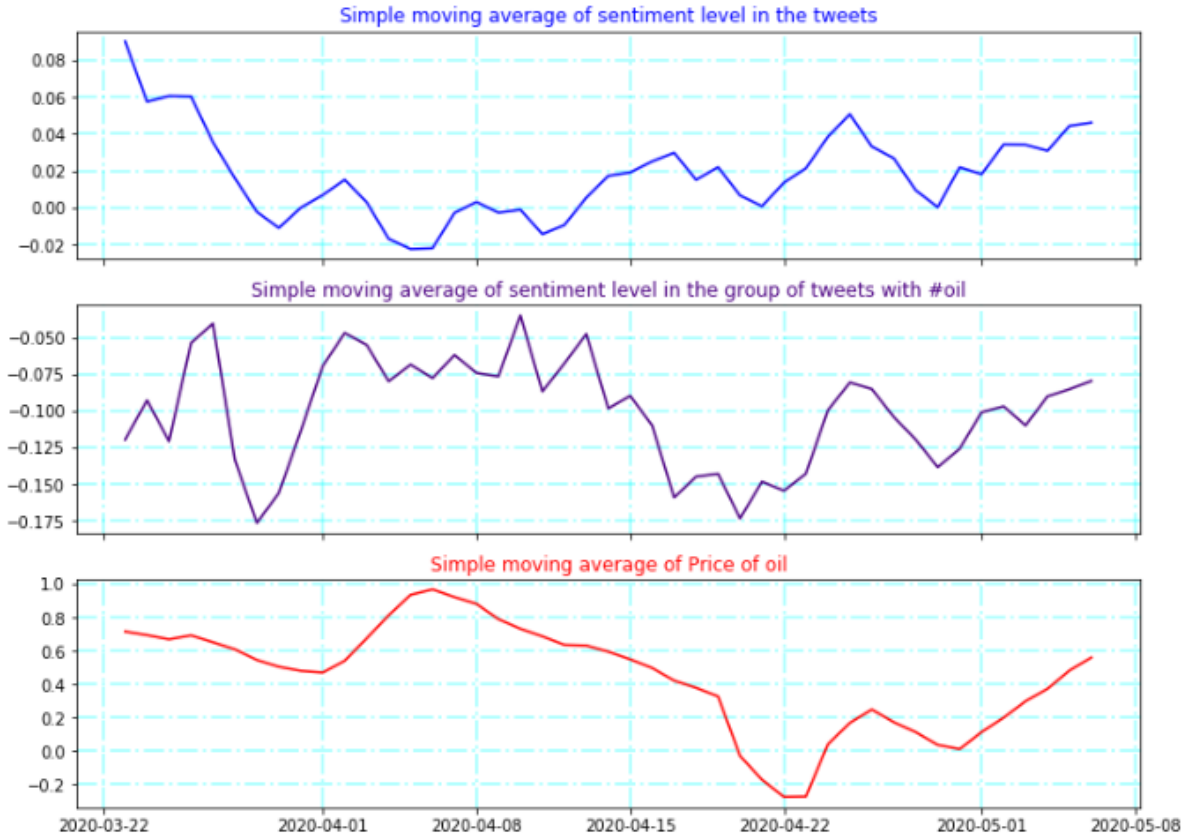


Figure 4.2.3. Plot of SMA model of time series for sentiment level in the tweets and price for oil.

The correlation coefficients of these 3 models are presented in the Table 4.2.1

Correlation table between Price for Oil and sentiment level in the tweets

Model	General sentiment in the tweets	Sentiment in the tweets with #oil
Daily frequency	0.0641	0.3488
Weekly frequency	0.1003	0.6198
SMA	-0.1608	0.577

Table 4.2.1. Correlation table between Price for Oil and sentiment level in the tweets.

The plot of correlation coefficients between sentiment level in tweets with hashtag #oil and price of oil with time tags is represented in the Figure 4.2.4

All correlation coefficients are situated in the interval [-0.3, 0.3] and are very small. Around 2-3 days, there is a small increase and after this reaction again is increasing or decreasing every few days, but remains insignificant.

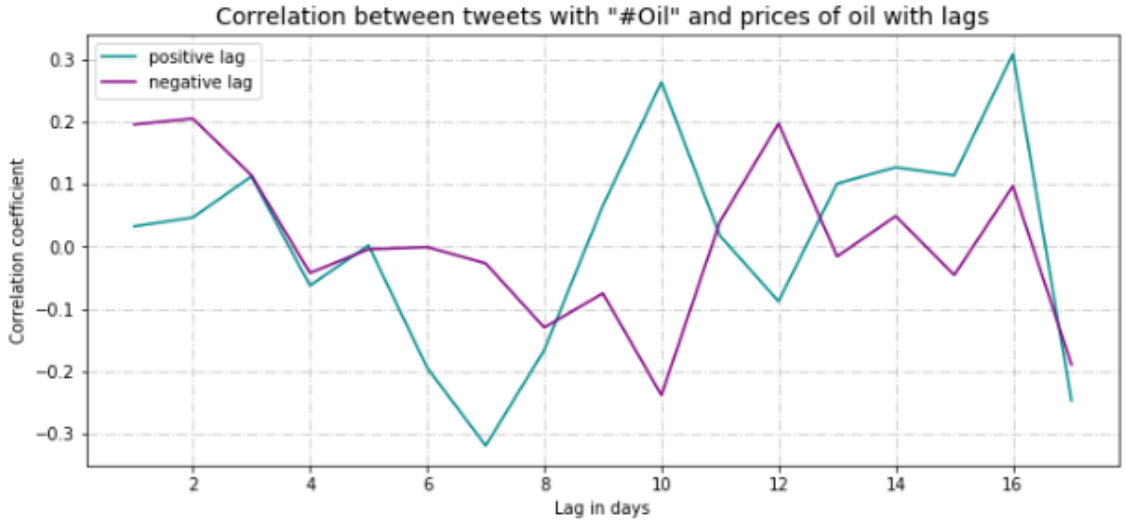


Figure 4.2.4 Correlation coefficients between sentiment level in the news about oil and price for oil with time lags

In general sentiment level in tweets collected for the project is noisier than in the news. Even weekly model shows better result only in the group with the special hashtag “#oil”.

4.3. ARIMA models of datasets.

It is very hard to estimate how sentiment in the news or in Social media is related to the price of oil, then the situation with oil at moment is full of uncertainty.

The COVID-19 pandemic brought with it a worldwide lockdown and distancing with use of petrol for transport dropped down. Many production facilities were closed. Production of oil was exceeding demand.

And, as if it wasn’t enough, there was the oil was between two biggest oil producers: Saudi Arabia and Russia.

These factors influenced price of oil to such a degree, that it became difficult to predict what is going to happen next.

Figure 4.3.1 represent an ARIMA (Autoregressive Integrated Moving Average) forecasting model for price of oil at the 5 May 2020. It shows that situation with the price for oil can go in the future in any direction. Of course, in such situation it is difficult to relay only on sentiment level estimation.

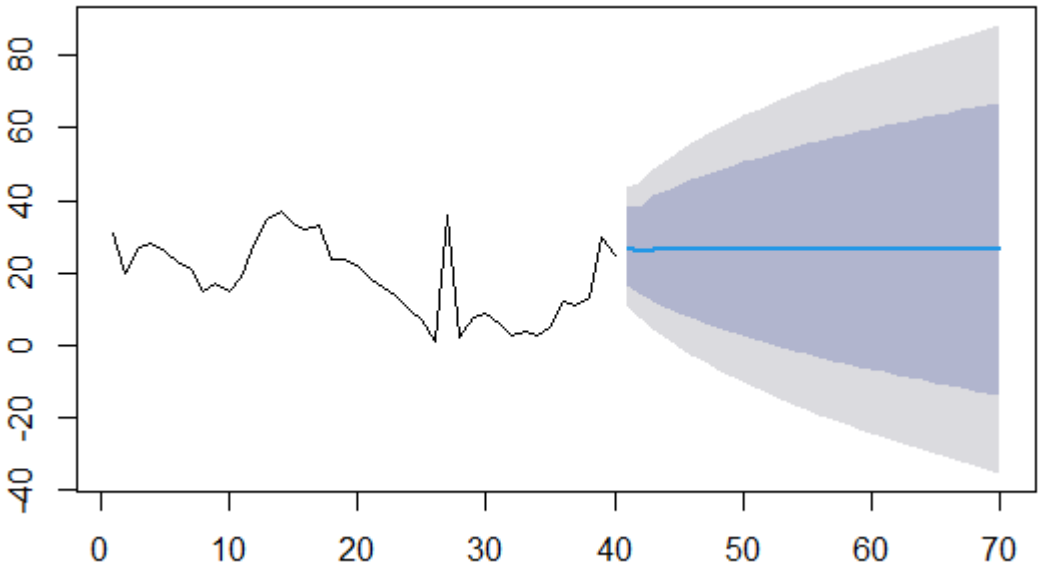


Figure 4.3.1 Price for oil forecasting using ARMA model.

Situation with COVID pandemic influenced not only forecasting for price of oil, but also sentiment level behaviour in the news and in the Social Media. In the Figures 4.3.2 and 4.3.3 we can see ARIMA forecasting for news and tweets.

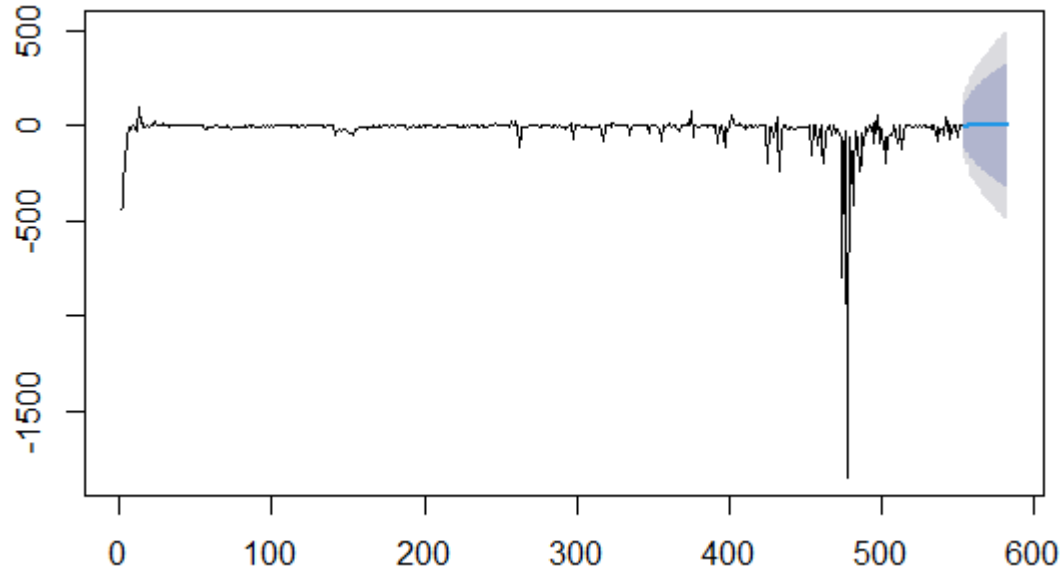


Figure 4.3.2 Sentiment level in the news about oil forecasting using ARMA model.

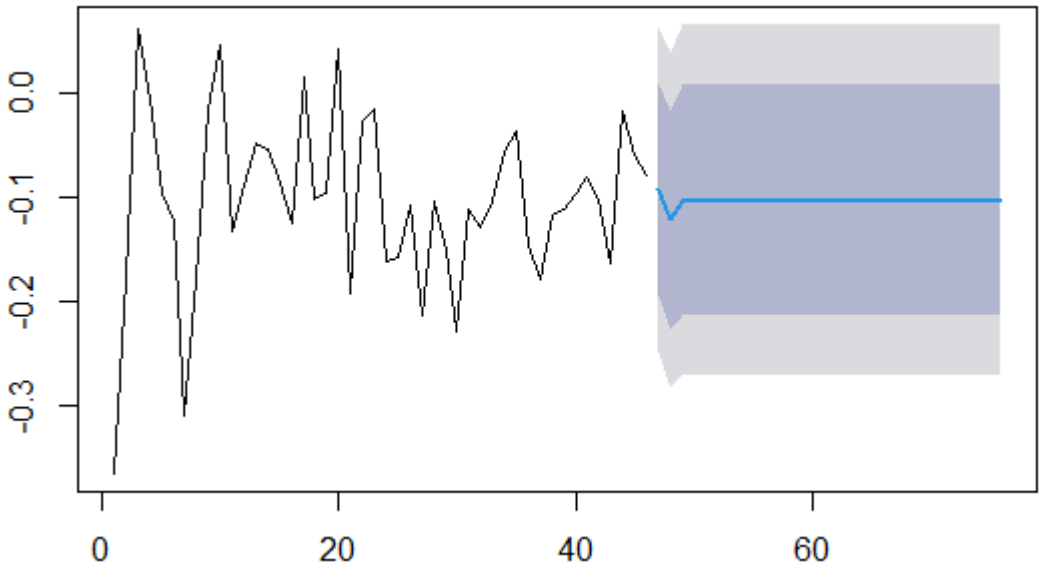


Figure 4.3.3 Sentiment level in tweets with hashtag "#oil" forecasting using ARMA model.

The expected forecast of tweets looks more stable and the conditional expected value of AR converges to the expected mean, as AR component of sentiment level forecast in the news can take a wider range of values.

Tweets in general tend to return to their mean value and behave more stable.

5. Main findings, recommendations and conclusion

The scope of the project was to do a sentiment analysis of the news from a financial site and from Social media and to find if there is a relationship between sentiment level and prices for energy sources.

News from Bloomberg site and tweets from Tweeter were collected, processed and analysed. As energy source was chosen price for oil as an important factor in energy industry.

When project started there were news about difficult situation in China where COVID-19 had an outbreak. At that moment world still hoped that it is not going to spread around the world and lead to the first in the history pandemic.

There were too many emotions going around in the period of time investigated for the project. Deserted streets, closed businesses, distancing, loss of jobs, friends and family members lost to the virus. World was boiling with emotions. It is very hard in this sea of tragedy and turmoil to find that information that was related to the project.

There were no definitive results showing that there is a direct link between sentiment level from the financial news and social media and price of oil.

We don’t know if such a situation was unique and won’t repeat itself or it is only the beginning of a series of similar events in the future. But we should study this situation from each direction.

During the research on this project a strong and defined link between sentiment analysis and financial changes was not found. But such link must exist.

The idea to find that link it is very attractive. Research in this direction should have a further development.

In order to get a better result, in case that such research will be continued, I would suggest the following:

1. Collection of data should be done for a longer period of time. 2-3 months collection of news or tweets it is not enough. This doesn’t require a longer research in time, but the access to a news archive. For Tweeter it requires access to a development account that can extract historical data.
2. Financial news should be collected or scrapped where it is possible from a few sources and not only from one.
3. Tweets should be collected not at random, but only from companies or Tweeter users that are in some way connected to the energy industry or to the finance.
4. For predicting prices dynamic it is not enough to consider only the sentiment level, but some additional factors too. May be the same futures that alone didn’t bring the expected result.

References

1. Alquist Ron, Kilian Lutz, and Vigfusson Robert, "Forecasting the Price of Oil". *Board of Governors of the Federal Reserve System International Finance Discussion Papers,* Number 1022, July 2011
2. Bradley Efron, Tibshirani Robert J. "An Introduction to Bootstrap". Chapman and Hall/CRC; 1 edition (January 1, 1993)
3. Bloomberg, "Terms of service", 2020, (https://www.bloomberg.com/notices/tos/)
4. Bloomberg green, 2020, (https://www.bloomberg.com/green?sref=TG1SYLLI)
5. Bloomberg. "Bloomberg the Company and its Products", 2020, (https://www.bloomberg.com/company/?utm_source=bloomberg-menu&utm_medium=bcom)
6. Chan Wesley, "Stock price reaction to news and no-news: Drift and reversal after headlines," *Journal of Financial Economics*, vol. 70(2), no. 2, pp. 223–260, February 2003.View at: Publisher Site | Google Scholar
7. EIA, "Oil: crude and petroleum products explained." US Energy Information Administration, 2020, (https://www.eia.gov/energyexplained/oil-and-petroleum-products/prices-and-outlook.php)
8. Fang Lily, Peress Joel. " Media coverage and the cross-section of stock returns", Willey online library, 28 Sep, 2009
9. Fontarella Clint. "The best 8 Sentiment Analysis Tools in 2020.", Originally published Jun 7, 2019 8:00:00 AM, updated April 16 2020, HubSpot, https://blog.hubspot.com/service/sentiment-analysis-tools
10. Gonzalez-Vallinas, Project "Extract Stock Sentiment from news Headlines", DataCamp, 2018, (https://learn.datacamp.com/projects/611)
11. Hansen Sarah. "Here's what Negative Oil Prices Really Mean". Forbes, 21 April, 2020, (https://www.forbes.com/sites/sarahhansen/2020/04/21/heres-what-negative-oil-prices-really-mean/#1322a3b05a85)
12. Hutto, C.J. & Gilbert, E.E. (2014). "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text." Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

13. MacFarlane Greg, “How Bloomberg makes money: Terminals, News, Business ”, Investopidia, 2020, (https://www.investopedia.com/articles/investing/102015/how-bloomberg-makes-billions-hint-not-just-news.asp)
14. Mehrotra Kishan, Moham Chilukuri, Huang HuaMing, “Amonaly detection principles and algorithms”. Springer, 2017
15. Mendelevitch Ofer, Stella Casey, Eadline Douglas, “Practical Data Science with Hadoop and Spark: Designing amd building Effective Analytics at Scale.” , Addison-Wesley Profesional, 2017
16. Montgomery Douglas, Jennings Cheryl, Kulahci Murat, “Forecasting the price of oil”, Economical Bulletin. Issue 4, 2015
17. Montgomery Douglass and others, “Introduction to Time Series Analysis and Forecasting (Wiley Series in Probability and Statistics) 2nd Edition, 2015
18. Symeonidis Symeon, “5 Things You Need to Know about Sentiment Analysis and Classification”. KDnuggets news, Mar, 2018.(https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html)
19. Schumaker, Robert P. and Maida, Nick (2018) "Analysis of Stock Price Movement Following Financial News Article Release," *Communications of the IIMA*: Vol. 16 : Iss. 1 , Article 1. Available at: https://scholarworks.lib.csusb.edu/ciima/vol16/iss1/1
20. Schumaker Robert P., “Analysis of Stock Price Movement Following Financial News”, *Communications of the IIMA*, volume 16, Issue 1, 2018
21. Surbhi Garg, Neetu Verma, “Study of Senti,ent Classification Techniques, April 2018, ResearchGate, (https://www.researchgate.net/publication/332343554_Study_of_Sentiment_Classification_Techniques)
22. Tableau, “7 great books about time series analysis”, 2020, (https://www.tableau.com/learn/articles/time-series-analysis-books)
23. Tweepy Documentation, 2020, (http://docs.tweepy.org/en/latest/)
24. Tweeter. “API reference index”, 2020, (https://developer.twitter.com/en/docs/api-reference-index)
25. Tweeter, Terms of use Policy , 2020, (https://developer.twitter.com/en/developer-terms/policy)

26. Uhr Patrick, Zenkert Johannes and Marjid, Fathi . " A Framework to utilize the Human Ability of Word Association for analysing Stock Market News Reports". Conference: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), At San Diego, CA, USA
27. Venner Json, "Pro Hadoop", Apress, 2009
28. Wikipedia, "Feature scaling", 2020, (https://en.wikipedia.org/wiki/Feature_scaling)
29. WHO, Coronavirus, 2020, (https://www.who.int/health-topics/coronavirus#tab=tab_1)
30. Worland, Justin (20 April 2020). "Oil Prices Won't Be Negative Forever. But the Oil Industry Will Never Be the Same". Time. Archived *from the original on 21 April 2020*. Retrieved 21 April 2020.
31. Yahoo Finance, "Crude Oil Jun 20 (CL=F) ", 2020, (https://finance.yahoo.com/quote/CL%3DF/history/)
32. Zinflou Arnaud, "Playing with time series data in Python", Medium, Jun 29, 2018, https://towardsdatascience.com/playing-with-time-series-data-in-python-959e2485bff8
33. Burchell Jodie, "Using VADER to handle sentiment analysis with social media text", t-redacty.io Blog, 2017, (http://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html)
34. Surbhi Garg, Verma Neetu, "Study of Sentiment Classification Techniques", 2018, researchgate.net (https://www.researchgate.net/publication/332343554_Study_of_Sentiment_Classification_Techniques)
35. Amazon, Mechanical Turk, Marketpalce, 2020, (https://www.mturk.com/)
36. Github, cjhutto/vaderSentiment, 2014, (https://github.com/cjhutto/vaderSentiment)
37. White Tom, "Hadoop. The definitive guide.", O'Really, 2012
38. Chen James, "Futures", Investopedia, 2020, (https://www.investopedia.com/terms/f/futures.asp)
39. Pandley Parul, "Simplifying Sentiment Analysis using VADER in Python (on Social Media Text), Medium, 2018, (https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f)
40. Koehersen Will, "Time Series Analysis in Python: An Introduction". Medium, 2018, (https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a)

41. Nau Robert, “Statistical Forecasting: notes on Regression and Time Series Analysis”, Chapter 5. “ARIMA models for time series forecasting”, Fuqua School of Business, Duke University, 2020, (https://people.duke.edu/~rnau/411home.htm)

42. Appendices:

Appendix 1. A sample of a tweet in JSON format:

{"created_at":"Fri Mar 20 17:55:26 +0000 2020","id":1241060796869816323,"id_str":"1241060796869816323","text":"RT @MediciSusan: #Bloomberg makes massive $18M transfer from campaign to #DNC https:\/\/t.co\/eZI0ekpQQJ","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1176458901010702336,"id_str":"1176458901010702336","name":"Eden Sarbanes","screen_name":"Eden_Sarbanes","location":"Maryland, USA","url":null,"description":"\ud83c\udf0aRESIST. Meritocracy + Community. DM\u2019s get unfollowed. Life is too short for trolls.","translator_type":"none","protected":false,"verified":false,"followers_count":2687,"friends_count":3617,"listed_count":4,"favourites_count":38084,"statuses_count":34289,"created_at":"Tue Sep 24 11:30:46 +0000 2019","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/1182091936280322049\/w6Po8sgK_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/1182091936280322049\/w6Po8sgK_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/1176458901010702336\/1571328778","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Fri Mar 20 17:51:52 +0000 2020","id":1241059900366741504,"id_str":"1241059900366741504","text":"#Bloomberg makes massive $18M transfer from campaign to #DNC https:\/\/t.co\/eZI0ekpQQJ","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":800880569487749122,"id_str":"800880569487749122","name":"Susan Medici","screen_name":"MediciSusan","location":"Planet Earth is on Fire","url":null,"description":"#NeverTrump Vote #Trump OUT","translator_type":"none","protected":false,"verified":false,"followers_count":5792,"friends_count":5817,"listed_count":29,"favourites_count":154238,"statuses_count":99091,"created_at":"Tue Nov 22 01:56:22 +0000 2016","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"prof

ile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/1231333099264528385\/gYtKd_XL_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/1231333099264528385\/gYtKd_XL_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/800880569487749122\/1584555548","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":2,"favorite_count":1,"entities":{"hashtags":[{"text":"Bloomberg","indices":[0,10]},{"text":"DNC","indices":[56,60]}],"urls":[{"url":"https:\/\/t.co\/eZI0ekpQQJ","expanded_url":"https:\/\/www.politico.com\/news\/2020\/03\/20\/mike-bloomberg-massive-18m-transfer-dnc-138771","display_url":"politico.com\/news\/2020\/03\/2\u2026","indices":[61,84]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en"},"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"Bloomberg","indices":[17,27]},{"text":"DNC","indices":[73,77]}],"urls":[{"url":"https:\/\/t.co\/eZI0ekpQQJ","expanded_url":"https:\/\/www.politico.com\/news\/2020\/03\/20\/mike-bloomberg-massive-18m-transfer-dnc-138771","display_url":"politico.com\/news\/2020\/03\/2\u2026","indices":[78,101]}],"user_mentions":[{"screen_name":"MediciSusan","name":"Susan Medici","id":800880569487749122,"id_str":"800880569487749122","indices":[3,15]}],"symbols":[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"1584726926425"}

Appendix 2. Dictionary with keywords to used to group news similar to tweets:

Renewable	Climate	Gas	Oil	Energy	Bloomberg	Coronavirus
wind	Tree	drillers	Oil	Electricity	Stocks	COVID
renewable	Emission	gas	crude	Energy	wall street	coronavirus
solar	Warm			Mining	bitcoin	lockdown
bio	Environment			Power	recession	pandemic
green	weather			fuel	crises	virus
biofuel,	climate				finance	distancing
battery	arctic				fund	health
	green				bond”	death
	meteorolo					medic
	dioxide					
	atmosphere					
	flood					
	plastic					
	Antarctic					
	Carbon					
	Pollution					
	methane					

List of Figures

Tables