|  |  |
| --- | --- |
| A view of a city at night  Description automatically generated  Text analytics: oil price | Abstract  Oil is among the commodities with the most volatile prices. Using various techniques and frameworks of text analytics in R, there are noteworthy insights collected in this report.  Delara Vafaeinejad  Text Analytics |

# Intro to the report

Text analytics regarding stocks is probably not the newest thing. However, oil stock prices could be the most relevant one if one intends to conduct analysis. The reason for that is because oil is a commodity that is highly related to qualitative factors and simply the news occurring across the globe. My purpose of doing this analysis is to help the people involved in this market, get to an understanding or possibly a prediction of oil price as soon as they listen to the newest headlines in the papers or on TV.

## The Frameworks Used

The data is collected based on more than 10 articles in a format of op-ed or news in two separate contexts (different timelines): “oil price going up” and “oil price going down”; These two, are my two text files being analyzed and compared in the analysis.

I have used several frameworks that are almost unilaterally conveying the same message that will be mentioned separately. The frameworks and the structure I have used is as follows:

* Importing the data and set them as data frames for tokenization
* Simple plotting the counts of words
* Correlogram and Correlation for the two data frames of oil up or down
* Sentiment analysis (with different variations)
* Word clouds
* Term frequency- Inverse document frequency

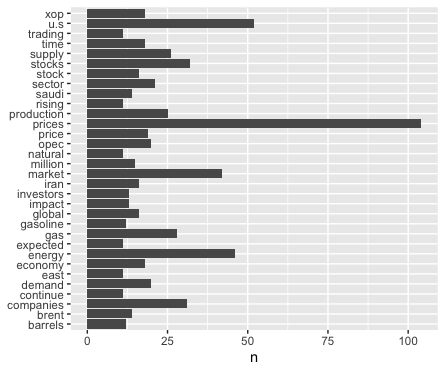
# Analysis of code results

## Tokenization

After tokenization of the data frames, a process of removing needless words, stop word, has been done. Running the plots, a few words have been added to the default stop words. And the result is as follows:

When oil price goes down

When oil price goes up

**A close up of a piece of paper

Description automatically generated**

These two simple count graphs show some certain key words that are related to the energy industry and is specific to oil and gas. Words such as xop[[1]](#footnote-1), brent[[2]](#footnote-2), bpd and so forth that are explained in the footnotes or could be simply googled. Here, the main point is the name of countries listed as the most repeated ones. Iran, Saudi, US and China which shows how these countries and their actions affect the market.

## Correlograms

A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

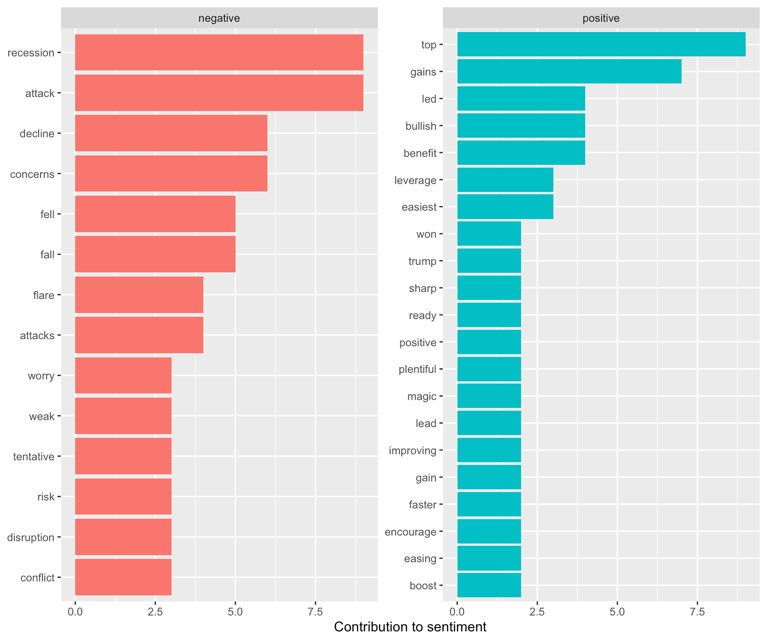
Description automatically generated

Although the results of the correlation, as shown in the second picture, was predictable, the correlogram has helpful insights in it. The result of this correlogram is truly helpful to notice that how words like attack is associated with higher prices with oil. Also, the words related to energy terminology are right on the line of the correlation; which is a proof of the result of the high correlation of the correlation formula, cor = 0.9507.

However, the certain words that catch the attention of analyst are the words attack, India, Iran and east[[3]](#footnote-3) that are associated with the upper oil prices, have some messages in it; higher oil prices will impact the economy of countries like China and India negatively and that is always their biggest concern. Also, history shows that sanctions on Iran has been always followed by higher price of oil, as well as war threat on the country.

## Sentiment Analysis

A close up of a logo

Description automatically generated

When oil price goes down:

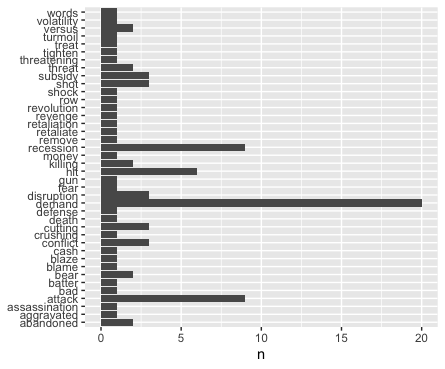
The bing sentiment lexicon that shows the attribute of negativity or positivity of a text is showing that negativity is higher in this file. Later we will see that in comparison with articles regarding higher price of oil, the case would be the same. However, the key point here is the word debt. The debt is the aftermath of oil-dependent countries after a decrease of price in oil.

When oil price goes up:

Although the comparison of negativity and positivity is the same as the lower oil prices, we want to point out the most contributed negative word that is recession here and in the above graph was debt. The word recession is referring to the result of high oil prices. When the price of oil goes up or the signs show this change, recession will signal a presence. In some cases, there has been a recession followed by high oil prices.

NRC sentiment analysis: Anger

The following graphs are showing the two data frames while and nrc sentiment analysis has been conducted. The anger sentiment has been chosen to prove that how oil prices will go up when this sentiment is seen more often in texts. It can be easily distinguished that the graph on the left side is nrc (anger) sentiment analysis for articles regarding spiking of oil prices.

A screenshot of a cell phone

Description automatically generated

The following sentiment analysis are also conducted to be analyzed separately for the two data sets.

When oil price goes down

When oil price goes up

## 

These graphs are clarifying what the negativity and positivity graph (mentioned above) could not explicitly say. Bing is so negative for lower price of oil and in a binary method, this lexicon assumes that most of the text is using negative comparing to the other text. However, this is only about one lexicon and for further analysis we have other methods to prove which text is more negative. These two visuals prove that all the three lexicons have lots of negativity that is related to volatility of prices in times of crisis across the world.

## Word clouds

The word clouds are only another form of approval of what has been stated so far. However, at one glance, word clouds might convey the messages on a very high level.

A screenshot of text

Description automatically generated

The general word cloud for both data sets

The following word clouds are so descriptive as they show that low price text is tending to be more about surprise and anger. While the high price text is more about disgust and surprise. Now, even joy is added to the higher price context and the reason for that is because in the vacation season, demand and as a result, oil prices will go high; this context was so out of mind.

When oil price goes down-nrc

A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generated

When oil price goes up- nrc

## Term Frequency -Inverse Document Frequency

Here is one of the most noteworthy points of the report. Tf-idf will show how significant some words are while they have not been necessarily the most frequent ones. The importance of a word is interesting as shown below:

# A screenshot of a cell phone Description automatically generated A screen shot of a computer Description automatically generated

Coronavirus topped the result of this research only because some of the articles were probably regarding the current news. This analysis is smart in the way that this word did not show up until now as a frequent word; however, now it is showing its importance which makes sense due to the effect it has. This virus lowers the price of oil as well as immunity of human body! We do not want to be distracted by this very unknown disaster and therefore we should study the other words. The word marketing seems to be like a buzzword, but it is not. Regarding the higher price of oil, there are some marketing efforts that energy companies should make so that price of oil does not decrease more than a certain level and they are always in a controversy with governments.

Behind all the words arranged in order of highest tf-idf there is a story that should be researched about for a better domain knowledge.

## Bigrams

The point of doing this code is to clear any mistakes in the analysis that is based on only one single word. For instance, as mentioned above, the word east was so frequent that shows how big of a player if this region in effects on oil prices. However, we did not know that until realizing the following result:

A screenshot of a cell phone

Description automatically generated

# Conclusion

The insights that text analytics could bring is absolutely valuable. Although further explanation of each content was out of the scope of this report, each graph or each result could convey solid messages if the code and data is set up properly.

So far, what is clear from the result of codes is that the volatility of the world is affecting oil prices in a vivid way and all players are mentioned above. The more negativity, the higher the price based on parts of the sentiment analysis, however nrc analysis proved that low price of oil is not necessarily the best thing for the world. From now, we would know that as soon as there is a military threat or sanction or tensions there is a higher oil price. We would consider many of these points if we intend to enter the energy stock market as well as working with related industries.

# The R-code

library(textreadr)

library(dplyr)

oil\_up<-read\_document("oil\_up\_t.txt")

oil\_up <- data\_frame(line = 1:length(oil\_up), text = oil\_up)

View(oil\_up)

oil\_down<-read\_document("oil\_down\_t.txt")

oil\_down <- data\_frame(line = 1:length(oil\_down), text = oil\_down)

View(oil\_down)

library(tidyverse)

library(tidytext)

library(ggplot2)

#tokenization

cust\_stop<-data\_frame(word=c('barrel','wsj.com','ydstie', 'wrote','oil','crude','year\'s','world','denbury'),

lexicon=c('cust','cust','cust','cust','cust','cust','cust','cust','cust'))

oil\_up\_token<-oil\_up%>%

unnest\_tokens(word, text)%>%

anti\_join(stop\_words) %>%

anti\_join(cust\_stop)

#count(word, sort=TRUE)

#top\_n(30)%>%

# ggplot(aes(word, n))+

# geom\_col()+

# xlab(NULL)+

# coord\_flip()

oil\_up\_token

oil\_down\_token<-oil\_down%>%

unnest\_tokens(word,text)%>%

anti\_join(stop\_words) %>%

anti\_join(cust\_stop)

#count(word,sort = TRUE)

# top\_n(30) %>%

# ggplot(aes(word, n))+

# geom\_col()+

# xlab(NULL)+

# coord\_flip()

oil\_down\_token

# heading to correlagrams and correlations

library(tidyr)

frequency\_up\_down <- bind\_rows(mutate(oil\_up\_token, author="Ups"),

mutate(oil\_down\_token, author= "Downs") )%>%

mutate(word=str\_extract(word, "[a-z']+")) %>%

count(author, word) %>%

group\_by(author) %>%

mutate(proportion = n/sum(n))%>%

select(-n) %>%

spread(author, proportion) %>%

gather(author, proportion, `Downs`)

# plot the correlograms:

library(scales)

ggplot(frequency\_up\_down, aes(x=proportion, y=`Ups`,

color = abs(`Ups`- proportion)))+

geom\_abline(color="grey40", lty=2)+

geom\_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+

geom\_text(aes(label=word), check\_overlap = TRUE, vjust=1.5) +

scale\_x\_log10(labels = percent\_format())+

scale\_y\_log10(labels= percent\_format())+

scale\_color\_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+

facet\_wrap(~author, ncol=2)+

theme(legend.position = "none")+

labs(y= "Ups", x=NULL)

##########################################

##doing the cor.test() ################

##########################################

cor.test(data=frequency\_up\_down[frequency\_up\_down$author == "Downs",],

~proportion + `Ups`)

######getting the lexicons

library(textdata)

afinn<-get\_sentiments("afinn")

nrc<-get\_sentiments("nrc")

bing<-get\_sentiments("bing")

loughran<-get\_sentiments("loughran")

nrc\_negative <- get\_sentiments("nrc") %>%

filter(sentiment == "negative") #what is your sentiment

nrc\_negative

#inner joining both approaches with negative sentiments

oil\_down\_token %>%

inner\_join(nrc\_negative) %>%

count(word, sort=T)

oil\_up\_token%>%

inner\_join(nrc\_negative)%>%

count(word, sort=TRUE)

#### we can see the articles regarding higher prices of oil has more negative sentiments

nrc\_anger <- get\_sentiments("nrc") %>%

filter(sentiment == "anger") #what is your sentiment

#inner joining with anger sentiments

oil\_down\_token %>%

inner\_join(nrc\_anger) %>%

count(word, sort=T)

# ggplot(aes(word, n))+

# geom\_col()+

# xlab(NULL)+

# coord\_flip()

oil\_up\_token%>%

inner\_join(nrc\_anger)%>%

count(word, sort=TRUE)

# ggplot(aes(word, n))+

# geom\_col()+

# xlab(NULL)+

# coord\_flip()

#####again this is another proof showing higher conflicts take oil prices higher

###############sentiment for oil\_down articles

afinn\_d <- oil\_down\_token %>%

inner\_join(get\_sentiments("afinn"))%>%

group\_by(index=line) %>%

summarise(sentiment=sum(value)) %>%

mutate(method="afinn")

afinn\_d

bing\_and\_nrc <- bind\_rows(

oil\_down\_token%>%

inner\_join(get\_sentiments("bing"))%>%

mutate(method = "Bing et al."),

oil\_down\_token %>%

inner\_join(get\_sentiments("nrc") %>%

filter(sentiment %in% c("positive", "negative"))) %>%

mutate(method = "NRC")) %>%

count(method, index=line ,sentiment) %>%

spread(sentiment, n, fill=0) %>%

mutate(sentiment = positive-negative)

bind\_rows(afinn\_d, bing\_and\_nrc) %>%

ggplot(aes(index, sentiment, fill=method))+

geom\_col(show.legend=FALSE)+

facet\_wrap(~method, ncol =1, scales= "free\_y")

###### now sentiments for oil\_up articles

afinn\_u <- oil\_up\_token %>%

inner\_join(get\_sentiments("afinn"))%>%

group\_by(index=line) %>%

summarise(sentiment=sum(value)) %>%

mutate(method="afinn")

bing\_and\_nrc <- bind\_rows(

oil\_up\_token%>%

inner\_join(get\_sentiments("bing"))%>%

mutate(method = "Bing et al."),

oil\_up\_token %>%

inner\_join(get\_sentiments("nrc") %>%

filter(sentiment %in% c("positive", "negative"))) %>%

mutate(method = "NRC")) %>%

count(method, index=line ,sentiment) %>%

spread(sentiment, n, fill=0) %>%

mutate(sentiment = positive-negative)

bind\_rows(afinn\_u, bing\_and\_nrc) %>%

ggplot(aes(index, sentiment, fill=method))+

geom\_col(show.legend=FALSE)+

facet\_wrap(~method, ncol =1, scales= "free\_y")

##############################################################

######## Most common positive and negative words #############

bing\_counts\_down <- oil\_down\_token %>%

inner\_join(get\_sentiments("bing")) %>%

count(word, sentiment, sort=T) %>%

ungroup()

bing\_counts\_down %>%

group\_by(sentiment) %>%

top\_n(10) %>%

ungroup() %>%

mutate(word=reorder(word, n)) %>%

ggplot(aes(word, n, fill=sentiment)) +

geom\_col(show.legend = FALSE) +

facet\_wrap(~sentiment, scales = "free\_y")+

labs(y="Contribution to sentiment", x=NULL)+

coord\_flip()

###############

bing\_counts\_up <- oil\_up\_token %>%

inner\_join(get\_sentiments("bing")) %>%

count(word, sentiment, sort=T) %>%

ungroup()

bing\_counts\_up %>%

group\_by(sentiment) %>%

top\_n(10) %>%

ungroup() %>%

mutate(word=reorder(word, n)) %>%

ggplot(aes(word, n, fill=sentiment)) +

geom\_col(show.legend = FALSE) +

facet\_wrap(~sentiment, scales = "free\_y")+

labs(y="Contribution to sentiment", x=NULL)+

coord\_flip()

###################### Worldcloud

library(RColorBrewer)

library(wordcloud)

oil\_down\_token %>%

with(wordcloud(word, n,max.words = 50))

###oil\_up world cloud

oil\_up\_token %>%

with(wordcloud(word, n,max.words = 50))

###################################################

#################oil down clouds#########################

library(reshape2)

oil\_down\_token %>%

inner\_join(get\_sentiments("nrc")) %>%

count(word, sentiment, sort=TRUE) %>%

acast(word ~sentiment, value.var="n", fill=0) %>%

comparison.cloud(colors = c("grey20", "gray80"),

max.words=100,

scale=c(0.5,0.5), fixed.asp=TRUE,title.size = 1)

library(reshape2)

oil\_down\_token %>%

inner\_join(get\_sentiments("bing")) %>%

count(word, sentiment, sort=TRUE) %>%

acast(word ~sentiment, value.var="n", fill=0) %>%

comparison.cloud(colors = c("grey20", "gray80"),

max.words=50,

scale=c(0.5,0.5), fixed.asp=TRUE,title.size = 1)

###################################################

#################oil up clouds#########################

library(reshape2)

oil\_up\_token %>%

inner\_join(get\_sentiments("nrc")) %>%

count(word, sentiment, sort=TRUE) %>%

acast(word ~sentiment, value.var="n", fill=0) %>%

comparison.cloud(colors = c("grey20", "gray80"),

max.words=50,

scale=c(0.5,0.5), fixed.asp=TRUE,title.size = 1)

library(reshape2)

oil\_up\_token %>%

inner\_join(get\_sentiments("bing")) %>%

count(word, sentiment, sort=TRUE) %>%

acast(word ~sentiment, value.var="n", fill=0) %>%

comparison.cloud(colors = c("grey20", "gray80"),

max.words=100,

scale=c(0.5,0.5), fixed.asp=TRUE,title.size = 1)

###########just a graph to consolidate both articles

combined\_oil <- bind\_rows(mutate(oil\_down\_token, make="down"),

mutate(oil\_up\_token, make= "up"))

oil\_modif <- combined\_oil %>%

count(make, word, sort=TRUE) %>%

ungroup()

oil\_modif2 <- oil\_modif %>%

group\_by(make) %>%

summarize(total=sum(n))

oil\_leftjoined <- left\_join(oil\_modif, oil\_modif2)

oil\_tfidf <- oil\_leftjoined %>%

bind\_tf\_idf(word, make, n)

oil\_tfidf # we get all the zeors because we are looking at stop words ... too common

oil\_tfidf %>%

arrange(desc(tf\_idf))

####

library(ggplot2)

ggplot(oil\_leftjoined, aes(n/total, fill = make))+

geom\_histogram(show.legend=FALSE)+

xlim(NA, 0.001) +

facet\_wrap(~make, ncol=2, scales="free\_y")

#################!!!tf idf######

all\_oil<-bind\_rows(mutate(oil\_down\_token,loc\_oil='down'),

mutate(oil\_up\_token,loc\_oil='up'),)

oil\_new <- all\_oil %>%

count(loc\_oil, word, sort=TRUE) %>%

ungroup()%>%

bind\_tf\_idf(word,loc\_oil,n)

oil\_new%>%

arrange(desc(tf\_idf))

oil\_new%>%

arrange(desc(tf\_idf)) %>%

mutate(word=factor(word, levels=rev(unique(word)))) %>%

group\_by(loc\_oil) %>%

top\_n(10) %>%

ungroup %>%

ggplot(aes(word, tf\_idf, fill=loc\_oil))+

geom\_col(show.legend=FALSE)+

labs(x=NULL, y="tf-idf")+

facet\_wrap(~loc\_oil, ncol=2, scales="free")+

coord\_flip()

#########################bigrams

library(dplyr)

library(tidytext)

library(tidyr)

oil\_bigrams <- all\_oil %>%

unnest\_tokens(bigram, word, token = "ngrams", n=2)

oil\_bigrams %>%

count(bigram, sort = TRUE)

# Reference[[4]](#footnote-4):

<https://www.cannontrading.com/tools/support-resistance-levels/crude-oil-guide-brent-vs-wti-whats-the-difference/>

<https://www.fool.com/investing/2020/01/08/why-oil-stocks-are-tanking-today.aspx>

<https://www.investopedia.com/5-reasons-energy-stocks-could-surge-4772280>

<https://www.cnbc.com/2019/04/03/why-oil-and-gasoline-prices-are-rising-faster-than-analysts-expected.html>

<https://economictimes.indiatimes.com/wealth/invest/crude-oil-on-the-boil-again-what-it-means-for-the-stock-market-and-economy/articleshow/73095214.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst>

<https://www.barrons.com/articles/oil-stocks-iran-attack-saudi-crude-iraq-51578068294>

<https://www.cnbc.com/2019/09/16/history-says-oil-stocks-keep-rising-this-is-easiest-way-to-bet-on-it.html>

<https://www.wsj.com/articles/disney-impresses-oil-falls-and-stocks-return-to-record-territory-11581080400?mod=searchresults&page=1&pos=9>

<https://www.wsj.com/articles/shale-gas-swamps-asia-pushing-lng-prices-to-record-lows-11580994013?mod=searchresults&page=1&pos=17>

<https://finance.yahoo.com/news/why-oil-stocks-getting-crushed-202900316.html>

<https://www.forbes.com/sites/gauravsharma/2020/02/07/perfect-market-storm-could-see-oil-prices-drop-below-30/#27c8b936f816>

<https://thehill.com/opinion/energy-environment/447334-why-oil-prices-are-dropping-this-summer>

1. SPDR S&P Oil & Gas Exploration & Production ETF (XOP) [↑](#footnote-ref-1)
2. **Brent** Blend is a combination of crude oil from 15 **different** oil fields **in the** North Sea. It is less “light” and “sweet” than **WTI**, but still excellent for making gasoline. ... **WTI** is also known as Light Sweet Crude, the majority of which is located **in the** Permian oil Field. [↑](#footnote-ref-2)
3. In bigram results, we would see that east means the Middle East [↑](#footnote-ref-3)
4. This reference list includes the articles that forms my two data sets [↑](#footnote-ref-4)