Visualising Spikes in Time Series and Sentiment Analysis

Stephen Lowry
B.A.(Mod.) Computer Science
Final Year Project April 2012
Supervisor: Prof. Khurshid Ahmad

School of Computer Science and Statistics
O’Reilly Institute, Trinity College, Dublin 2, Ireland
DECLARATION

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

__________________________________________  ______________________
Name                                              Date
PERMISSION TO LEND

I agree that the Library and other agents of the College may lend or copy this thesis upon request.

______________________________________________  __________________________
Name Date
I would like to thank Professor Khurshid Ahmad for the expertise and advice he provided.
Abstract

Crude oil is an actively traded commodity. It can be bought and sold in a number of ways. One of which is through future contracts. This project focuses on crude oil future contract prices. The price of oil is based on a number of factors. Supply and demand is one of the main contributors but traders can also use market sentiment as an indicator of when to buy and sell contracts. This project investigates the correlation between sentiment contained within news articles and the price of oil.
# Table of Contents

Acknowledgement ................................................................................................................... 4

Abstract ................................................................................................................................... 5

1. Introduction ............................................................................................................................. 8
   1.1 Oil ......................................................................................................................................... 9
   1.2 Oil Prices .............................................................................................................................. 11
      1.2.1 Future Contracts ........................................................................................................... 11
      1.2.2 Oil Price History ......................................................................................................... 13
   1.3 Sentiment Analysis ............................................................................................................ 14

2 Literature Review ..................................................................................................................... 16
   2.1 Oil Prices ............................................................................................................................ 16
   2.2 Efficient Market .................................................................................................................. 17
   2.3 Sentiment Analysis ............................................................................................................ 17
   2.4 Motivation .......................................................................................................................... 18

3 Method ...................................................................................................................................... 20
   3.1 Technologies Used .............................................................................................................. 20
      3.1.1 Retrieving an Article ................................................................................................... 20
      3.1.2 Extracting the words .................................................................................................. 21
      3.1.3 Graphing ..................................................................................................................... 22
   3.2 Sentiment in News .............................................................................................................. 22
   3.3 Building a dataset .............................................................................................................. 22
   3.4 Testing the datasets .......................................................................................................... 25
   3.5 Data Analysis ..................................................................................................................... 26
      3.5.1 Compression ................................................................................................................. 27
      3.5.2 Graphing ...................................................................................................................... 27

4 Experiments & Evaluation ....................................................................................................... 29
5 Afterword........................................................................................................................................35
Bibliography ......................................................................................................................................38
List of Tables and Figures..................................................................................................................40
1. Introduction

The aim of this project is to investigate the correlation between sentiment in news articles and oil prices. This objective is completed by:

- Identifying which words relate to positive sentiment and which words relate to negative sentiment with the context of oil and oil prices.
- Retrieving news articles using data retrieval techniques.
- Building a positive word dataset and a negative word dataset using text processing.
- Analysing the datasets using sentiment analysis techniques.

Chapter 1 is named Introduction. This chapter of the report will introduce the reader to the project and the domain of oil and oil prices. As quite a bit of context is required to understand the project, the introduction section is split into a few different parts. A background of oil and information about oil consumption and production is given. This shows the reader how much oil is used worldwide, how important it is and how large the companies dealing with oil are. Next oil prices are discussed including crude oil future contracts. The crude oil future market is explained along with the type of people who actively trade oil contracts. Key financial terms are explained to help the reader understand the context of the project. To give an overall understanding of the context behind the project, a brief history of oil prices is discussed and visualised. An introduction to sentiment analysis is also given. This includes definitions and explanations to help the reader understand the project.

Chapter 2 is named Literature Review. In this chapter, information from this area of study is conveyed. The idea is to give the reader an understanding of what other people are doing in the domain of oil and the field of sentiment analysis research. This chapter also discusses the motivation behind the project.

Chapter 3 is named Method. This chapter aims to explain the idea and approach taken to complete this project and how objectives were completed. It describes the technologies used and their applications.

Chapter 4 is named Experiments and Evaluation. In this chapter, the reader is shown information from experiments carried out. These experiments are a result of applying the method described in chapter 3. The experiments are then evaluated and a conclusion is drawn from them.
Chapter 5 is named Afterword. This chapter concludes the report and gives information about further work. Ideas for further work are presented and discussed.

1.1 Oil

Crude oil or petroleum is a natural resource found underground. It is a non-renewable energy source similar to coal and natural gas.

Crude oil is extracted from the ground and refined to create a product that has many uses. It can be used for fuels including petrol jet fuel, kerosene, fuel oil and diesel fuel. Other uses include lubricants, waxes and pesticides.

It is a widely traded commodity. Consumption rates have risen steadily in the last 50 years.

Table 1 Top Five Countries by Oil Consumption

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>BBL/Day</th>
<th>Date of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>19,150,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>2</td>
<td>European Union</td>
<td>13,680,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>3</td>
<td>China</td>
<td>9,400,000</td>
<td>2011 est.</td>
</tr>
<tr>
<td>4</td>
<td>Japan</td>
<td>4,452,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>5</td>
<td>India</td>
<td>3,182,000</td>
<td>2010 est.</td>
</tr>
</tbody>
</table>

The United States is the largest consumer of oil. It consumes more than the all the countries of the European Union put together and twice the amount consumed by China, the third largest consumer of oil.

Table 2 Top Five Countries by Oil Production

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>BBL/Day</th>
<th>Date of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Saudi Arabia</td>
<td>10,520,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>2</td>
<td>Russia</td>
<td>10,270,000</td>
<td>2010</td>
</tr>
<tr>
<td>3</td>
<td>United States</td>
<td>9,688,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>4</td>
<td>Iran</td>
<td>4,252,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>5</td>
<td>China</td>
<td>4,073,000</td>
<td>2011</td>
</tr>
</tbody>
</table>

Even though the United States is the largest consumer of oil, it is only the third largest producer. Japan is within the top five countries based on oil consumption rates but is not within the top twenty countries based on oil production. [1]
Table 3 Top Five Countries by Oil Exports

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>BBL/Day</th>
<th>Date of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Saudi Arabia</td>
<td>7,635,000</td>
<td>2009 est.</td>
</tr>
<tr>
<td>2</td>
<td>Russia</td>
<td>5,010,000</td>
<td>2010 est.</td>
</tr>
<tr>
<td>3</td>
<td>Iran</td>
<td>2,523,000</td>
<td>2009 est.</td>
</tr>
<tr>
<td>4</td>
<td>United Arab Emirates</td>
<td>2,395,000</td>
<td>2009 est.</td>
</tr>
<tr>
<td>5</td>
<td>European Union</td>
<td>2,196,000</td>
<td>2009 est.</td>
</tr>
</tbody>
</table>

Saudi Arabia is the largest producer and the largest exporter of oil. It is followed by Russia and Iran. The United States is not within the top ten countries based on oil exports. Although the countries in the European Union are outside the top ten producers of oil, they are within the top five exporters.

Some countries, such as Ireland, do not produce any oil. Ireland consumes 159,700 bbl/day.

As can be seen, oil is traded in large quantities on a daily basis. For countries that do not produce any or enough oil, there is a dependence on imports. Some countries, such as Saudi Arabia, are able to export large amounts of the oil they produce because they are producing far more than they are consuming.

The oil industry contains some of the largest and most important companies in the world. These companies are highly profitable and have a large effect on the economy and the environment.

BP is an oil company that extracts and produces oil. It is one of the largest companies in the world, employing nearly 80,000 people around the world. It is vertically integrated and is involved in the exploration, production, refining, distributing and marketing of oil. It has a refining throughput of around 2,426 thousand barrels per day.

BP supplies its products to retail and industry in around 100 countries. [2]

With this information it is obvious to see that oil is an important entity to businesses and countries.
1.2 Oil Prices

The price of oil is a result of a number of factors. Oil can be bought and sold in a number of different ways. This paper focuses on crude oil future contracts being traded on a financial market. In this context, supply and demand has an effect on oil price to a certain extent but it is not the sole reason for price changes. Another primary factor is market sentiment.

Supply and demand is an economic model for price determination. If supply increases, or demand goes down, the price of oil goes down. If supply decreases, or demand goes up, the price of oil goes up.

Market sentiment is the overall feeling or attitude of the market. If the sentiment, or outlook, is positive then price will go up. If the sentiment, or outlook, is negative then the price will go down. This sentiment is from a trader’s point of view.

1.2.1 Future Contracts

Oil can be bought and sold in a few different ways. One way of buying and selling oil is through future contracts. A future contract is an agreement between two parties to exchange an entity for a fixed price, at a fixed delivery location and on a fixed date in a future. Future contracts can range from one month in the future and onwards. This is in contrast with spot contracts which, basically, are deals made on the spot with a delivery being made in the following days to the contract agreement.

Future contracts are traded on a financial market which is called a futures market. Crude oil contracts are traded on a futures market. On this market, contracts are traded in a similar way to how stock is traded on the stock market. Contracts are continuously bought and sold. There is a set amount of barrels of crude oil that can be traded on each contract. Other entities traded in a similar way include wheat, gold and currency indexes.

In a futures market, there are two main players: hedgers and speculators. These can both buy and sell contracts but are differentiated based on their aims.

Hedgers use the market as a way to reduce risk. An example of a hedger could be an airline company. If their oil supply runs out in September next year, they might want to get a new supply of oil in June. They have a few options.
They can wait until June and order oil and get it delivered straight away. In this situation they are subject to whatever the price of oil is in June. If oil price is rising, they are at a risk of paying a high price.

If they think the price of oil will rise between now and June, they have the option of buying a future contract. This will allow them to pay an agreed price now for a fixed delivery location and a fixed date in the future (June next year in this case). The result of this contract is decided by the direction that oil price goes. If oil price rises, the airline company ends up paying less than the price in June which is a good result. If the oil price falls then the airline has actually paid more than the value of oil in June. This result isn’t ideal but the company still knew the price it agreed to pay so it was not at the whim of changing oil prices.

The company has avoided risk of unknown changes in oil price.

The other player in the market is known as speculators.

Speculators use the market for profit. They buy and sell contracts with the aim of making a positive return on their investment. They are essentially betting on changes in the price of oil. If the price of oil goes up after a contract is bought, it can be sold and a profit is made. If the trader thinks the price will continue to rise, they can hold onto the contract and try to make a larger profit. If the price of oil goes down after the contract is bought the trader has made a loss. The trader can either hold onto the contract in hope that the price of oil will rise and eventually turn into a profit or they can sell to limit the loss made from the trade.

It is important to note that speculators do not necessarily want a delivery of crude oil. Their main aim is to trade contracts hoping to make a profit. Traders use the crude oil market because it is volatile and because it is another market to diversify their portfolio. The same traders might trade in multiple futures, options and stock.

For traders to make a positive return on their investment they need to know the right times to buy and sell contracts. To do this they can try to analyse and attempt to forecast future price movement. A trader may look at external events that affect price movement or try to find patterns in the market.
1.2.2 Oil Price History

Figure 1 gives the reader an idea how oil price has changed in the years between 1993 and 2012. Figure 2 shows the log return rates for the time period between 2002 and 2012. A rise can be seen between 1993 and 2004 with some peaks and declines within that time frame. After 2004 the price oil rises from under 100 dollars to over 120 dollars in 2006. A sharp rise and fall can be seen in the years 2007 until 2009 with the price seeming to begin rising again towards 2012.

Factors influencing oil price is a topic of much debate. Surveys have been done that suggest factors such as stagnant supply, unexpected economic growth from developing countries such as China and India, low interest rates and a weak U.S. dollar all were associated with and potentially contributed to the oil price rise and fall in 2007/2008. [3]
1.3 Sentiment Analysis

Traditionally in financial markets, excessive rises in time series are cancelled by traders not buying and excessive falls in time series are cancelled by traders not selling. This is a form of efficient market hypothesis, which is discussed further in chapter 2. It essentially means that prices control future prices. This is not always the case as can be seen by viewing the crude oil future price history.

Noise in the market can impact on prices. This results in prices not controlling prices because there is another influence. One form of noise is sentiment.

Sentiment of the market is the feeling or attitude of the market. Sentiment analysis is an application of natural language processing. It is an active area of research that attempts to retrieve sentiment from text. More information on the work being done in sentiment analysis is discussed in chapter 2.
The aim for this project is to detect whether a day’s news is positive or negative based on the news articles published on that day. A day is classified as positive, negative or neutral. The method to do this is discussed in chapter 3.

Sentiment analysis can also be referred to as opinion mining. It has been used to analyse movie reviews, product reviews and comments on news or blog articles. This allows companies to see how well their product is being received. Since it is automated, thousands of texts can be analysed quickly. This has obvious advantages over getting a team of people to read through the texts.

This paper looks at whether or not sentiment in news articles has any correlation to the rise and fall of crude oil prices. The assumption is based on the fact that traders use the market sentiment as one of their tools for deciding whether to buy or not.
2 Literature Review

2.1 Oil Prices

Oil is one of the world’s most important commodities and oil prices have a huge effect on both local economies and the world economy.

The IMF (International Monetary Fund) Research Department studied the effect that higher oil prices have on the global economy. [4] The study concludes that a shock in oil prices would affect the global economy through supply and demand effects and also have an effect on inflation in terms of higher wage claims. This would lead to central banks increasing interest rates to offset inflationary pressures.

These results are backed up by other research into oil and the macroeconomy. Higher oil price makes production more expensive and means that firms purchase less oil. The result of this is that output is reduced. [5]

Japan is one of the largest consumers of oil. It accounts for 7% of the world’s total annual oil consumption. This is important considering its low oil production rates. Oil market conditions have the possibility of having a knock-on effect on the country’s economy. [5]

Factors that affect crude oil price changes have been examined and researched in depth. James D. Hamilton notes that changes in oil price have historically been permanent, difficult to predict and governed by different regimes at different points in time. [6] His work looks at resource depletion and the role it plays. It concludes that resource scarcity was negligible in 1997 but is becoming more and more important due to newly industrialised countries.

Oil prices also affect the stock prices of alternative energy companies. It might seem obvious that rising oil prices should be good for alternative energy companies but a simulation by Henriques and Sadorsky showed that a shock to technology stock prices has a larger impact on alternative energy stock prices than does a shock to oil prices. [7] Regardless of whether technology stock shocks or oil price shocks produce a greater change, it is evident that oil price has a knock on effect in multiple domains and areas.
2.2 Efficient Market

The Efficient Market Hypothesis deals with the theory of whether movements in markets can be predicted or not. This is a debated topic with experts on both sides having different views. An efficient market is one where the prices fully reflect all information available that affects the price. This means that no trader has access to information that another trader does not. As a result, it means that the market cannot be beaten or that prices cannot be predicted.

Predicting prices of stock or predicting movement of prices is a widely studied and researched topic. Alfred Cowles wrote a paper on his research that analysed a number of agencies that have attempted to select common stocks that provide a positive return in investment. Records from forty five professional agencies were analysed for a four and a half year period. These agencies included insurance companies, financial companies and financial publications. It was noted that “statistical tests of the best individual records failed to demonstrate that exhibited skill, and indicated that they more probably were results of chance.” but also “There is some evidence, on the other hand, to indicate that the least successful records are worse than what could reasonably be attributed to chance.” [8]

Whether stock or future prices can be predicted or not is outside the scope of this project. The aim is to provide information about the correlation between sentiment in news and oil prices.

2.3 Sentiment Analysis

While oil price fluctuates based on supply and demand, it may also fluctuate on sentiment in the market. In this paper, sentiment in news article is looked at and correlated with changing oil prices.

News can be good or bad. Sentiment analysis is used to find out which it is. Since computers cannot understand the meaning behind a word this project attempts an approach using statistical data to decide whether an article is good or bad by comparing it to previous articles which are already know to be good or bad.

In the area of sentiment analysis research, previous work includes work done by Namrata Godbole, Manjunath Srinivasiah and Steven Skiena. [9] In their research they develop a sentiment analysis system. Not only do they analyse the word in question for polarity but they also assign the same polarity to any synonyms of the word and the opposite polarity to any
antonyms. A path from the word to its antonym is discovered and the algorithm runs twice. On the first pass it calculates a preliminary score for each word. On the second pass it calculates the number of polarity changes. The fewer changes, the more trustworthy the path is. A threshold is set and only paths that fall within its range are counted. They analyse the sentiments around popular and controversial public figures such as political figures and criminal figures. Their findings show that the same person can be viewed different in blogs than in newspapers to which they conclude as bias in either the blogs or the newspapers towards that person.

The World Wide Web is made up of content comprising of text, images and video. As it has a large volume of text, it is the ideal resource for carrying out sentiment analysis. Unfortunately, it also contains a lot of noise. This makes the task of automatic recognition of sentiment more difficult. Automatic Sentiment Analysis in On-line Text is a paper by Boiy, Hens, Deshacht and Moens in which they give an overview of various techniques used to approach the problems in sentiment analysis. [10]

In their research they discuss using a web search (AltaVista) to find word tuples. This technique looks at the number of documents returned from a search and from that it calculates the orientation (positive or negative) of a word.

It discusses retrieving more than a single word from a sentence. This is referred to as N-grams. A word N-gram is a subsequence of N words from a given sequence. This allows the analysis to have a greater context for each word.

A conclusion is given that states that in some circumstances improvements over state-of-the-art methods for sentiment recognition in texts are possible.

2.4 Motivation

Previous research into oil and the effects of changing oil price show that it is a very important commodity. Changing oil price has the potential to have a knock on effect in multiple different domains including effects on the global economy.

It is important that movement in oil price is studied and observed.

The area of sentiment analysis research is active. Research is active in multiple techniques to enhance how effective sentiment analysis is. This project uses a lexicon based approach towards sentiment analysis.
This project makes an attempt to use sentiment analysis to investigate the correlation between sentiment contained within news articles and changes in oil price.
3 Method

This section of the report describes the method of building a system to analyse news articles. The polarity of a day is the result of the analysis on the news articles that were published on that day. The technologies used are explained in detail as well as their application.

All scripts, programs and code written to carry out the task will be referred to as a system in this report.

The approach to making this system is split into two parts. The first part is a building stage. In this stage two datasets are created or trained using data retrieval, natural language processing and technical analysis techniques. The second stage is the testing stage. In this part the previously trained datasets are used to decide the polarity for a specific date. This uses a statistical approach.

In the context of this project, the term dataset refers to a collection of words.

3.1 Technologies Used

The system was written in Python 2.7, an open source programming language. [11] Python 2.7 was chosen because it offers a range of extensible libraries which were suited for the task undertaken by this project. The development platform was Ubuntu 11.10.

3.1.1 Retrieving an Article

The first step in building the datasets is to find the articles for a specific date and retrieve all the words from each article. To do this the system uses a Python module called urllib2 for fetching URLs (Uniform Resource Locators). [12] A URL is a reference to an Internet resource. It is more commonly known as a website address. A modified URL containing a user’s input (single date or a date range) is fetched from the BBC World News website. This performs the equivalent action of a person going to the BBC World News website and searching for a date through the web interface. (Figure 3) If the user entered a single date, the system would only retrieve the news articles for that date. If a date range was input then the
The system would step through each date and retrieve and analyse each date until it reached the end date given.

The result of fetching a URL is the HTML source code of a web page that lists the articles published on that date. The system then uses a Python library called Beautiful Soup. [13] This allows the system to analyse the contents of a webpage in its HTML source form. From this, the titles of the articles are identified along with the URLs to each article. urllib2 is used again to retrieve each article’s HTML source code. Once again, Beautiful Soup is used. This time it is used to extract the words in the article while avoiding other text such as HTML code, navigation text and other content on the page.

![Figure 3 Search result for oil from the BBC News website](Image)

### 3.1.2 Extracting the words

The system has the words for an article but not every word is important and some extra content might have been retrieved by mistake. The next step the system performs is to filter out words that should not enter the dataset. This process uses an open source Python module called Natural Language Toolkit (NLTK). This provides linguistic data and documentation for research and development in natural language processing and text analytics. [14]
3.1.3 Graphing

All visualisation and graphing done for this report and for the project used a Python 2D plotting library, matplotlib. [15]

3.2 Sentiment in News

The analysis is carried out on publicly-available news reports. The news source used for this report is the BBC World News website. [16] Normal, publicly-available articles were retrieved and analysed from this source.

To analyse sentiment in news articles, the system looks at the words in the article. This presents the first problem. The system must know which words are positive and which words are negative in relation to oil. The aim of the system is to build or train two datasets. One dataset will be full of words that relate to positive sentiment and the other dataset will be full of words that relate to negative sentiment.

3.3 Building a dataset

For the project, two datasets were created. One dataset contained positive words and the other contained negative words.

To create a dataset, words in articles that are analysed are marked as positive or negative. If words were found to be neutral they were discarded. The log return of the close price for crude oil futures decides the positivity or negativity of an article.

After the system extracts all the words from the articles published on a specific date, it checks the log return for the following day.

If the log return is negative, this means that oil price went down the day after the articles were published. This implies that these articles contained negative sentiment and all the words published on that date are put into the negative words dataset.

If the log return is positive, this means that oil price went up the day after the articles were published. Conversely, this implies that the articles contained positive sentiment. As a result, the words published on that date are put into the positive words dataset.

\[
\text{if logreturn} > 0 : \text{sentiment is positive}
\]
if logreturn == 0: sentiment is neutral

if logreturn < 0: sentiment is negative

This process can be repeated daily over a range of dates. When this dataset has been created, it can be used to decide the polarity of other articles using a method detailed in the testing stage.

A dataset is a collection of words with the following properties:

- It does not include non-essential words
- Words are stemmed to their base
- Words may appear multiple times
- The same word may appear in both the positive dataset and the negative dataset

After words are extracted from articles, word processing is carried out. All words are converted to lowercase. This is to standardise the words. For example, “Home” and “home” would be regarded as two different words to a computer because of the capitalisation of the first letter. When both words are converted to lowercase, they are regarded as the same word. This is the ideal result.

After words are converted to lowercase, punctuation within and around the words is removed. Again, this is to standardise each word and to make them equal. For example, “bad.” and “bad” would not be regarded as the same. To a human they are the same but to a computer they are regarded as two different strings. Punctuation is removed to solve this problem.

Examples of this:

“don’t” becomes “dont”

“bad.” becomes “bad”

It is important to note that punctuation may play a role in detecting sentiment in text. An exclamation mark or a question mark can change or emphasise the meaning of a word. This was not included in this system but is discussed in chapter 5 regarding further work.

Words that are common in the English language are also removed. All languages include words that are known as “glue words”. These are words that hold a sentence together. These include conjunctions, prepositions and certain pronouns. Generally, these types of words won’t have an effect on the sentiment and meaning of a sentence. As a result, these words can be discarded. The opposite of these words are sometimes referred to as “working words”.

23
These include nouns, verbs and adjectives. Working words are generally regarded as being important to the sentiment and meaning of a sentence.

A definitive list of glue words is not used in the system but the idea is used to remove a set of words. This set of words is known as stop words. It removes words that commonly appeared but wouldn’t be regarded as important.

The idea is that these words won’t be more positive than negative or more negative than positive. They will show up in both datasets a similar number of times. By removing them, the system does not have to analyse them which means the process may run faster. If they are not removed, the outcome should not be affected.

Along with removing common words, “junk words” must also be removed. Junk words might include pieces of HTML code that got retrieved with the article’s contents or URLs within the article. Words like this are useless and will not have any effect on the outcome.

Finally, the last step to text processing is to stem each word. Stemming is the process of removing affixes in words. For example, the words “complicated”, “complications”, “complicating” and “complicates” would all be stemmed to “complic”. When one of these words appears, it is stored in the datasets as its stemmed value. The result of this process means that each of these words will be regarded as the same word. This will mean that a word mentioned in its various forms will have higher importance than if each word were regarded as separate words. This process is done with the Natural Language Toolkit module for Python which was previously mentioned. It is important because while a word might be written in various ways, the meaning of the word is similar and may produce the same sentiment.

After the word processing is complete, the leftover words are the words that need to be classified as positive words or negative words. As mentioned previously, the polarity for a set of words is decided by the log return value for the following day.

To calculate the log return for the following day, the time series information is analysed. The log is taken of the close price divided by the close price of the previous day.

\[
logreturn = \log\left(\frac{\text{close price of today}}{\text{close price of previous day}}\right)
\]
3.4 Testing the datasets

The building stage can be thought of as a continuous process which can be expanded and improved nearly indefinitely. For the purpose of this report and for testing purposes the building stage was carried out for a year and a half. This means that for a year and a half, each day’s articles were retrieved and analysed. The result is two large datasets full of words. Each dataset contains over 500,000 words.

Now, instead of relating words to the return rate value of the crude oil future prices for the following day, the previously trained datasets may be used to extract sentiment for a day.

The overall sentiment for a day is decided based on calculations using the frequency of the words in each set.

Articles are retrieved in an identical method to the building stage and words are extracted using the same word processing techniques. Now with a list of words from a specific date, calculations on the frequencies of the words can be carried out.

Each word is assigned a positivity score and a negativity score. The positivity score is a result of getting the frequency of the word in the positive word dataset and dividing it by the total frequency of the word in the positive and negative word datasets.

The negativity score is a result of getting the frequency of the word in the negative word dataset and dividing it by the total frequency of the word in the positive and negative word datasets. Chapter 4 presents examples of words and their individual scores.

\[
\text{PositivityScore} = \frac{\text{freq.\text{word in positive dataset}}}{(\text{freq.\text{word in positive dataset}} + \text{freq.\text{word in negative dataset}})}
\]

\[
\text{NegativityScore} = \frac{\text{freq.\text{word in negative dataset}}}{(\text{freq.\text{word in positive dataset}} + \text{freq.\text{word in negative dataset}})}
\]
Each positivity score can be summed up to give an overall positivity score and each negativity score can be summed up to give an overall negativity score.

\[
\text{Overall Positivity Score} = \sum_{i=0}^{n} \text{Positivity Score}_i + \text{Positivity Score}_{i+1} + \ldots + \text{Positivity Score}_n
\]

\[
\text{Overall Negativity Score} = \sum_{i=0}^{n} \text{Negativity Score}_i + \text{Negativity Score}_{i+1} + \ldots + \text{Negativity Score}_n
\]

The polarity for a date is decided by whichever overall score is higher. If the overall positivity score is higher, the sentiment is declared as being positive. If the overall negative score is higher, the sentiment is declared as being negative.

This test process can be executed for a single date or for a range of dates. If a single date is input, the system will produce the sentiment for that date. If a range of dates is given, the system will step through the dates and produce a sentiment result for each date.

This method is known as a lexicon based approach to sentiment analysis. It is commonly referred to as the bag-of-words-approach. Relations between individual words are not considered. The sentiment of every word is determined and then combined with an aggregation function. [10]

3.5 Data Analysis

The data analysed and used in the above method is time series data for continuously traded crude oil future contracts. This time series data consists of five minute oil price data in the form of CSV (comma separated values) formatted text files.

These files are analysed using Python and using the matplotlib library to visualise information.
3.5.1 Compression

The files for the oil data are large and contain a lot of information. In order to analyse the data for different time intervals a python program was written to compress the data to the input time.

From example, the data files contain five minute data which is a piece of information for every five minutes. The system allows the user to compress this information to hourly or daily information data.

3.5.2 Graphing

Using the matplotlib library for Python, graphs were generated to display a variety of information. The time series information is able to be graphed showing various different days, months (Figure 4) and years (Figure 5).

![Future Prices](image)

Figure 4 Crude Oil Future Prices for April 2011
Figure 5 Crude Oil Future Prices for the year 2002
4 Experiments & Evaluation

After applying the method to build the positive word dataset and the negative word dataset, two complete datasets were successfully built. The building process was run continuously for a year and six months’ worth of news articles. This produced two datasets, both containing over five hundred thousand words.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Words</td>
<td>562,962</td>
</tr>
<tr>
<td>Negative Words</td>
<td>557,687</td>
</tr>
</tbody>
</table>

An example of the contents of a dataset can be seen in Figure 6 and Figure 7. These two word clouds are a visualisation of samples of a dataset. They were generated with a word cloud generator [17]. These are just small samples of a dataset because a whole dataset is too large to visualise with this generator. In these word clouds, the frequency of a word is shown by the word size. The larger a word, the more frequent it is. It is apparent that the largest and therefore most frequent word in these samples is the word “oil”. This is expected because the articles that are searched for are explicitly related to oil. Other large words include places such as “middle east”, “world” and “iraq”. These are large because of where the news is reported from and where it is reported about.

An example of the stemming technique that was mentioned earlier can be seen in the word clouds.

Other words in the word clouds that would appear to make up the oil domain ontology include “threat”, “fear”, “deal”, “warn”, “sanction”, “saddam”, “gaddafi”, “war”, “attack” and “govern”. These words may appear in both datasets and as such they get a positivity score and a negativity score during the test stage.
Figure 6 Visualisation of a dataset

Figure 7 Visualisation of a dataset
Table 5 Words and their frequencies in the positive word dataset

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>oil</td>
<td>2839</td>
</tr>
<tr>
<td>year</td>
<td>1803</td>
</tr>
<tr>
<td>not</td>
<td>1797</td>
</tr>
<tr>
<td>world</td>
<td>1680</td>
</tr>
<tr>
<td>had</td>
<td>1630</td>
</tr>
<tr>
<td>more</td>
<td>1625</td>
</tr>
<tr>
<td>after</td>
<td>1511</td>
</tr>
<tr>
<td>say</td>
<td>1462</td>
</tr>
<tr>
<td>about</td>
<td>1415</td>
</tr>
<tr>
<td>one</td>
<td>1381</td>
</tr>
<tr>
<td>over</td>
<td>1361</td>
</tr>
<tr>
<td>some</td>
<td>1269</td>
</tr>
<tr>
<td>last</td>
<td>1238</td>
</tr>
<tr>
<td>other</td>
<td>1223</td>
</tr>
<tr>
<td>could</td>
<td>1222</td>
</tr>
<tr>
<td>new</td>
<td>1214</td>
</tr>
</tbody>
</table>

Table 6 Words and their frequencies in the negative word dataset

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>oil</td>
<td>2934</td>
</tr>
<tr>
<td>not</td>
<td>1835</td>
</tr>
<tr>
<td>year</td>
<td>1816</td>
</tr>
<tr>
<td>world</td>
<td>1744</td>
</tr>
<tr>
<td>had</td>
<td>1719</td>
</tr>
<tr>
<td>more</td>
<td>1652</td>
</tr>
<tr>
<td>after</td>
<td>1539</td>
</tr>
<tr>
<td>say</td>
<td>1511</td>
</tr>
<tr>
<td>about</td>
<td>1449</td>
</tr>
<tr>
<td>one</td>
<td>1416</td>
</tr>
<tr>
<td>over</td>
<td>1358</td>
</tr>
<tr>
<td>last</td>
<td>1285</td>
</tr>
<tr>
<td>some</td>
<td>1364</td>
</tr>
<tr>
<td>could</td>
<td>1254</td>
</tr>
<tr>
<td>busi</td>
<td>1235</td>
</tr>
<tr>
<td>two</td>
<td>1234</td>
</tr>
</tbody>
</table>
Words appearing in Table 5 and Table 6 are the most frequent words in the datasets. As can be seen these words are not very controversial and some appear as the most common in both datasets. This is to be expected and as a result the words that are very common to both have little effect on the overall polarity of the day’s articles. Some of these words could be added to the list of stop words mentioned earlier.

Table 7 Words and their positivity and negativity scores

<table>
<thead>
<tr>
<th>Word</th>
<th>Positivity Score</th>
<th>Negativity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>saddam</td>
<td>0.5705</td>
<td>0.4295</td>
</tr>
<tr>
<td>war</td>
<td>0.5068</td>
<td>0.4932</td>
</tr>
<tr>
<td>fear</td>
<td>0.5121</td>
<td>0.4879</td>
</tr>
<tr>
<td>fall</td>
<td>0.4974</td>
<td>0.5026</td>
</tr>
<tr>
<td>doubt</td>
<td>0.4926</td>
<td>0.5074</td>
</tr>
<tr>
<td>horror</td>
<td>0.5806</td>
<td>0.4194</td>
</tr>
<tr>
<td>demand</td>
<td>0.4698</td>
<td>0.5302</td>
</tr>
<tr>
<td>seriou</td>
<td>0.4275</td>
<td>0.5302</td>
</tr>
<tr>
<td>fundamentalist</td>
<td>0.6190</td>
<td>0.3810</td>
</tr>
<tr>
<td>sanction</td>
<td>0.5222</td>
<td>0.4778</td>
</tr>
<tr>
<td>bribe</td>
<td>0.5570</td>
<td>0.4430</td>
</tr>
<tr>
<td>crash</td>
<td>0.4551</td>
<td>0.5449</td>
</tr>
<tr>
<td>bomb</td>
<td>0.5303</td>
<td>0.4697</td>
</tr>
<tr>
<td>ulcer</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>discovery</td>
<td>0.4749</td>
<td>0.5251</td>
</tr>
<tr>
<td>disput</td>
<td>0.5198</td>
<td>0.4802</td>
</tr>
<tr>
<td>volatil</td>
<td>0.4832</td>
<td>0.5168</td>
</tr>
<tr>
<td>stockpil</td>
<td>0.4082</td>
<td>0.5918</td>
</tr>
</tbody>
</table>

Table 7 shows a selection of words in their stemmed form and their individual positivity and negativity scores. If these words appeared in an article, these are the scores they would contribute to the overall positive score and overall negative score discussed above. From the table we can see words such as saddam, horror and bribe have a higher positivity score meaning they have positive sentiment. It is important to note that this is positive sentiment in terms of traders making a profit. Positive sentiment correlates to prices rising. The most extreme scoring is from the word ulcer. This is more than likely an error. The data retrieval section of building the dataset probably encountered a health topic regarding some type of oil to help ulcers rather than a topic about crude oil. Since the word is not commonly used in oil articles, there isn’t enough information to statistically relate it to either positive or negative sentiment. Words like this or new words (a new president’s name for example) will disrupt
the results from the dataset. The next most extreme word is “fundamentalist”. This is correlated with positive sentiment with oil. Words will not always follow our intuition but sometimes they do. For example, the words “crash” and “doubt” correlates with prices falling.

If Table 7 were an article containing just those words, the system would mark it as having positive sentiment using the method outlined in chapter 3.

See Figure 8 and Figure 9 for visualisations of sentiment relating to oil price. These graphs visualise the correlation found between sentiment and price. 1.0 indicates positive sentiment and -1.0 indicates negative sentiment. When sentiment is shown as positive, it can be seen that a rise in price follows. When sentiment is shown as negative, it can be seen that a fall in price follows. As sentiment remains high, oil price keeps rising. When the sentiment turns negative, the price drops.

As detailed in chapter 3, the datasets created were tested. These findings can be seen in Table 8. The success rate ranges from 40% to 63.64%. The datasets are sensitive to new words being introduced in a late stage of the building process. For example, if the building stage runs for a yearlong period and it encounters a new word on the last week of the process, there won’t be enough statistical information to score that word properly. An unrelated and uncommon word such as “ulcer” in Table 7 can also disrupt results.

<table>
<thead>
<tr>
<th>Date Range</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/05/2004 – 25/05/2004</td>
<td>57.1% Correct</td>
</tr>
<tr>
<td>01/10/2007 – 31/10/2007</td>
<td>40% Correct</td>
</tr>
<tr>
<td>09/04/2008 – 23/04/2008</td>
<td>63.64% Correct</td>
</tr>
<tr>
<td>06/05/2008 – 15/05/2008</td>
<td>62.5% Correct</td>
</tr>
<tr>
<td>01/08/2009 – 11/08/2009</td>
<td>62.5% Correct</td>
</tr>
<tr>
<td>01/03/2010 – 31/03/2010</td>
<td>56.2% Correct</td>
</tr>
<tr>
<td>01/05/2010 – 31/05/2010</td>
<td>40% Correct</td>
</tr>
<tr>
<td>01/06/2010 – 31/06/2010</td>
<td>50% Correct</td>
</tr>
<tr>
<td>01/01/2011 – 31/01/2011</td>
<td>45% Correct</td>
</tr>
<tr>
<td>01/06/2011 – 30/06/2011</td>
<td>45.45% Correct</td>
</tr>
</tbody>
</table>
Figure 8 Sentiment and Oil Price for January 2006

Figure 9 Sentiment and Oil Price for June to July 2006
5 Afterword

The aim of this project was to investigate the correlation between sentiment in news articles and oil prices. This objective was completed by:

- Identifying which words relate to positive sentiment and which words relate to negative sentiment with the context of oil and oil prices.
- Retrieving news articles using data retrieval techniques.
- Building a positive word dataset and a negative word dataset using text processing.
- Analysing the datasets using sentiment analysis techniques.

All of these objectives were completed successfully and the aim of the project was satisfied.

The result of running the building process is two large datasets full of words that either have a positive relation to oil or negative relation to oil. These datasets were successfully built and were able to be used or tested.

The conclusions mentioned previously give reason to believe that if the datasets could be improved it might be possible to improve the overall rate of success for the system.

The datasets have been built using certain techniques but more processing or different techniques could be applied in attempt to better the results.

The system uses a single news source, the BBC World News website. While being regarded as a quality provider of news, it may not cover every single event that might be important to oil and oil price. To remove potential bias in news sources and to make sure the system has a full variety of oil news then it would be best to use a range of news sources. Unfortunately, adding in new sources is not a simple event.

News websites do not have a distinct layout and cannot be approached in the same manner. The same technologies that are mentioned previously can be used but the approach would need to be modified. In addition, JavaScript on a page may interrupt the system.

An approach of retrieving news from RSS feeds may be possible and might simplify getting news from multiple news sources.

With multiple news sources, the volume of news could also be measured. A study into whether the amount of oil news correlates to oil price could be carried out.
The system collects words by themselves. It takes every word out of its sentence and out of its context. It discards sentences and sentence structure. An example of this is the following sentence.

“The outlook is not good”

The current system takes the word “good” out of context and will put it in either the positive word dataset or negative word dataset. The system could be modified to look at two words instead of one.

So from the sentence above the following word pairs would be stored in a dataset:

“outlook not”

“not good”

This takes into account that “The” and “is” would be removed with the stop words. If the system stored them in the dataset as “outlooknot” and “notgood” then during the testing phase they could be matched by similar word phrases. This may produce more accurate positivity or negativity scores.

The technique of storing more than a single word in the dataset could work for any number of words including whole sentences. This might have some disadvantages such as very large datasets in terms of storage and it might be very computationally expensive or slow to build and test them. As mentioned previously, this is known as N-grams. [10]

The system discards punctuation in and around words as mentioned previously. It may be possible to extract sentiment from punctuation. For example, a question mark might change the sentiment of a word. This combined with storing more than a single word may provide a much better context for looking at sentiment.

Again, as the system only looks at words by themselves it is ignoring all properties that a full sentence has. This includes sentence length which could be analysed for importance.

Nouns, verbs and adjectives will be the words that give the highest positivity or negativity scores but the system does not differentiate between them. It may be possible that a noun is more important than a verb or that an adjective is more important than a noun. This could be accounted for.

The system looks at the correlation between news articles on a certain date and the log return for oil prices on the following date. This means news propagates instantly or in one day.
Investigations could be done into the correlation of news and oil price with a news propagation rate of two days. Similarly, propagation rate of one week or two weeks could be looked at. Another, but more difficult, thing to investigate might be which news articles propagate faster and which propagate slower.

Further study could be made into the pattern of sentiment in news and patterns in oil price. This would involve finding the sentiment in news for a range of dates and attempting to recognise a pattern within the information. A technical analysis could be performed on the crude oil future price data and attempts could be made to find a pattern within that information. The study could look into the correlations between the patterns and report on the findings.

All of these changes or improvements might play a role in enhancing the quality of both datasets. The datasets would be far larger as a result and have a more varied approach. This may enable the system to achieve a higher success rate.
Bibliography


List of Tables and Figures

Figure 1 Crude Oil Future Prices from 1993 until 2012 ............................................................ 13
Figure 2 Daily Log Returns of Crude Oil Future Prices from 2002 until 2012 ...................... 14
Figure 3 Search result for oil from the BBC News website ..................................................... 21
Figure 4 Crude Oil Future Prices for April 2011 .................................................................. 27
Figure 5 Crude Oil Future Prices for the year 2002 .................................................................. 28
Figure 6 Visualisation of a dataset ....................................................................................... 30
Figure 7 Visualisation of a dataset ....................................................................................... 30
Figure 8 Sentiment and Oil Price for January 2006 ................................................................. 34
Figure 9 Sentiment and Oil Price for June to July 2006 .......................................................... 34

Table 1 Top Five Countries by Oil Consumption ........................................................................ 9
Table 2 Top Five Countries by Oil Production ........................................................................... 9
Table 3 Top Five Countries by Oil Exports .............................................................................. 10
Table 4 Dataset Word Count .................................................................................................... 29
Table 5 Words and their frequencies in the positive word dataset ........................................... 31
Table 6 Words and their frequencies in the negative word dataset ........................................ 31
Table 7 Words and their positivity and negativity scores ......................................................... 32
Table 8 Success Rate of Analysing Sentiment ........................................................................ 33