ELECTRONIC, ELECTRICAL & SYSTEMS ENGINEERING

UNIVERSITY OF
BIRMINGHAM

BEng Final Year Project (EE3P)

Ahmed Zaidi

1153609

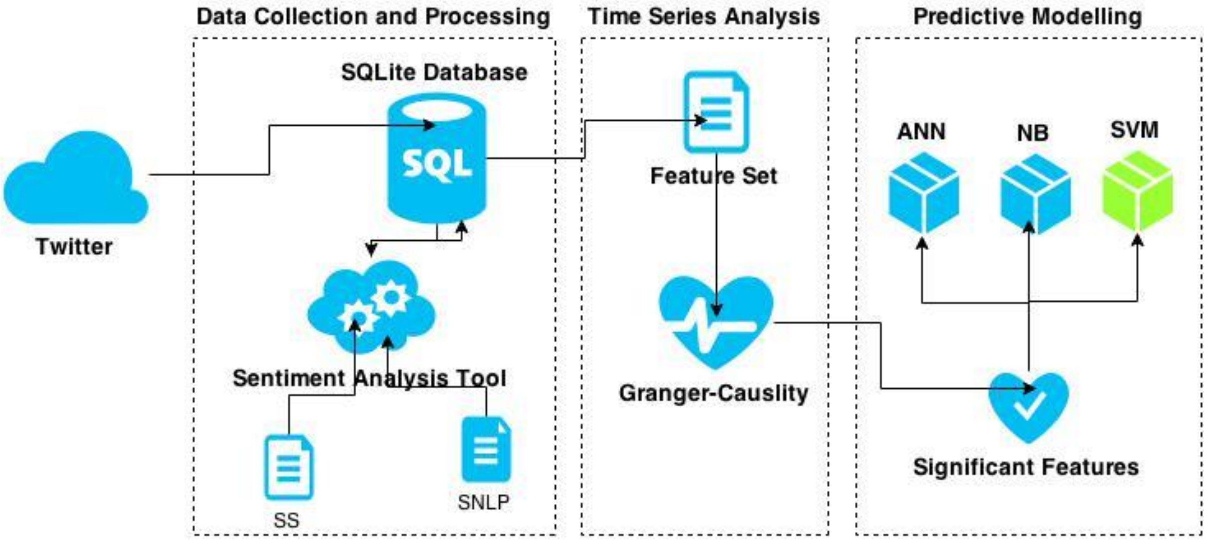
Forecasting Weekly WTI Crude Oil using Twitter Sentiment of US Foreign Policy and Oil Companies

Dr. Mourad Oussalah

Abstract

The drop in crude oil price during late 2014 has had a significant impact on all nations. While some countries have reaped the benefits of low oil prices, others have suffered greatly. As a result, it is no surprise that many academics have attempted to develop reliable models to forecast crude oil price. In the age of information and social media, the role of Twitter and Facebook has become increasingly more relevant in understanding our environment. Many academics have exploited this wealth of data to extract features including sentiment and word frequency to build reliable forecasting models for financial instruments such as stocks. These methodologies, however, remain unexplored for the prediction of crude oil prices. The purpose of this investigation to develop a novel model that uses sentiment of United States foreign policy and oil companies' to forecast the direction of weekly WTI crude oil prices. The investigation is divided into three parts: 1) a methodology of collecting tweets relevant to US foreign policy and oil companies'; 2) a statistical analysis of the novel features using Granger Causality Test; 3) the development and evaluation of three machine learning classifiers including Naïve Bayes, ANNs, and SVM to predict the direction of weekly WTI crude oil. The findings of the statistical analysis showed strong correlation between the novel inputs and WTI crude oil price. The results of the statistical tests were then used in the development of the predictive model. SVM was found to provide best forecasting performance. Furthermore, using these novel features, the predictive accuracy exceeded that of existing models mentioned in literature.

Graphical Abstract



Acknowledgements

I would like to express my special appreciation and thanks to my advisor Dr. Mourad Oussalah, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been invaluable.

I would also like to thank Ahmed Fetait, PhD candidate at the University of Warwick, who spent countless hours mentoring me through this journey. I truly appreciate your help.

List of Table

List of Figures

List of Acronyms

ANN – Artificial Neural Networks

SVM – Support Vector Machines

NB – Naïve Bayes

USFP – United States foreign policy

OC – Oil Company

OPEC - Organization of the Petroleum Exporting Countries

WTI – Western Texas Intermediate

MLP – Multilayer Perceptron

SS – SentiStrength Sentiment Analyser

SNLP – Stanford NLP Sentiment Analyser

Contents

Chapter 1

Introduction

1.1 Motivation

The drop in crude oil price during late 2014 has had a significant impact on all nations. While some countries have reaped the benefits of low oil prices, others have suffered greatly. Oil and gas revenues represent about 70 percent of Russia's export income. For every dollar the crude oil prices drops, Russia loses approximately $2 billion in annual revenues. European nations, on the other hand, have benefited from the lower oil prices. In light of the importance of crude oil to the global economy, it is not surprising that economists have spent a great deal of resources trying to predict its movements. Forecasting future prices will allow companies[1] mitigate their risks against price fluctuations.

Crude oil prices are impacted by a variety of external and internal factors. Some factors include production rates, foreign sanctions, economic growth and seasonal consumption. Traditionally, economists have focused on historical oil prices, supplementary commodity prices (e.g. coal, natural gas), inventory levels, and financial instruments prices (futures and options) as a method of predicting crude oil prices. Few academics have modelled the price of oil as a function of non-oil variables (e.g. US dollar exchange rate gap). However, there still remains a gap on the impact of the overall sentiment of foreign policy on oil price, specifically US foreign policy. In mid-2012, the US-led sanctions imposed against Iran reduced their oil production from 2.4 million barrels a day to 1.4 million barrels a day.

In the age of information and social media, websites such as Facebook and Twitter contain a plethora of knowledge that can be used to understand market behaviour and extract sentiment of individuals, companies and even countries. Studies have shown that Twitter sentiment of companies can used as inputs in predicting future stock prices. This methodology, to the best of my knowledge, has not been applied to crude oil prices. Therefore, this study aims to fill

[1] *Particularly Airline companies as oil prices represent are large aspect of their total cost*

the gap in the domain by applying social media forecasting methods with inputs of US foreign policy and oil companies' sentiment to forecast the direction crude oil prices.

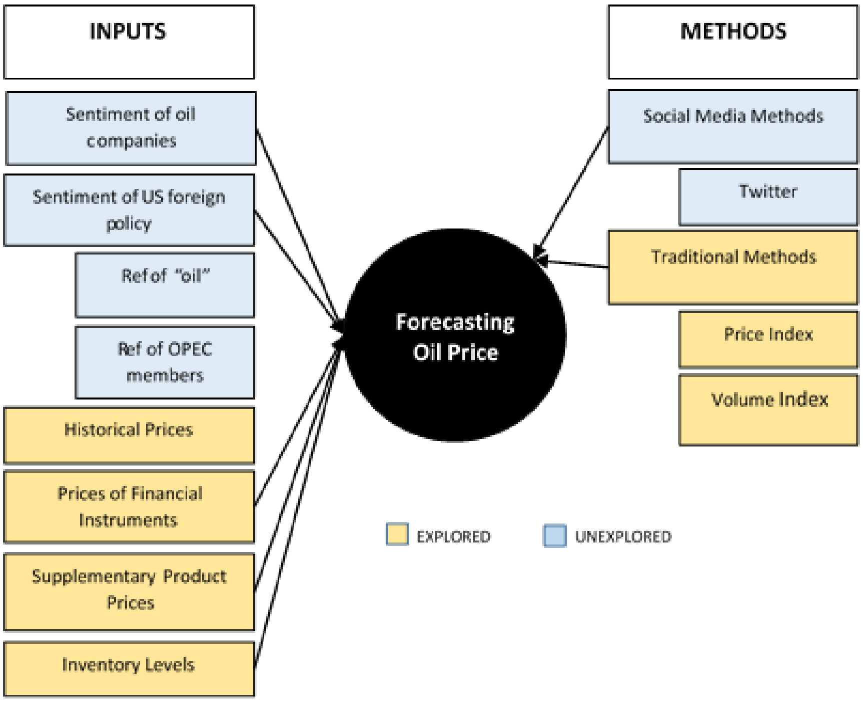


Figure 1.1 - A chart showing existing methodologies input of forecasting models for crude oil

1.2 Purpose of Investigation

The purpose of this investigation is to improve on the existing methods for forecasting the directional shift of crude oil price. The investigation attempts to achieve this aim by introducing **Twitter** sentiment of **US foreign policy (USFP)** and **oil companies' (OC)** as an input into the machine learning forecasting model for crude oil prices.

The objectives for this investigation are as follows:

1. Carry out a comprehensive literature review on the existing techniques used to forecast crude oil price
2. Develop and evaluate a novel methodology and tool to extract and store US foreign policy and oil companies' sentiment from Twitter into a relational database
3. Conduct a statistical causation and correlation study between US foreign policy and oil companies' sentiment with weekly WTI crude oil prices
4. Build and evaluate a predictive model to forecast weekly WTI crude oil prices using US foreign policy and oil companies' as an input
5. Create a hedging strategy using the predictive model developed to mitigate risk for companies
6. Identify limitations of the investigation and opportunities for future works

1.3 Contribution

This research offers **four** novel contributions to the area of crude oil predictive modelling:

1. A study that demonstrated that the sentiment of US foreign policy and oil companies' tweets between 2011 and 2014 has a statistically significant correlation with WTI crude oil prices.
2. Through additional feature extraction from the tweets, a study that shows the frequency of references to "oil" and OPEC members in the USFP tweets have an inverse correlation with the price of WTI crude oil.
3. Through the use of machine learning models, it was shown that USFP and OC sentiment, frequency of references to "oil" and OPEC countries in USFP tweets, serve as significant inputs in the directional forecasting model of weekly WTI crude oil prices.
4. An investigation that demonstrated that the cumulative impact of USFP and OC sentiment and the frequency of references to "oil" and OPEC members in the USFP tweets takes **seven weeks** to work through the WTI crude oil price. This finding is consistent with studies done on the relationship between WTI crude oil price and gasoline prices[2].

[2] Amadeo (2012), GlobalPetrolPrices.com

1.4 Report Structure

The structure of the report is as follows:

- **Chapter 2** will provide a literature review into the existing methods of forecasting crude oil prices.
- **Chapter 3** will provide a background into machine learning models, specifically ANNs, SVM and Naïve Bayes
- **Chapter 4** will provide a background of sentiment analysis, specifically SentiStrength and Stanford NLP
- **Chapter 5** will provide an overview of the methodology of data collection, storing, and processing using Twitter API, SQLite, Excel, MATALB, and the sentiment analysers
- **Chapter 6** will provide an overview of the *Time Series Analysis* study. This is an experiment that uses Granger-Causality test to identify the correlation between various features and weekly WTI crude oil prices.
- **Chapter 7** will outline the *Predictive Modelling* study. This is an experiment that builds and evaluates a predictive model to forecast weekly WTI crude oil prices using various features include sentiment.
- **Chapter 8** will discuss the overall investigation. It will address the extent to aims and objectives were achieved. The chapter will also discuss the limitations of the investigation. Finally it was identify future works based on the findings of this investigation.

Chapter 2

Review of Literature

The aim of the investigation is to improve on the accuracy for existing methods of forecasting crude oil price by using Twitter sentiment of USFP and OC as inputs. Therefore, in order to meet this objective, a comprehensive literature review must be conducted. Chapter 2 provides a background on the models mentioned in literature to forecast crude oil prices. This chapter will also discuss the limitations of these models.

2.1 Background on Forecasting Crude Oil Price

In literature, models have tried to forecast a variety of different grades of crude oil. The main types include Western Texas Intermediate (WTI), Brent, Dubai, Oman, and Urals. For this investigation, **WTI** crude oil prices will be forecasted. This is due to the amount of historical data available for this oil grade. Furthermore, as WTI has been widely used in literature providing a more comprehensive comparison of results. WTI is traded on the New York Mercantile Exchange (NYMEX) and has been used historically as the benchmark in oil pricing.

Oil price forecasting models can be categorized into two main types: **quantitative** and **qualitative** models. Quantitative models are based on quantitative historical data and mathematical models. Furthermore, they generally forecast short to medium term oil prices. Qualitative models combine the inputs from quantitative models with additional inputs such as isolated events e.g. natural disasters, political factors (elections, revolutions). The majority of this review will focus on assessment of quantitative models in literature. Within quantitative models there are **econometric models** and **non-standard models**. Econometric models are further divided into **time-series models, financial models,** and **structural models.** The division of econometric models is based on the input variables used to forecast the oil price.

2.1.1 Qualitative Models

In literature, time-series models have been implemented using Naïve models, exponential smoothing models and autoregressive models (ARIMA, ARCH/GARCH). Pindyck (1999) incorporates an autoregressive model to forecast crude oil prices from 1887 to 1996. However, results indicate poor forecasting ability. Radchenko (2005), who built on Pindyck (1999)'s model attributed the inadequate forecasting to the inability to consider OPEC behaviour. Wang et al. (2005) and Xie et al. (2006) compare linear ARIMA models with non-linear artificial neural networks (ANN) and support vector machines (SVM) and discovered that ARIMA out performs only in very short horizons while ANN and SVM models have superior performance for medium and long term forecasting. Fernandez (2010) concludes similar findings in his research where ARIMA model underperforms in medium to long term forecasting. There is a general consensus in literature that linear models are less reliable in forecasting oil prices as compared to non-linear models. This is not surprising given the non-linear nature of the oil price and its variables.

Mohammadi and Su (2010) use various GARCH and exponential GARCH models to forecasting weekly data of crude oil spot price. The results indicate that the nonlinear GARCH models outperform the other models. Silva et al. (2010) implements a hidden Markov model (HMM) to forecast medium term crude oil price movements. Using this non-linear approach, the author achieves a mean forecasting accuracy of 57%. The conclusion drawn from literature is that time-series models do not provide an accurate means of predicting medium to long term oil prices. This is a fundamental issue as most investors are interested in hedging their risk against the long term fluctuation of prices as opposed to short term.

The next type of econometric models are financial models. These models use the relationship of financial instruments such as futures and forward contacts to predict future crude oil spot prices[3]. Chin et al. (2005) examine energy futures prices to accurately forecast future spot prices. The research suggests that future prices are unbiased predictors of spot prices and outperform time-series models. In contract, Chernenko et al (2004) finds that futures prices were not an efficient method of predicting future spot prices. Although there seems to be a

[3] *spot prices* are current prices at which a security can be bought or sold

strong correlation between spot prices and futures prices, the results from literature reveal that futures prices cannot consistently and accurately determine future spot prices and is therefore are not a reliable method of forecasting future oil prices. However, Chen (2014) addresses this issue by using an adaptation of financial models to predict oil prices. The author uses oil-sensitive stocks using a linear regression model to forecast future WTI crude oil prices. His investigation reveals AMEX oil index is a superior predictor to alternative inputs with a classification accuracy 63% for the directional forecast for 1-month ahead WTI crude oil price.

The third type of econometric models are structural models. These models use a range of variables to predict the price of crude oil. Some of these variables include OPEC behaviour, inventory levels, oil consumption and production. In literature, many structural models have used economic activity, interest rate and other non-oil variables as the input for the forecasting model. Ye et al. (2006) use inventory levels and non-linear models to accurately predict the 1-month ahead nominal[4] WTI forecasting price. This was an adaptation of the previous work Ye et al. (2005) did with linear models. The new study revealed that the non-linear model performed significantly greater. Mirmirani and Li (2004) uses oil supply, petroleum consumption, money supply and WTI crude oil futures prices to forecast the movements of US oil price. The results reveal that ANN model outperforms the linear VAR model. The limitations with structural models is that despite providing strong relationships between variables in certain models, in many cases the future values of those variables may be required to determine future oil prices.

Non-standard models or computational models are non-linear techniques to forecast prices of crude oil. These models use methods that do not clearly fit into any specific sub-econometric models (time-series, structural, financial) and therefore get categorized under "non-standard models". Shambora and Rossiter (2007) predicts the direction of daily crude oil prices using an ANN model that build with the price of crude oil futures contracts as the input. The study reveals that ANN model performance is statistically superior to other traditional models with a classification accuracy of 53.10%.

[4] nominal price refers to whether the price of oil will increase or decrease

2.1.2 Qualitative Models

There are very few studies on the qualitative models in literature. Primarily due to the motivation behind developed forecasting models. Investors are most interested in a model that is consistently applicable rather than a model that is based on historical isolated events. Qualitative models use additional inputs such as natural disasters and political factors to forecast the price of price of crude oil. Wang et al. (2005) use a novel approached called TEI@I to predict WTI crude oil prices. The author investigates the effect of infrequent events an irregular events on oil prices by implementing techniques such as Web-based Text Mining (WTM). The results indicate that the non-linear TEI@I produced superior results to the linear ARIMA model. Ghaffari and Zare (2009) use historical oil spot prices and Adaptive Network-based Fuzzy Inference System (ANFIS) to forecast the direction of daily WTI crude oil prices from 5/1/2007 to 5/31/2007. This is done by removing short term disturbances experienced in the spot prices of oil over time. The model achieves a 68.18% classification accuracy, superior to other models that predict the sign of oil price movement including Morana (2001) with 46.67%, Gori et al. (2007) with 45.76% and Fan et al. (2006) with 54.54%. However, this model does not provide any practical application as investors require models that are able to predict the price of oil over a long-horizon. Although, qualitative methods have achieved considerable accuracy, they are based on irregular and infrequent events and therefore are not are limited in their practical application.

2.2 Conclusion

The literature review has revealed many findings about the existing methods of forecasting crude oil. Firstly, non-linear models outperform linear models in medium and long term horizons. Therefore, as investors are more concerned with the long term risk mitigation, the intrinsic value of a model is dependent of its forecasting horizon; the longer the better. The second finding is that forward and futures[5] contract prices and historical oil prices[6] are poor indicators of future oil price. The literature review also reveals that text mining has only been used in methods to predict the impact of irregular and isolated events[7]. Furthermore, although qualitative models account for the impact of political factors on oil price, they have done so in a limited capacity.

[5] Chernenko et al (2004)
[6] Wang et al. (2005); Xie et al. (2006)
[7] Wang et al. (2006)

Forecasting techniques outside of the crude oil domain have revealed other interesting methods within this domain. With the abundance of data on Twitter and Facebook, social media has played an increasingly significant role in understanding the market. Many academics have tapped into the social media corpus as a novel way to forecast various aspects of business, politics and finance. Zhang et al. (2010) used sentiment of tweets to identify the directional shift in Dow Jones, S&P 500, and NASDAQ. Gilbert and Karahalios (2010) used emotions to predict the stock market movement. Boolen et al (2011) gathered 9 million tweets to investigate whether public mood can be correlated with the value of Dow Jones Industrial Average (DJIA) over time. Literature reveals that Twitter has proven to be reliable method of forecasting political, societal and financial factors. However, to the best of my knowledge, Twitter has not been used in the forecasting of WTI crude oil. A search on Science Direct database, ProQuest database and with the following queries oil+sentiment+forecast or oil+twitter+forecast returns zero results. A similar search on Google Scholar does not return any relevant results either.

The literature review on forecasting crude oil techniques combined with the increasing interest in social media as a variable in forecasting various aspects of finance, we can justify many decisions made during development of the investigation question. This investigation will use non-linear machine learning classifiers as literature has suggested they provide stronger forecasting ability. The inputs of the models will be based on previously unexplored but theoretically sound factors that impact crude oil: political factors (US foreign policy), oil company behaviour, and OPEC behaviour[8]. The investigation will use social media as a primary source of data for the extraction of these variables. As literature shows that sentiment has been used to successfully forecast the price of financial instruments[9], the investigation will use sentiment as the key indicator of US foreign policy and oil company behaviour.

Chapter 3 and *Chapter 4* are dedicated to providing a detailed overview and technical understanding on how machine learning classifiers, and sentiment analysis tools work, respectively.

[8] The lack of OPEC behaviour consideration was identified as a detriment in ARIMA models by Radchenko (2005)

[9] Boolen et al (2011)

Chapter 3

Background Machine Learning Classifiers

As mentioned in literature, non-linear machine learning classifiers provide significantly greater predictive capabilities as compared linear methods. As a result this chapter aims to provide a background into the three popular machine learning classifiers: Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Naïve Bayes (NB). ANN and SVM were selected for this investigation for their superior performance in forecasting crude oil[10] and the number of literature comparisons available. The following authors have used ANNs or SVMs to forecast crude oil: Yu et al. (2008); Tehrani, Khodayar (2011); Movagharnejad et al (2011); Jammazi, Aloui (2012). The Naïve Bayes classifier was selected due to its superior classification accuracy[11].

3.1 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are a group of learning models inspired by the biological neural networks. In this investigation we will utilize a Multi-Layer Perceptron (MLP) model. MLP is a Feed-Forward Neural Network that is made up of: a) an input layer with input units; b) hidden layer with hidden units; c) an output layer with output units. Each unit is known as a *perceptron* or an artificial neuron.

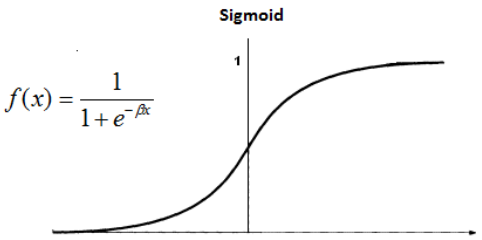


Figure 3.1 - *A graphical representation of the sigmoid function (Saedsayad.com, 2015)*

Suppose i_1, i_2, i_n are inputs to units and o_u,o_2, o_n are outputs to perceptron *u*. Perceptron *u* will only "fire" a signal (o_u) if $i_u > 0$. The variable i_u can be represented by the following equation:

$$i_u = \sum_{n=1}^{N} o_n w_{n,u}$$

[10] See literature review, Chapter 2, for additional information.

[11] The Naïve Bayes classifier had strong predictive capabilities in forecasting WTI crude oil prices. For more details please see results, Chapter 7.

The output value o_u of a perceptron is determined by the activation function g which in MLPs is represented by the sigmoid function (non-linear) of *g (refer to Figure 2.1)*:

$$o_u = g(i_u)$$

$$g(x) = \frac{1}{1 + e^{-\beta x}}$$

The combining of perceptrons or artificial neurons is known as artificial neural networks. An MLP consists of multiple layers of neurons where each layer is fully connect to the next one. See *Figure 2.2* for a visual representation. A single-hidden layer MLP consist of one input layer, one hidden layer, and one output layer. This model is used to characterize data using *linear* decision boundaries. A two-layer MLP contains two hidden layers and can be used to characterize data using arbitrary decision boundaries[12].

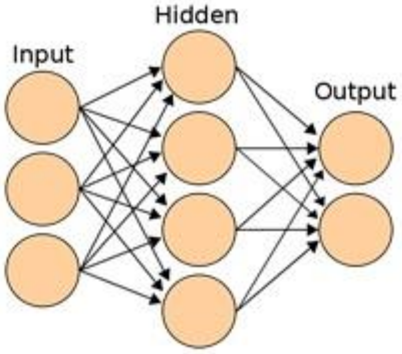


Figure 3.2- *A figure showing the structure of ANNs (Stanford, 2013)*

When building an MLP, one must define the following parameters: *number of layers, number of input units, number of hidden units, and number of output units.* Determining the optimal parameters for an MLP model is done through experimenting and tweaking[13]. Once the parameters have been optimized, the performance of the MLP is determined by the weights w for each of the nodes in the MLP which is determined automatically through a method called **Error-Back-Propagation (EBP)**. The training dataset will contain input vectors *i* with a target output vectors *t(i).* The input vector *i* will propagate through the network to produce an output *o(i)*. The error *E* will then be calculated using the following formula:

$$E = |\, t(i) - o(i)|$$

EBP then calculates $\frac{\partial E}{\partial w}$ for each *w* by propagating back up through the network. Completing this process with each of the instances in the training data, the average of $\frac{\partial E}{\partial w}$ is subtracted from original w. This process is repeated until the error falls below a defined threshold. The number of units in the output layer is the number of classes in the classification problem. The

[12] (Russell, 2014)

[13] (Russell, 2014)

class is determined by the output pattern of the output units realized at the output layer of the MLP.

3.2 Support Vector Machines (SVMs)

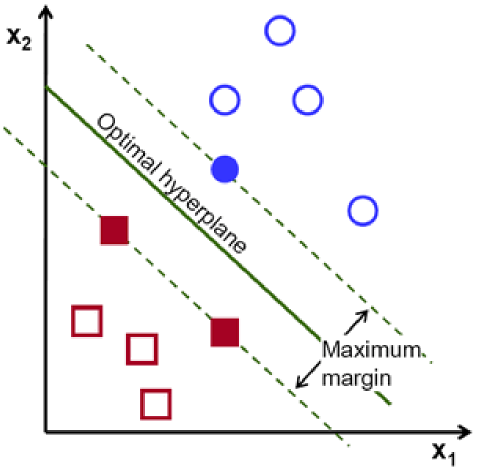


Figure 3.3- A figure showing the main concept behind SVM (maximizing margin) *(OpenCV, 2015)*

Support Vector Machines (SVMs) have gained considerable popularity in the machine learning domain due their superior performance in many classification problems. To put it simply, in a linear separable dataset, SVM techniques determine the optimal hyper-plane to divide two classes. In non-linear separations, the **kernel trick** is applied. The optimal hyper-plane is one that leaves the maximum margin between the two classes. The two elements in the training set that represent the maximum margin are known as **support vectors** *(refer to Figure 2.3 – support vectors are the filled in shapes)*. The support vectors are the only points in the training set that influence the optimality of the classification. The hyper-plane in a **2-dimensional** linearly separable dataset can be defined as follows:

$$g(x) = ax + by + c$$

In **higher dimensions,** the hyper-plane can be represented by the following equation:

$$g(\vec{x}) = \vec{\omega}_0^T \vec{x} + \omega_0$$

For both equations the classification of the instance in a binary class problem is determined by the following inequalities:

$$g(\vec{x}) \geq 1, \qquad \forall \vec{x} \in class\ 1$$

$$g(\vec{x}) \leq -1, \qquad \forall \vec{x} \in class\ 2$$

In order to optimize the SVM's performance we must maximize the margin between the two support vectors and the hyper-plane. The distance between the hyper-plane and each support vector can be represented by the following equation:

$$\frac{|g(\vec{x})|}{||\vec{\omega}||} = \frac{1}{||\vec{\omega}||}$$

Therefore the total margin can be represented by$\frac{2}{||\vec{\omega}||}$. In order to maximize the distance we must minimize$||\vec{\omega}||$. This can be done through a non-linear optimization technique known as **Karush-Kuhn-Tucker (KKT)** conditions, using **Langrange multipliers** λ_i.

3.3 Naïve Bayes

The Naïve Bayes classifier has often used as a method of classification due to its simple but effective implementation. It is based on the fundamental principles of the **Bayes Theorem:**

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

$P(c|x)$ is the **posterior probability**, $P(x|c)$ is the **likelihood,** $P(x)$ is the **predictor prior probability**, and $P(c)$ is the **class probability.** The key assumption that the Naïve Bayes classifier makes is that $P(c|X) = P(x_1|c) * \; P(x_2|c) * \ldots * P(x_n|c) * P(c)$ where $P(x_n|c)$ is the likelihood of attribute n occurring in class c. The training dataset is used to estimate $P(x_n|c)$ for each attribute.

In a binary classification problem, for each instance in the test dataset, $P(c|X)$ is computed for both classes. Each instances is classified as the class with the higher value of $P(c|X)$.

3.4 Conclusion

This chapter has provided a detailed overview of how the three selected classifiers for this investigation work. As mentioned in literature, non-linear models perform significantly better than linear models. ANNs and SVM are particularly prevalent in existing methods of forecasting crude oil. The novelty in this project lies in the types of inputs we provide the classifiers, sentiment of USFP and OC. As a result, the next chapter will provide a detailed overview of the tools that can be used to extract the sentiment from Twitter.

Chapter 4
Background on Sentiment Analysis

With the abundance of data on Twitter and Facebook, social media has played an increasingly significant role in understanding the market. Data on social media, through the use of Application Program Interfaces (APIs), has been easy to retrieve and manipulate. This allows statisticians and financial analysts to analyse large datasets that are more representative of the market. The common feature extraction from social media data is sentiment, through the use of natural language processing techniques. The ability for users to influence other users and thus the business through social media is a widely accepted notion amongst marketing and communication experts[14]. As a result, there is a growing importance in understanding how sentiment on social media influences different aspects of business e.g. stock market. This, coupled with the fact that literature has shown that sentiment has not previously been used to forecast WTI crude oil prices, brings novelty and justification to this investigation. With the main novel inputs in our classifying model being sentiment of United States foreign policy and oil companies' sentiment it is important to understand how the tools that extract sentiment work. This chapter sets out to provide an overview of the two different sentiment analysis tools that will be used in this investigation.

There are various sentiment analysis tools that can be used to extract sentiment from text. The tools are generally divided into two categories: *machine learning-based* and *lexicon-based*. For this investigation we will use **one** lexicon-based tool (SentiStrength) and **one** machine learning-based tool (Stanford NLP Sentiment Analyser). This is to eliminate the bias that may be prevalent by using just one type of sentiment analyser. Stanford NLP Sentiment Analyser was selected due to its superior performance in analysing short English phrases and its novel way model (RNTN) to classify sentiment[15]. SentiStrength was selected due to superior performance as compared to other machine learning algorithms in identifying positive sentiment.

[14] Gilbert and Karahalios (2010)
[15] (Stanford, 2013)

4.1 SentiStrength (SS)

SentiStength is a lexicon-based classifier developed at the University of Wolverhampton that uses additional linguistic information and rules to detect the sentiment strength of short English text. SentiStrength splits up the sentences by punctuation. The algorithm combines a lexicon, a look up table contains a list or words with associated strengths from 2 to 5, and additional rules such as: spelling correction algorithms, booster word lists, idioms list, negating words list, repeated letters list, emoticons list with polarities to identify sentiment of a phrase. The list or words was optimized by academics at the University of Wolverhampton using machine learning methods that trained and tested the list using various social media websites including Twitter, YouTube, and MySpace to name a few[16].

For each phrase, SentiStrength outputs two integers (positive strength and negative strength) ranging from 1(no sentiment) to 5(high sentiment). For example, “I love Birmingham, but only when it’s not raining” would output a 3 for positive and a 1 for negative. The output value of each integer is determined by the word with the highest strength. For example, if a sentence contains multiple positive words {2, 4, and 5} the positive integer will take the score of the highest positive word i.e. 5. The same applies for the negative integer.

For the purposes of this experiment, we will be calculating an aggregate sentiment which is defined as the difference between the positive and negative strength. This was done in order to standardize the variable thereby making it easier to compare to the other sentiment analyser being used in this investigation (SNLP).

$$sentiment = positive - negative$$

While many sentiment analysers detect the polarity of a phrase, very few provide sentiment strengths[17] . SentiStrength was chosen due to its ability to classify sentiment strength. In addition, as we are using Twitter as our main source of data, the fact that SentiStrength was designed for short informal texts makes it an ideal choice.

[16] Thelwall (2010)
[17] Pang & Lee (2005); Strapparava & Mihalcea (2008); Wilson et al. (2006)

4.2 Stanford NLP Sentiment Analysis (SNLP)

The other sentiment analyser beings used in this investigation is the machine learning-based Stanford NLP Sentiment Analyser. The SNLP is based on a new model that uses Recursive Neural Tensor Network (RNTN) and Stanford Sentiment Treebank. The Stanford Sentiment Treebank is the first corpus with fully labelled parse trees that allows for a complete analysis of the compositional effects of sentiment in language. The RNTN model was introduced to address the limitation of previous sentiment analysers that did not provide an accurate classification for shorter phrases, such as Tweets. RNTN uses tri-gram word vectors which act as inputs of the classification model.

The SNLP model, unlike existing Naïve Bayes, bi-gram Naïve Bayes, and SVM, accounts for word order. Most sentiment analysers analyse words in isolation. The authors of the model argue that by ignoring word order, important information is lost. For example the sentence "This movie was actually neither funny, nor super witty", contains two positive words "funny" and witty". However, the overall sentence is negative and SNLP identifies it as negative[18]. This same sentence tested in SentiStength returned a positive 4 and a negative 1, resulted in an aggregate value of positive 3.

The SNLP model maintains the highest classification accuracy at 80.7% when predicting fine-grained sentiments. Furthermore, the sentiment analyser does the necessary pre-processing including lower-cased, stop-word removal, HTML tags, and non-English removal. The sentiment analysis works on a 5-class scale (1- very negative, 2 – negative, 3 – neutral, 4 – positive, and 5 – very positive).

The SNLP model was chosen due to its significantly higher classification accuracy for when predicting sentiment. In addition, it also one of the few sentiment classifiers that provide sentiment strength rather than just polarity.

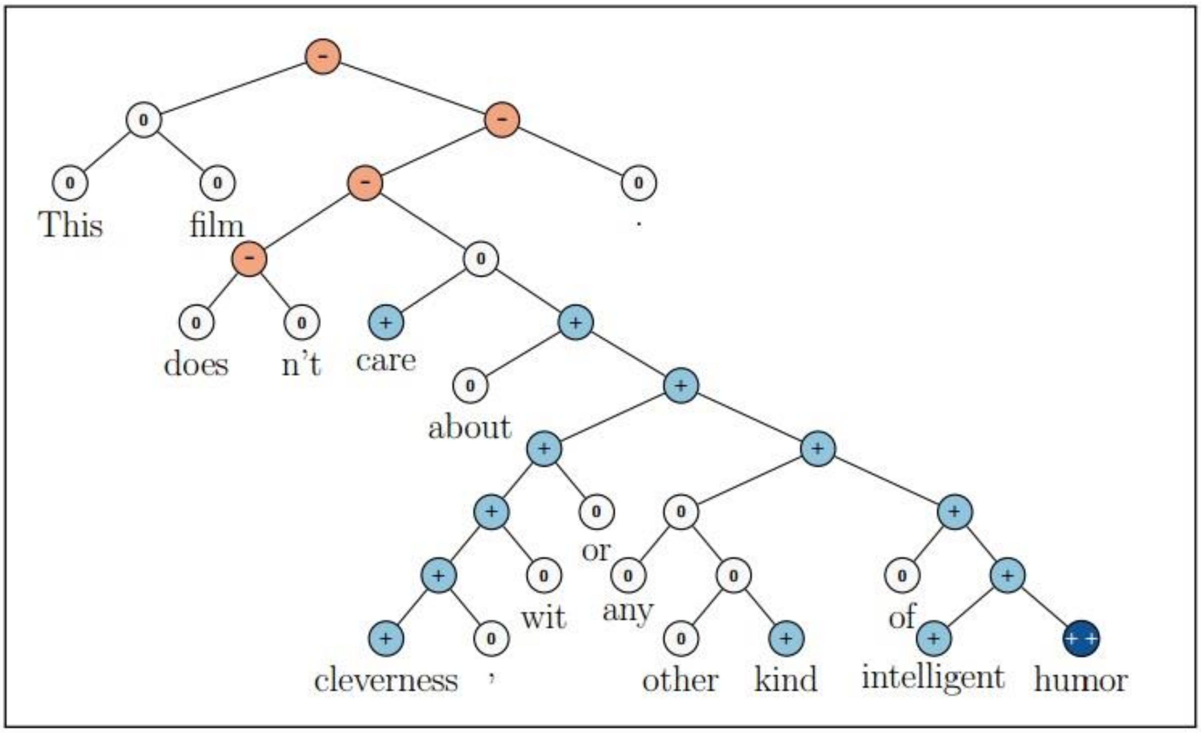


Figure 4.1 - *Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence. (Socher, et al., 2013)*

4.3 Conclusion

Sentiment has become a popular input in forecasting models, but remains unexplored as feature of future oil price. Therefore this investigation aims to bridge that gap by using SentiStrength and Stanford NLP to extract sentiment from Twitter. In order to minimize bias, one lexicon-based (SentiStrength) and one machine learning-based (Stanford NLP) tool were chosen. Both analysers were chosen due to the fact that they output strength as opposed to just polarity. SentiStrength was chosen as it has been specifically optimized for short informal English text, similar to the content of Tweets. Stanford NLP was chosen due to its high classification accuracy and novel methodology of accounting for word order.

The next chapter will discuss how the data will be collected from Twitter, the criteria that will be used to filter relevant data from irrelevant data, the pre-processing measures are taken to standardize the data, and the technical details on how to manipulate the aforementioned analysers to extract sentiment from the collected data.

Chapter 5

Data Acquisition and Processing

The decision to use USFP and OC sentiment as inputs to the forecast model was derived from the gap in the literature review. A significant component of this investigation is determining how to collect and identify what data is relevant. This chapter will outline the methodology and tools used for the acquiring the USFP and OC data from Twitter. Furthermore, it will provide an overview of the process of data standardizing and sentiment analysis using the tools described in *Chapter 4*.

5.1 Data Filtering and Criteria

The aim investigation is to identify whether Twitter sentiment of **USFP** and **OC** is useful in forecasting the directional shift of weekly WTI crude oil prices. A large part of assessing these factors involves acquiring the relevant data from Twitter. In addition, it also requires an assessment of what constitutes as USFP and OC data. The three key assumptions made when collecting data related to USFP and OC are:

- *Only tweets from credible or influential Twitter accounts constitute as relevant information when forecasting crude oil* - this claim suggests that tweets about USFP or OC from an unreliable and uninfluential users on Twitter have no significant impact on the WTI crude oil prices.
- *USFP data on Twitter is represented by the tweets made* ***by*** *US foreign policy and strategy think tanks' Twitter accounts*– this claim suggests that US foreign policy and strategy think tank tweets are a reliable source and representative of USFP data
- *OC data on Twitter is represented by the tweets made* ***by*** *the largest oil companies' and associations' Twitter accounts*– this claim suggests the largest oil companies and associations are a reliable source and representative of OC data.

The decisions made to select the best Twitter accounts for USFP and OC data are highlighted in *Section 5.1.1* for USFP data and *Section 5.1.2* for OC data.

5.1.1 USFP Data

A list of all the foreign policy and strategy think tanks based in the United States was acquired (EIA.gov). For each member of the list, a Twitter account was searched. In the case where there was no Twitter account or a relevantly inactive Twitter account, the member was removed from the list. Otherwise, the username of the member was recorded. This process was repeated for 100 think tanks. The final list was comprised of 76 think tanks that were actively operating Twitter accounts. Some of these think tanks included *The Center for Strategic and International Studies* and *The Council of Foreign Relations.* Please refer to *Appendix A* for a complete list of think tanks.

5.1.2 OC Data

In order to identify the relevant oil companies and associations a list of the 25 biggest oil companies by revenue was retrieved from Forbes. This, along with the rankings from Platts 250 was used to produce a list of 52 oil companies and associations that were operating active Twitter accounts. Additionally, well known energy administrations such as the United States Energy Information Administration (USEIA) were also included in the list. Please refer to *Appendix B* for a complete list of oil companies.

5.2 SQLite Database

Once the list of usernames was finalized, a SQLite database was setup to store the relevant tweets that are collected. SQLite is a relational database management system that unlike other database management systems, is not a client-server database engine. SQLite was chosen due to its simple and quick implementation process. Furthermore, SQLite is compatible with a variety of applications including MATLAB[19]. For the purposes of this investigation the creation of the SQLite database was done through Java. This is for seamless integration with SentiStrength, which is only available in Java. The pseudo code for the Java application is outlined below.

[19] Used to process data

```
c = connection to database;
s = statement pathway to database
sql statement =
             "CREATE TABLE USFP " +
             "(ID INTEGER PRIMARY KEY," +
             " TWEET TEXT NOT NULL," +
             " USER TEXT NOT NULL," +
             " DATE TEXT NOT NULL," +
             " SENTI INTEGER,"+
             " NLP INTEGER)";
Send sql statement to database via pathway;
Close statement pathway;
Close connection to database;
```

The pseudo code above creates a SQLite database called USFP that will contain all the tweets relating to USFP.

- ID – is a unique identification for each tweet that is stored
- TWEET - is the tweet that is retrieved
- USER - the username of the tweet sender
- DATE - is the date and time that the tweet was posted
- SENTI - is a placeholder for the future SentiStrength sentiment score
- NLP - is a placeholder for the future Stanford NLP sentiment score

The same process was repeated for the oil companies under the database titled OC and JUSTINB. JUSTINB are the latest tweets from Justin Bieber's Twitter account and will serve as a controlled variable the *Time Series Analysis* study (additional details in Chapter 6).

5.3 Twitter API

5.3.1 Data Acquisition

After completing the creation of the database, a tool was created to retrieve the required tweets from Twitter and store them in the SQLite database. In order to understand this process, it is important to introduce the Twitter API and how it operates.

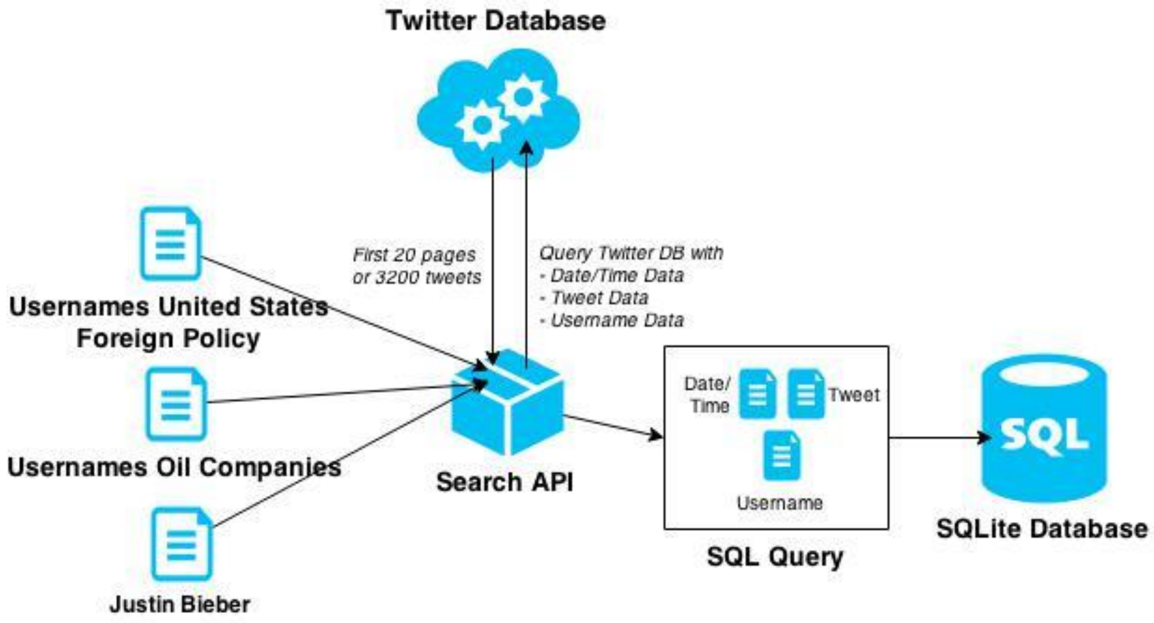
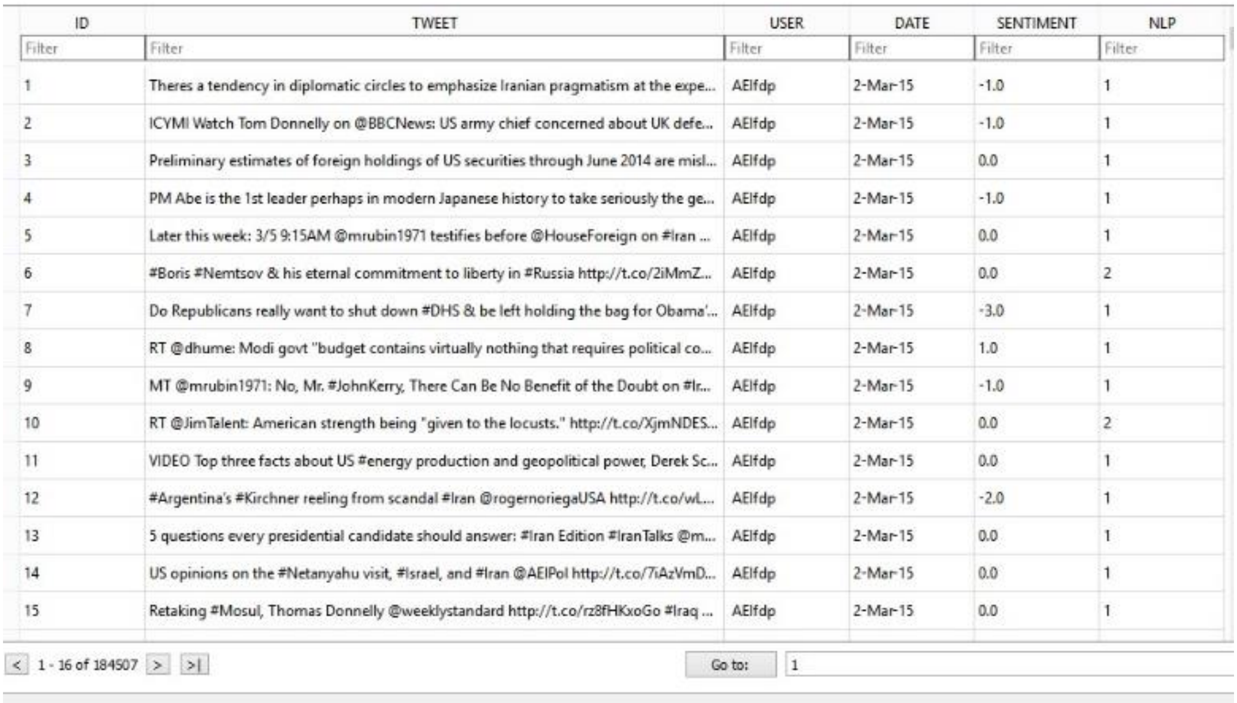


Figure 5.1 - *A block diagram showing the process used to retrieve tweets from Twitter and store them in a SQLite database*

An API is an Application Programming Interface that provide software developers with the building blocks needed to incorporate various aspects of the program. In the case of Twitter there are two main services: the Search API and the Streaming API. The Search API allows users to retrieve historical tweets and Streaming API allows the retrieval for live tweets. For the purposes of this investigation we will use the Search API. The Twitter API is available on all platforms, but for the purposes of this investigation Twitter API's java package (twitter4j) was used. This is to allow seamless integration of the sentiment analysis tool, SentiStrength, which is only available in Java. The approach taken to retrieve tweets in this investigation was the user_timeline approach. Each user's timeline on Twitter is divided into pages. This approach allows you to return the first 20 pages or the 3,200 most recent tweets (whichever comes first). In order to make this process more efficient a Java tool was created. The tool combined the aspects of the Twitter API and SQLite. See *Figure 5.1* for more details regarding the data collection process.

This process was repeated for both USFP and OC usernames. Data was also collected for Justin Bieber's username. Bieber's data will be used as a controlled variables in *Chapter 6*. At the end of data collection process, 184,507 tweets were collected for USFP from March

2015 going back to December 2009. 121,989 tweets were collected for OC from March 2015 to August 2009. See *Figure 5.2* for a screenshot of the SQLite database containing the tweets.



ID	TWEET	USER	DATE	SENTIMENT	NLP
Filter	Filter	Filter	Filter	Filter	Filter
1	Theres a tendency in diplomatic circles to emphasize Iranian pragmatism at the expe...	AEIfdp	2-Mar-15	-1.0	1
2	ICYMI Watch Tom Donnelly on @BBCNews: US army chief concerned about UK defe...	AEIfdp	2-Mar-15	-1.0	1
3	Preliminary estimates of foreign holdings of US securities through June 2014 are misl...	AEIfdp	2-Mar-15	0.0	1
4	PM Abe is the 1st leader perhaps in modern Japanese history to take seriously the ge...	AEIfdp	2-Mar-15	-1.0	1
5	Later this week: 3/5 9:15AM @mrubin1971 testifies before @HouseForeign on #Iran ...	AEIfdp	2-Mar-15	0.0	1
6	#Boris #Nemtsov & his eternal commitment to liberty in #Russia http://t.co/2iMmZ...	AEIfdp	2-Mar-15	0.0	2
7	Do Republicans really want to shut down #DHS & be left holding the bag for Obama'...	AEIfdp	2-Mar-15	-3.0	1
8	RT @dhume: Modi govt "budget contains virtually nothing that requires political co...	AEIfdp	2-Mar-15	1.0	1
9	MT @mrubin1971: No, Mr. #JohnKerry, There Can Be No Benefit of the Doubt on #Ir...	AEIfdp	2-Mar-15	-1.0	1
10	RT @JimTalent: American strength being "given to the locusts." http://t.co/XjmNDES...	AEIfdp	2-Mar-15	0.0	2
11	VIDEO Top three facts about US #energy production and geopolitical power, Derek Sc...	AEIfdp	2-Mar-15	0.0	1
12	#Argentina's #Kirchner reeling from scandal #Iran @rogernoriegaUSA http://t.co/wL...	AEIfdp	2-Mar-15	-2.0	1
13	5 questions every presidential candidate should answer: #Iran Edition #IranTalks @m...	AEIfdp	2-Mar-15	0.0	1
14	US opinions on the #Netanyahu visit, #Israel, and #Iran @AEIPol http://t.co/7iAzVmD...	AEIfdp	2-Mar-15	0.0	1
15	Retaking #Mosul, Thomas Donnelly @weeklystandard http://t.co/rz8fHKxoGo #Iraq ...	AEIfdp	2-Mar-15	0.0	1

Figure 5.2 - *A screenshot of the SQLite database containing tweets*

5.3.2 Standardizing and Refining the Data

After the tweets collection was completed, the data in the SQLite database was standardized for compatibility purposes. This includes formatting the dates and removing duplicate tweets. The data in the SQLite database was exported to Excel in .csv format. The dates were standardized using the following format *dd-mmm-yy*, which is one of the recognized date formats in MATLAB[20]. Excel's built-in duplicate removal function was used to remove all tweets containing identical messages. The .csv file was then uploaded into the SQLite database to replace the existing data.

[20] MATLAB will be used in *Chapter 6* for the time-series analysis.

5.3.3 Sentiment Analysis and Additional Feature Extraction

The tweets contained in the SQLite database were then analysed for sentiment strength. This process is completed using two tools: SentiStength (lexicon-based) and Stanford NLP Sentiment Analyser (machine learning-based). Both tools have publicly available Java libraries.

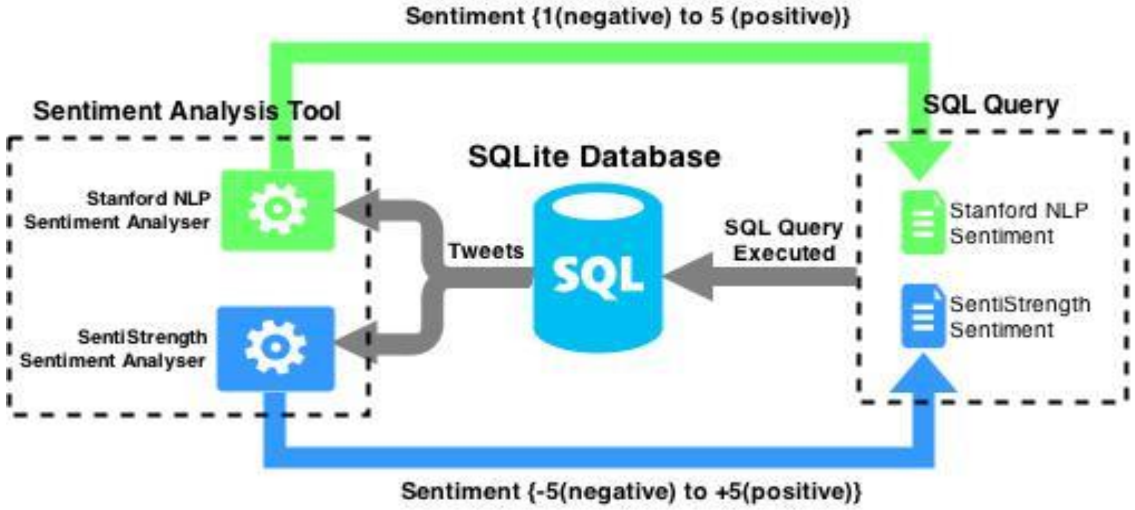

Sentiment {1(negative) to 5 (positive)}
Sentiment Analysis Tool
Stanford NLP Sentiment Analyser
SentiStrength Sentiment Analyser
SQLite Database
Tweets
SQL
SQL Query Executed
SQL Query
Stanford NLP Sentiment
SentiStrength Sentiment
Sentiment {-5(negative) to +5(positive)}

Figure 5.3 - *A block diagram showing the process used to retrieve tweets from the SQLite database and update them with a sentiment strength from both Stanford NLP and SentiStrength sentiment analysers*

In order to make the processes of sentiment analysis more efficient, a Java tool was created (See *Figure 5.3* for the block diagram). The tool retrieves the tweets from the SQLite database and determines the sentiment strength of each tweet using both SNLP and SS analysers. The output from both analysers is then incorporated into a SQL query. The SQL query is execute and updates the database with the sentiment for each tweet. The SS sentiment is stored in column "SENTI" and SNLP sentiment in column "NLP".

Literature review suggested that OPEC behaviour[21] is a key impacting factor of WTI crude oil price. One novel method of quantifying OPEC behaviour is by monitoring the frequency of references to OPEC members in USFP tweets. Refer to *Appendix C* for the list of OPEC members. Another feature that is extracted is the frequency of references to "oil". Studies have shown that often references to a particular word can indicate future values of certain products or stocks[22]. Therefore, as we are trying to forecast crude oil, a decision was made to

[21] Radchenko (2005)
[22] Moat et al. (2013)

monitor the frequency of the word “oil” in order to identify a correlation[23]. After the sentiment scores for each tweet was stored in the SQLite database, the data was exported in .csv format. Opening the .csv file in Excel, a formula was written to calculate the number of references to “oil” and OPEC members in each tweet. The As the investigation is concerned with ***weekly*** WTI crude oil prices, for each Tweet posting date/time, a formula was used to determine the week start date. All of this data (tweets, sentiment, and frequency of “oil” and OPEC members) was collated and saved in a .csv file called *featureset.csv*.

```
=COUNTIFS(A1, “oil*”)
=COUNTIFS(A1, “Algeria”, “Saudi Arabia”, “KSA”, “United Arab Emirates”, “UAE”…)
=A1-WEEKDAY(A1,2)+1
```

5.4 WTI Crude Oil Prices

The WTI crude oil prices were obtained from the United States Energy Information Administration. Refer to the *Appendix D* for list of weekly oil prices starting the week commencing 3rd of January 2011 and ending the week commencing 2nd of March 2015 (218 weeks).

5.5 Conclusion

This chapter outlined the methodology and tools (Twitter Search API) used to collect the data from Twitter. The main assumptions made during data collection were: 1) Only tweets from credible or influential Twitter accounts constitute as relevant information when forecasting crude oil 2) USFP data on Twitter is represented by the tweets made **by** US foreign policy and strategy think tanks’ Twitter accounts. 3) OC data on Twitter is represented by the tweets made **by** the largest oil companies’ and associations’ Twitter accounts. The chapter also outlined the data refining process undertaken prior to conducting the sentiment analysis. The decision to extract additional features such as frequency of “oil” and references to OPEC members was also justified. The next chapter (*Chapter 6)* will use the data collected and processed from Twitter to identify whether or not there is a correlation between these novel inputs (USFP and OC sentiment and frequency of “oil” and OPEC member references) and weekly WTI crude oil.

[23] Correlation analysis will be conducted in *Chapter 6*

Chapter 6

Time Series Analysis Study

The literature review has revealed a gap in the types of inputs WTI crude oil forecasting models have used in the past. This has motivated us to investigate various alternative and novel inputs that can potentially predict the price WTI crude oil price more accuracy. As indicated by literature, the successful use of sentiment from social media to forecast financial instruments has motived us to use sentiment as our primary feature. This, combined with the impact of factors such as, USFP and OC behaviour, have resulted in the decision to investigate the correlation between sentiment of USFP and OC and WTI crude oil prices. The relevance of OPEC behaviour to predictive accuracy, as suggested in literature[24], justifies the decision to monitor the frequency of references to OPEC members. Similarly, as mentioned in *Chapter 5*¸ studies have shown that certain word frequencies correlate with the rise and fall of financial instrument prices. This justifies the decision to investigate the correlation between the frequency of "oil" and WTI crude oil price. Using these novel inputs, a comprehensive time-series analysis was conducted. The purpose of this chapter is to provide a detailed overview of the Time Series Analysis study which attempts to statistically identify a correlation between weekly WTI crude oil prices and novel inputs defined above.

6.1 Aim

The aim of this study is to determine whether the sentiment of **USFP** and **OC** and **frequency of references to "oil"** and **OPEC** members in USFP tweets provide **statistically** significant information to forecast weekly **WTI** crude oil prices. By doing so, we can determine the optimal inputs that can be used to build the predictive model in *Chapter 7*. The features that do not provide any statistical relationship with WTI crude oil will not be used in the forecasting model. The study uses weekly crude oil prices as opposed to daily as long term horizon forecast is more useful in industrial application than short term. Given that the data set is limited to approximately 4 years, monthly prices would reduce our sample size significantly thereby potentially preventing us from obtaining accurate and reliable results.

[24] Radchenko (2005)

6.2 Hypothesis

There are four main **null** hypotheses being tested in this experiment:

- USFP sentiment is ***not*** statistically significant in forecasting weekly WTI crude oil prices
- OC sentiment is ***not*** statistically significant in forecasting weekly WTI crude oil prices
- Frequency of references to “oil” is ***not*** statistically useful in forecasting weekly WTI crude oil prices
- Frequency of references to OPEC members is ***not*** statistically useful in forecasting weekly WTI crude oil prices
- Justin Bieber’s sentiment is ***not*** statically significant in forecasting weekly WTI crude oil prices. **Control Variable –** we know for certain the Bieber’s sentiment should not have any impact on WTI crude oil prices. Therefore the robustness of our methodology can be tested by using this variable**.**

6.3 Methods

The first step to test the significant of each features was to import the *featureset.csv*[25] file into MATLAB as a table using the “Import Data” option. Each of the features (USFP sentiment, OC sentiment, “oil” frequency, OPEC frequency) were then processed in MATLAB to obtain the weekly mean or sum *(depending on the feature)* using the commands below. Sum was obtained for “oil” and OPEC frequency, while the mean was calculated for USFP and OC sentiment. This command was executed until the weekly means and sums for all the features had been calculated. A similar command was used to extract additional features from the data such as *standard deviation* and *variance*.

```
varfun(@mean,tbl_name,'InputVariables','SENTI','GroupingVariables',WEEK’)
varfun(@sum,tbl_name,'InputVariables','OILFREQ','GroupingVariables',WEEK’)
```

```
varfun(@std,tbl_name,'InputVariables','SENTI','GroupingVariables',WEEK’)
varfun(@var,tbl_name,'InputVariables','SENTI','GroupingVariables',WEEK’)
```

[25] *Featureset.csv* is a file containing all the extracted features from Twitter. These include {USFP sentiment (SS and NLP), OC sentiment (SS and NLP), frequency of “oil”, frequency of OPEC members}.

The values returned by the commands above were added to the *featureset.csv* file. See *Appendix E* for a screenshot of the *featureset.csv* file. The updated *featureset.csv* file was re-uploaded to MATLAB. Each column in the *featureset.csv* file were stored as a column vector in MATLAB. *Table 6.1* contains a list of the column vectors, contents description, and value type.

***Table 6.1** - Description of MATLAB column vectors' and value type*

Column Vector	Contents	Value size and type
usfp_senti	United States foreign policy sentiment (SentiStrength)	218x1 double
usfp_nlp	United States foreign policy sentiment (Stanford NLP)	218x1 double
usfp_oilfreq	frequency of the term “oil” in United States foreign policy tweets	218x1 double
usfp_opecfreq	frequency of references to OPEC members in United States foreign policy tweets	218x1 double
oil_senti	oil companies’ sentiment (SentiStrength)	218x1 double
oil_nlp	oil companies’ sentiment (Stanford NLP)	218x1 double
jb_senti	Justin Bieber sentiment (SentiStrength	66x1 double
jb_nlp	Justin Bieber sentiment (Stanford NLP)	66x1 double
jb_date	weeks corresponding to the tweets available from Justin Bieber’s twitter account	66x1 cell
jb_oilprice	weekly WTI crude oil prices corresponding to the weeks from jb_date	66x1 double
oil_price	weekly WTI crude from 3rd January 2011 to 2nd March 2015	218x1 double
dates	start of week dates between 3rd January 2011 and 2nd March 2015	218x1 cell

After the columns have been uploaded to MATLAB, the statistical analysis can begin. In order to determine the significant of each feature, we will be using a Granger-Causality. Granger-Causality is a statistical hypothesis test that uses a bivariate linear regression method to model a stochastic process. The test uses two time-series, *X* and *Y*. Time series *X* is said to “Granger-cause” time series *Y* if time-series *Y* can be better predicted using historical values of both time series *X* and *Y* as opposed to the historical values of Y in isolation.

The Granger-Causality Test can be defined by the following equation:

$$Y_t = \alpha + \sum_{i=1}^{n} \beta_{1t-i} Y_{t-i} + \sum_{i=1}^{n} \beta_{2t-i} Y_{t-i} + \varepsilon_t$$

$$X_t = \alpha + \sum_{i=1}^{n} \beta_{3t-i} X_{t-i} + \sum_{i=1}^{n} \beta_{4t-i} Y_{t-i} + \varepsilon_t$$

The variable *n* represents the maximum number of lags to be considered in the model. The matrix *β* represents the coefficients of the model and ε_t is the residual value of the time series (prediction error). An f-test is used on the statistical model to determine whether or not time-series *X* "Granger-causes" times series *Y*. If the **f-statistic** outputted from the test is greater than the **critical value** then the **null hypothesis** can be rejected. If the null hypothesis is rejected then that means the feature is statistically significant and will be incorporated into the predictive model in *Chapter 7*.

Two main assumptions can be made about the Granger-Causality Test and the relationship between time-series *X* and time-series *Y:*

- The cause happens prior to the effect – i.e. if X is causing Y, then X must be occurring before Y
- The cause has unique information about the future values of its effect – i.e. if X causes Y then X has unique information about the future values of Y.

The Granger-Causality test was chosen due to its wide application in forecasting oil prices[26] as well as its computational simplicity. However, the Granger-Causality is not a built in function in MATLAB, as a result an external .m[27] was downloaded from MATLAB Central website. Refer to *Appendix F* to the see the MATLAB code. The Granger-Causality test was applied to weekly WTI crude oil prices and each of the features in *Table 4.1* that have a value

[26] Ye et al. (2006); Gillman and Nakov (2009); Kilian and Murphy (2010);

[27] MATLAB file format

type of *double* using the following formula where *y* was replaced with a feature column vector, alpha was 0.05 and lag values between 1 and 10 were used (1 lag = 1 week).

```
[F,c_v] = granger_cause(oil_price,y,alpha,max_lag)
```

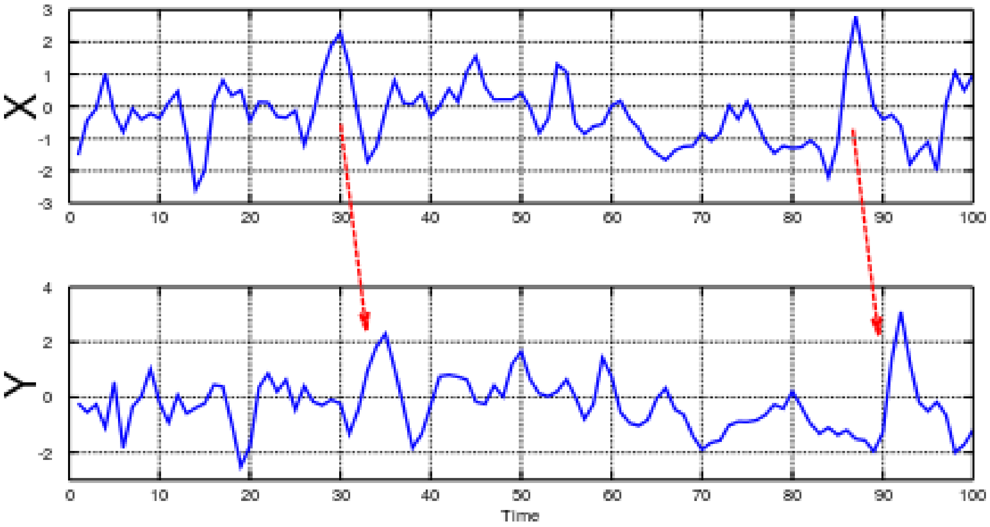


Figure 6.1 - *A graph showing time-series X causing time-series Y. This graph suggests that the patterns between X and Y are repeated after a brief lag. Thus past values of X can be used to predict future values of Y.*

6.4 Results

The results of the Granger-Causality Test have been recorded in *Table 6.2* through to *6.9*. The graphical representations of each of the attributes mentioned in *Table 6.1* can be seen in Figures *6.2 to 6.7*. The tables contain three columns: **lag**, **f-statistic**, and **critical value**. The lag is the number of weeks that the feature takes to have an impact on the price of weekly WTI crude oil. The f-statistic is the output of the f-test. If the value of the f-statistic is larger than the critical value, then at the specified lag, the null hypothesis can be rejected[28].

[28] If the null hypothesis is rejected, then the feature is statically significant and can be used as an input in the predictive model.

Table 6.2 - *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and USFP as per* ***SS*** *analyser (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	**3.9817**	3.8851
2	**4.1303**	3.8853
3	**4.1303**	3.8853
4	**4.1303**	3.8853
5	**4.1303**	3.8853
6	**7.8631***	2.1422
7	**7.8631***	2.1422
8	**4.5240**	3.8866
9	**4.5240**	3.8866
10	**4.5240**	3.8866

The results of the Granger-Causality Test using USFP sentiment from the SS analyser in *Table 6.2* showed that the f-statistic was greater than the critical value at lags of 1-10. Therefore, we can reject the null hypothesis that USFP sentiment as per SS analyser does not provide statically significant information in forecasting weekly WTI crude oil price. The rejection of the null hypothesis indicates that USFP as per SS "Granger-causes" the future value of WTI crude oil and as result would be an acceptable input for the forecasting model in *Chapter 7*. Table *6.2* also shows that lag 6 and 7 have the largest f-statistic of 7.2927 suggesting that the impact of USFP sentiment likely takes 6 to 7 weeks to work through the weekly price of WTI crude oil. Comparing the results of *Table 6.2* with the graph in *Figure 6.2* we can see a clear visual correlation with the weekly price of WTI crude oil and the USFP sentiment as per SS analyser. The weekly price of WTI crude oil and USFP sentiment remain relatively constant until around October 2014, when the price of crude oil begins to drop along with the sentiment. Towards the middle of February 2015, the sentiment of USFP begins to pick up and so does the price of WTI crude oil. This suggests a positive correlation between the two variables.

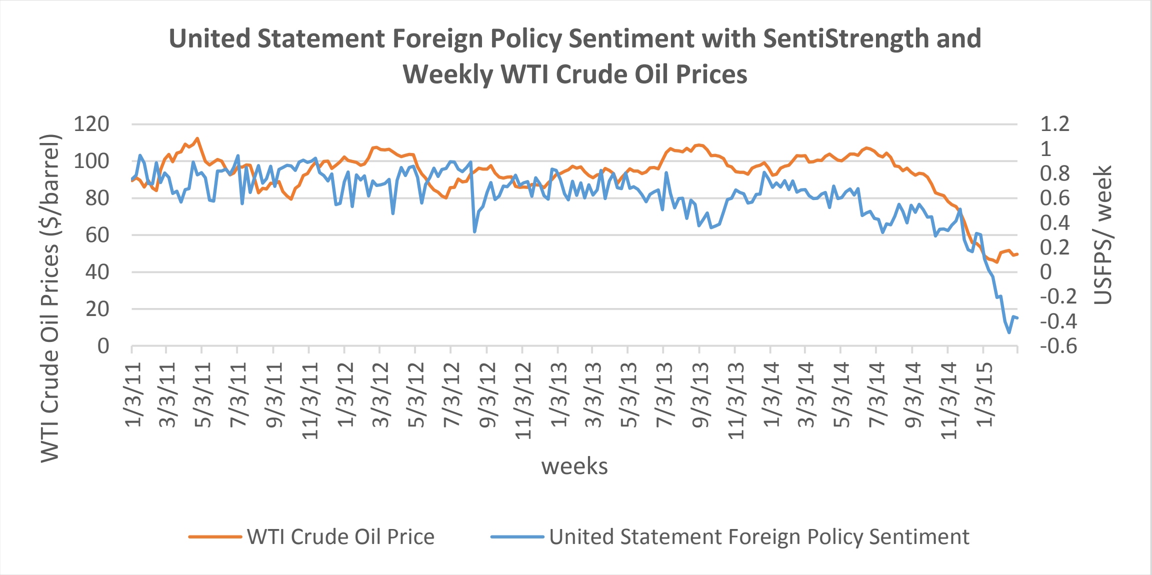


Figure 6.2 - *A graph showing the relationship between weekly WTI crude oil price and United States foreign policy sentiment as per SentiStrength analyser*

Granger-Causality Test on USFP sentiment from the SNLP analyser in *Table 6.3* revealed that at lag 6 and lag 7, the f-statistic is 5.9497 and 5.9603, respectively. These values are greater than the critical value and therefore provide statistically significant information to forecast weekly WTI crude oil price. However, the USFP sentiment is not statistically significant at lag of 1-5 and 8-10. Therefore, the null hypothesis can only be rejected at lags of 6 and 7. This finding suggests that USFP as per SNLP is only a statistically significant input for the forecasting model at lags of 6 and 7. *Figure 6.3*, unlike *Figure 6.2*, does not provide a visual correlation between USFP sentiment and weekly WTI crude oil price. However, a visual correlation does not necessarily suggest that the data is not statistically significant.

Table 6.3- *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and USFP as per **SNLP** analyser (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	0.7176	3.8851
2	1.2493	3.8853
3	1.2493	3.8853
4	1.2493	3.8853
5	1.2493	3.8853
6	**5.9497**	2.1422
7	**5.9603***	2.0538
8	0.6209	3.8866
9	0.6209	3.8866
10	0.6209	3.8866

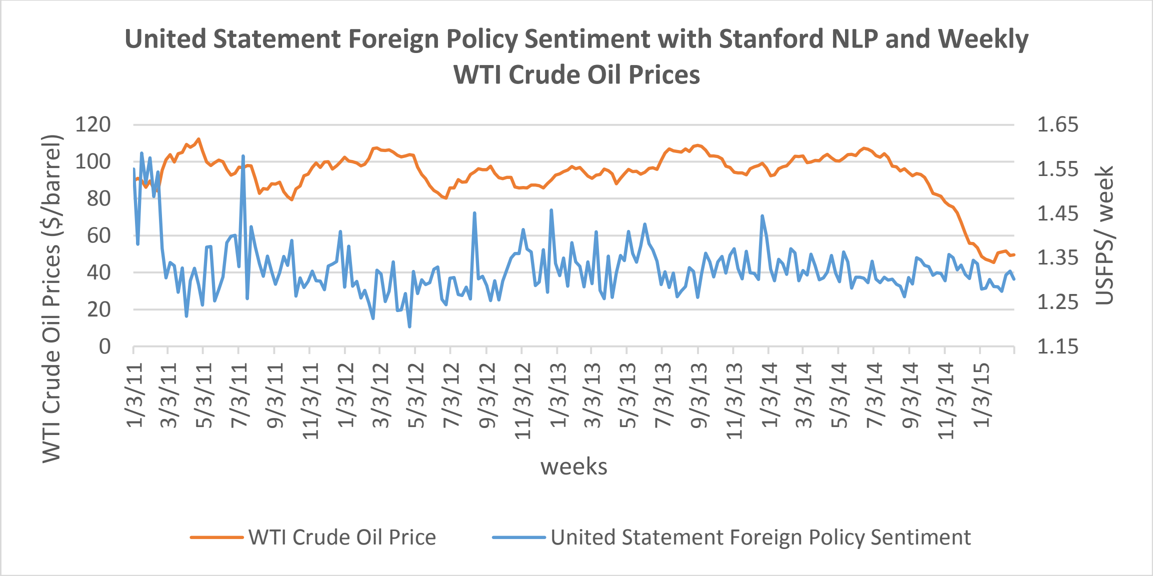


Figure 6.3- *A graph showing the relationship between weekly WTI crude oil price and United States foreign policy sentiment as per Stanford NLP analyser*

Table 6.4- *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and the frequency of the term "oil" in USFP tweets (alpha 0.05; bold = significant; * = highest f-statistic)*

Lag (n)	f-statistic	critical value
1	**6.0291**	3.8851
2	**5.7702**	3.8853
3	**5.7702**	3.8853
4	**5.7702**	3.8853
5	**5.7702**	3.8853
6	**7.2927***	2.1422
7	**7.2927***	2.1422
8	**5.3300**	3.8866
9	**5.3300**	3.8866
10	**5.3300**	3.8866

The results of the Granger-Causality Test in *Table 6.4* show that the f-statistic for lags 1- 10 are greater than the critical value. This allows us to reject the null hypothesis which states that the frequency of "oil" in USFP tweets do not provide statistically significant information for forecasting weekly WTI crude oil prices. As the f-statistic is greater than the critical value at all lags, we can conclude that frequency of "oil" is a statistically significant input for forecasting crude oil at all lags. However, similar to *Table 6.2,* in *Table 6.4* the f-statistic at lags 6 and 7 is the largest (f-statistic: 7.2927). This suggests that the predictive power of this feature is strongest at lag of 6 or 7. The graph of frequency of "oil" and weekly WTI crude oil prices in *Figure 6.4* provides a clear indication of the relationship between the two variables. As the frequency of "oil" increases (blue line), the price of WTI crude oil decreases (orange line). Thus, based on *Figure 6.4* we can conclude that the frequency of "oil" and weekly WTI crude oil price have a negative correlation.

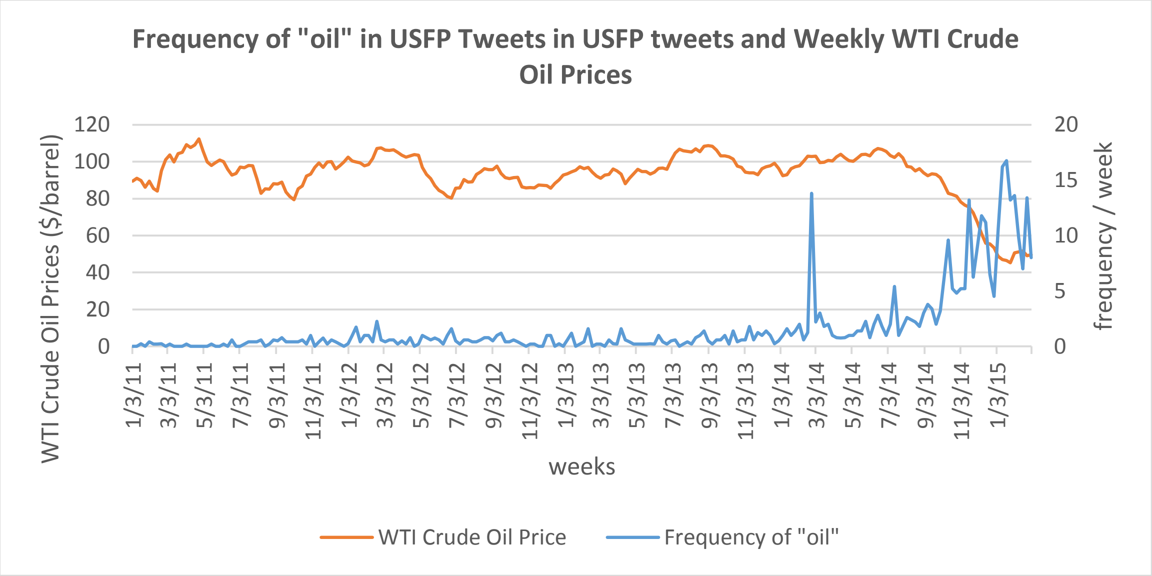


Figure 6.4- *A graph showing the relationship between weekly WTI crude oil price and the frequency of references to "oil" in USFP tweets*

Table 6.5 - *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and the frequency of the OPEC member references in USFP tweets (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	**6.0132**	3.8851
2	**6.6712**	3.8853
3	**6.6712**	3.8853
4	**6.6712**	3.8853
5	**6.6712**	3.8853
6	**7.7699***	2.1422
7	**7.7699***	2.1422
8	**5.6853**	3.8866
9	**5.6853**	3.8866
10	**5.6853**	3.8866

Similar results were observed in *Table 6.5* where the f-statistic for lags 1-10 are all greater than the critical value. Therefore, we can reject the null hypothesis that the frequency of OPEC members is not does not provide statistically significant information in forecasting weekly WTI crude oil prices. This indicates that OPEC member references is a statistically significant input for the predictive model. However, the f-statistic is most significant at lags of 6 and 7 (f-statistic: 7.7699). Therefore, once again, the impact of OPEC member references take 6 to 7 weeks to work through the price of WTI crude oil. The graph in *Figure 6.5* shows that the frequency of references to OPEC members also has a negative correlation with weekly WTI crude oil prices. As the frequency of OPEC member reference increases (blue), the price of WTI crude oil decreases (orange).

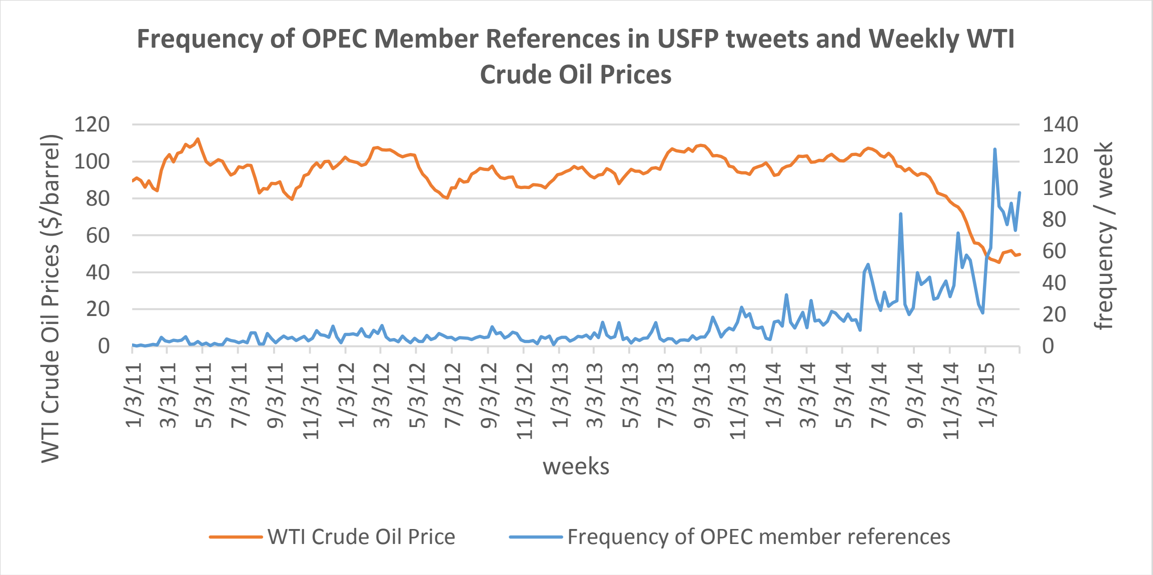


Figure 6.5- *A graph showing the relationship between weekly WTI crude oil price and the frequency of references to OPEC members in USFP tweets*

Table 6.6 shows the output of the Granger-Causality Test between OC sentiment as per SS analyser and weekly WTI crude oil price. The test reveals that OC sentiment is only statistically significant input in forecast weekly WTI crude oil price at lags of 6 and 7 (f-statistic: 6.2269). Therefore, the null hypothesis can be rejected for lags of 6 and 7. The graph in *Figure 6.6* reveals somewhat similar movements between the OC sentiment and WTI crude oil price. Particularly during October 2014, where both the OC sentiment and the price of WTI crude oil fall steeply and February where both variables begin to increase.

Table 6.6 - Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and OC sentiment as per SS analyser (alpha 0.05; bold = significant; *highest f-statistic)

Lag (n)	f-statistic	critical value
1	0.3918	3.8851
2	0.4342	3.8853
3	0.4342	3.8853
4	0.4342	3.8853
5	0.4342	3.8853
6	**6.2268***	2.1422
7	**6.2268***	2.1422
8	0.8022	3.8866
9	0.8022	3.8866
10	0.8022	3.8866

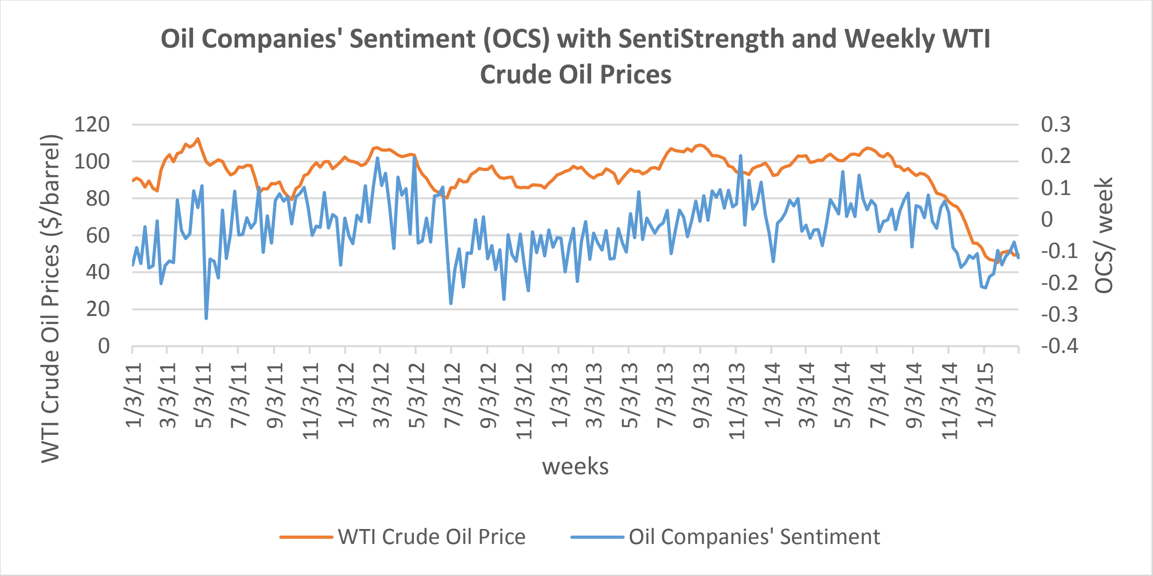


Figure 6.6 - *A graph showing the relationship between weekly WTI crude oil price and oil companies' sentiment as per SentiStrength analyser*

The data from *Table 6.7* suggests at lags of 2 to 7 we can reject the null hypothesis which states that OC sentiment as per SNLP does not provide statistically significant information in forecast weekly WTI crude oil prices. The f-statistic is largest for lags of 6 and 7 with a value of 8.0339. This suggests that OC sentiment as per SNLP is most significant in forecasting WTI crude oil prices at lags of 6 and 7. However, analysing the graph in *Figure 6.7,* no distinct visual correlation is not present.

Table 6.7- *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and OC sentiment as per SNLP analyser (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	0.4005	3.8851
2	**5.9510**	3.0383
3	**5.9510**	3.0383
4	**5.9510**	3.0383
5	**5.9510**	3.0383
6	**8.0339***	2.1422
7	**8.0339***	2.1422
8	0.8408	3.8866
9	0.8408	3.8866
10	0.8408	3.8866

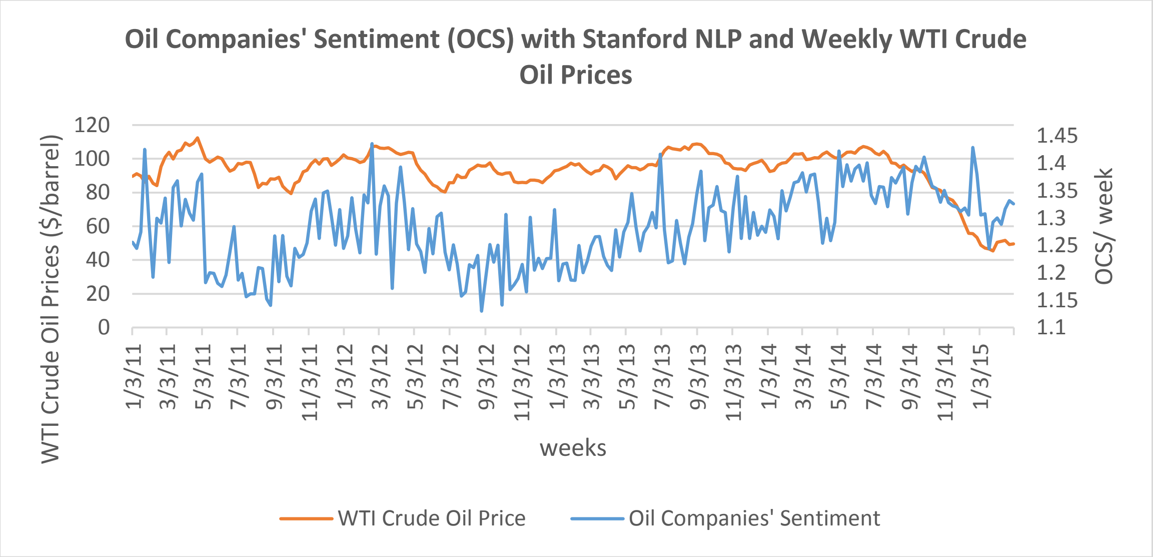


Figure 6.7- *A graph showing the relationship between weekly WTI crude oil price and oil companies' sentiment as per Stanford NLP analyser*

Tables 6.2 through to *6.7* suggest that each feature tested using Granger-Causality Test is statistically significant in forecasting WTI crude oil prices at lags of 6 and 7. Although, some features were statistically significant in other lag values as well, the lags of 6 and 7 consistently returned the largest f-statistic. There is one exception to this observation. In *Table 4.3* the f-statistic of lag 7 (5.9603) is slightly greater than lag 6 (5.9497) suggesting that 7 is the optimal number of lags between weekly WTI crude oil and all the features used in this investigation.

In order to verify the robustness of the data we had to select a variable that for certain would not have any impact on the weekly WTI crude oil prices or a control variable. The variable chosen for this test was Justin Bieber's sentiment on Twitter. The f-statistic values outputted from the Granger-Causality Test performed on Bieber's sentiment, determined by both SS and SNLP sentiment analysers, revealed no correlation between sentiment and weekly WTI crude oil price. Therefore the null hypothesis cannot be rejected and thus supporting the credibility of the data.

Table 6.9 - *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and Justin Bieber's sentiment as per SS analyser (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	0.5451	3.9934
2	0.4743	3.9959
3	0.4743	3.9959
4	0.4743	3.9959
5	0.4743	3.9959
6	0.4743	3.9959
7	0.4743	3.9959
8	0.4743	3.9959
9	0.4743	3.9959
10	0.4743	3.9959

Table 6.8 - *Granger-Causality Test – statistical significance (f-statistic) at lags of 1 – 10 weeks between weekly WTI crude oil price and Justin Bieber's sentiment as per SNLP analyser (alpha 0.05; bold = significant; *highest f-statistic)*

Lag (n)	f-statistic	critical value
1	0.8603	3.9934
2	1.4933	3.9959
3	1.4933	3.9959
4	1.4933	3.9959
5	1.4933	3.9959
6	1.4933	3.9959
7	1.4933	3.9959
8	1.4933	3.9959
9	1.4933	3.9959
10	1.4933	3.9959

6.5 Discussion

This study was conducted to determine the statistical significance of the novel inputs' ability to forecast weekly WTI crude oil prices. The results of this study revealed several findings. The first finding is the direct correlation between USFP as per SS and weekly WTI crude oil price. This suggests that as the sentiment of USFP decreases so does the oil price. The fall in USFP sentiment can be attributed to various factors including economic or financial decline. The decline of the global economy can result in reduced consumption and demand and excess of supply. These circumstances will drive the price of oil down to reach equilibrium.

The second finding of the study is that there is an inverse correlation between the frequency of "oil" and OPEC references and WTI crude oil. The third finding of the study is that all the attributes tested in this investigation provide some statistically significant or "unique" information about the future values of WTI crude oil price at the lag of 7. As a result, all of the attributes are statistically significant inputs for the predictive model in *Chapter 7* at a lag of 7. The lag of 7 also indicates that the overall time that a change in feature's value takes to impact WTI crude oil price is 7 weeks. This is in line with the theoretical nature of demand and supply of oil. For example, a threat of a sanction may not immediately impact the supply of oil and thus the price of oil. A study conducted by GlobalPetrolPrices.com and Kimberley Amadeo, a US economy expert, identified a 7 week lag between the change of WTI crude oil price and the price of gasoline.

6.6 Conclusion

Overall study produced several results that provide an essential foundation needed in the next experiment. The main conclusion we can draw from this study is that all of the features used in the study "Granger cause" weekly WTI crude oil prices at the lag of 6 and 7. Therefore we can reject all **four** null hypothesis of the experiment. We can also conclude that the lag of 7 is the optimal lag for all features i.e. it takes 7 weeks for the features to have a full impact on the weekly price of WTI crude oil. Therefore, in *Chapter 7,* the Predictive Modelling study, all of the features tested during this experiment will be included as one of the input for the forecasting model of weekly WTI crude oil prices. Furthermore, to support our findings regarding the lag of 7, we will build a test model using the lag of 1 and evaluate its predictive capability.

Chapter 7

Predictive Modelling Study

The results of the study in *Chapter 6* suggested that USFP and OC sentiment and the frequency of the references to “oil” and OPEC members all have a statistically significant correlation with weekly prices of WTI crude oil, thus making them statistically significant inputs for the WTI crude oil forecasting model. Furthermore, based on the finding regardin the lag of 7, the predictive model will be built to forecast the price oil 7-weeks ahead. Literature suggests that non-linear models, particularly ANNs and SVM models[29] possess greater forecasting abilities than linear models. As a result, throughout this chapter, we will outline the process by which the statistically significant inputs of *Chapter 6* can be used to build a forecasting model to effectively predict the directional shift of weekly WTI crude oil prices.

7.1 Aim

The aim of this study is to **build** and **evaluate** *three* machine learning forecasting models to classify the direction of weekly WTI crude oil prices using the various statistically significant Twitter extracted features tested in *Chapter 6*.

7.2 Hypothesis

The hypothesis for this experiment is that the direction of weekly WTI crude oil prices **can** be *effectively* (better than 50-50 coin toss) predicted using **machine learning techniques** and modelling weekly WTI crude oil price as a function of USFP and OC sentiment and frequency of “oil” and OPEC member references.

[29] Fernandez (2010)

7.3 Methods

In order to understand the methodology used to build the predictive model is important to clearly define what the predictive model is trying to classify. The model is trying to solve a **binary classification problem** or a problem with two possible outcomes. The problem can be defined as follows:

- *What will the directional shift of the weekly WTI crude oil price in* ***seven*** *weeks?*
    - ***Increase*** *or* ***Decrease?***

As mentioned in *Chapter 6*, a lag of 7 provides the most statistically significant results in the Granger Causality Test. Therefore, the model will be built to predict the price of weekly WTI crude oil price for *t+7* or 7-weeks ahead.

For the purposes of this investigation, the machine learning models will be built using Weka 3.6. Weka is an open source software developed at the University of Waikato that contains functions for a number of machine learning algorithms and data mining tasks. It allows users to pre-process their data and contains tools for classification, regression, clustering, association rules, and visualizations. Weka was chosen due to its relatively simple interface, comprehensive set of tools for users to experiment with.

Before building the model, we must label our feature set in accordance with the two classes defined above, "*Increase*" or "*Decrease*". This will be done in the *featureset.csv* file created previously. An additional column titled "Classification" was then created. As the classification of instance at time *t* is based on the price of weekly WTI crude oil at *t+7* the following excel formula was used:

```
=IF(B2[WTI crude oil price @ t]<B9[WTI crude oil price @ t+7], "Decrease", "Increase")
```

This process was completed for the first 211 out of the 218 instances, as the last 7 instances do not have future oil prices to compare against. The fully labelled of the feature set data will represent **SET 1**. 70 instances (approximately half from *Increase* class and the other half from *Decrease*) were then randomly sampled and stored in a separate .csv file. This file will represent **SET 2**. The instances sampled and stored in SET 2 were then removed from SET 1, resulting a total of 141 instances in SET 1 and 70 in SET 2. The purpose of the creation was two sets is to ensure a robust testing procedure for the machine learning models. SET 2 will remain unseen and will only be used to test the model once the parameters have been optimized. SET 1 will be used for the training a d testing purposes. Both SET 1 and SET 2 are .csv files (Weka compatible format for training and testing data).

SET 1 was then opened in Weka using the “Open File” function. Once successfully opened, the number of instances and attributes appeared. For SET 1 there are 141 instances and 20 attributes.

Table 7.1 - *A list of attributes in the feature set with description*

Attribute	Description
oil_price	Weekly WTI Oil Price
usfp_senti	United State foreign policy sentiment as per SS
usfp_senti_var	Variance of United States foreign policy sentiment as per SS
usfp_senti_std	Standard deviation of United states foreign policy sentiment as per SS
usfp_nlp	United State foreign policy sentiment as per SNLP
usfp_nlp_var	Variance of United States foreign policy sentiment as per SNLP
usfp_nlp_std	Standard deviation of United states foreign policy sentiment as per SNLP
usfp_oil freq	Frequency of "oil" in United States foreign policy tweets
usfp_opec freq	Frequency of OPEC member references in United States foreign policy tweets
usfp_oil_var	Variance of frequency of "oil" in United States foreign policy tweets
usfp_opec_var	Variance of frequency of OPEC member references in United States foreign policy tweets
usfp_oil_std	Standard deviation of frequency of "oil" in United States foreign policy tweets
usfp_opec_std	Standard deviation of Frequency of OPEC member references in United States foreign policy tweets
oil_senti	Oil companies' sentiment as per SS
oil_nlp	Oil companies' sentiment as per SNLP
oil_senti_var	Variance of oil companies' sentiment as per SS
oil_sent_std	Standard deviation of oil companies' sentiment as per SS
oil_nlp_var	Variance of oil companies' sentiment as per SNLP
oil_nlp_std	Standard deviation of oil companies' sentiment as per SNLP
Classifcation{Increase, Decrease)	Class

Now, that we had uploaded the training set into Weka were able to begin the building of the machine learning model. In order to build the machine learning models, the classification tool box in Weka was opened. The three different classifiers that will be used for this investigation are Support Vector Machines (SVM), Multilayer Peceptron (MLP), and Naïve Bayes (NB). SET 1 will be used to train, test, and optimize each of these classifiers. The methodology used for testing in this study is a **10-fold cross validation** method. SET 1 will be divided into 10 equal parts. 9 parts will be used for training and the remaining part will be used for testing. This process is done for each of the 10 parts in SET 1. The average performance of all 10 folds is the performance of the model. This will serve as an indicator of whether the parameters need to be updated. The aim of optimizing the parameters is to ensure highest possible accuracy. Once the parameters of the model have been optimized in accordance with SET 1, an unseen data set, SET 2, will be used to the evaluate the model classification accuracy. The accuracy obtained on the unseen test set, SET 2, will be the true accuracy of the classification model. *Figure 7.1* shows a visual representation of how the sets will be divided, trained and tested.

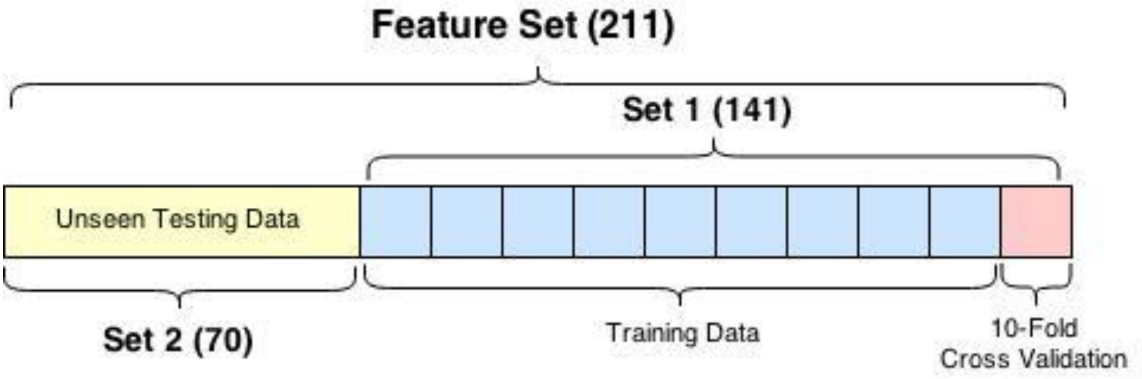


Figure 7.1 – A visual representation of the feature set. Set 1 will be used for parameter optimization through training and 10-fold cross validation. Set 2 is an unseen test that on which the model will be evaluated on.

Support Vector Machine

In Weka, there are several configurations of Support Vector Machine classifiers available. The one used for this investigation is Sequential minimal optimization (SMO). The classifier is only suited for binary classification problems. The SMO model requires the configuration of the parameters listed in *Table 7.2*. The model was evaluated and optimized according to these parameters until the maximum classification accuracy was achieved. Please see *Section 7.4* for the optimal parameters and classification accuracy.

Table 7.2 - *A list of parameters and configurations that need optimized for the SVM model to perform accurately*

Parameter/Configurations	Function	Value Type
buildLogisticModels	Fit logistic model into output (for proper probability estimates)	True/False
C parameter	Large values of C, the optimization will choose a smaller-margin between the support vectors to ensure more training points get classified correctly. Small values of C, the optimization will look for a large margin between the support vectors even if it means a misclassification of more training points.	R (real number)
Kernel	Kernel function or "kernel trick" is a mathematical equation that is used to transform the data in a higher dimension where a hyperplane is able to linearly separate the two classes.	-PolyKernel -Puk (Pearson VII) -RBFKernel -StringKernel

Naïve Bayes

Weka offers various types of Naïve Bayes classifiers in their classification tool box. For this investigation, the "NaiveBayes (NB)" classifier will be used. The NB classifier relies on class estimator probabilities obtained from the training data to classify the instances in the test data. Weka allows for the use of supervised discretization of the data when training and testing the classifier. Supervised discretization is the process of converting continuous or numerical features to nominal ones. This is usually done through the Fayyad & Irani's MDL method which uses mutual information to recursively define the best bins.

Table 7.3 - *A list of parameters and configurations that need optimized for the NB model to perform accurately*

Parameter/Configurations	Function	Value Type
useSupervisedDiscretization	Process of converting numerical values to nominal values	True/False

Multilayer Perceptrons (MLP)

There are a number of variations in the type of MLP models available in the classification tool box. For this investigation, we will use the “Multilayer Perceptron (MLP)” model. The MLP models node output values are determined by a sigmoid activation function as mentioned in *Chapter 3*. In order to optimize the MLP model a number of configuration and parameters need to be adjusted. *Table 7.4* contains a list of parameter and configurations required by the model. The optimal settings for these parameters are available in *Section 7.4.*

Table 7.4 - *A list of parameters and configurations that need optimized for the MLP model to perform accurately*

Parameter/Configurations	Function	Value Type
hiddenLayers/units	Defines the number of hidden layers in the neural network. Traditionally 1 hidden layer is used for feature sets that are linearly separable in a two-dimensional space. 2 hidden layers are used for feature sets that operate in a higher-dimension to be separated by a hyperplane.	#units in first hidden layer, # units is second hidden layer…,# of units in *n* hidden layer
learningRate	Controls the size of the weights and bias changes during the learning. Increase the learning rate applies a greater portion of the respective adjustment to the old weight. The model, as a result, will learn quicker, but if there is significant variability in the attribute values then it may not learn well.	$0 \leq R < 1$
Momentum	Momentum adds an additional factor the determination of the amount that the weights are adjusted. By increasing momentum, we can add small amounts of the previous weight adjustment to the current weight adjustment. It can improve learning rate in some conditions. It also allows for the smoothing of unusual conditions in the training set. High values risk overshooting the local minima while low values risk slow the system.	$0 \leq R < 1$

7.4 Results

After using SET 1 to train and the data and optimize the model parameters, the classifier was test with SET 2. The output of the results are contained within *Table 7.5*.

Table 7.5 - *A table showing a detailed accuracy of classifier performance by class {CA = classification accuracy, ROC = ROC Area, 1 lag = 1 week, *best performing classifier, bold =lag with highest accuracy}*

	Lag 1				**Lag 7**			
Classifiers	CA (%)	ROC	Precision	Recall	CA (%)	ROC	Precision	Recall
SVM*	55.71	0.52	0.56	0.56	74.29	0.74	0.75	0.74
Increase	50.00	0.52	0.52	0.50	83.80	0.74	0.72	0.78
Decrease	60.50	0.52	0.59	0.61	63.60	0.74	0.78	0.70
Naïve Bayes	52.86	0.54	0.55	0.53	67.14	0.69	0.78	0.97
Increase	68.80	0.54	0.49	0.69	40.50	0.69	0.94	0.41
Decrease	39.50	0.54	0.60	0.40	97.00	0.69	0.59	0.67
MLP	44.29	0.44	0.44	0.44	61.43	0.72	0.61	0.62
Increase	84.40	0.44	0.44	0.84	67.60	0.72	0.63	0.68
Decrease	10.50	0.44	0.44	0.11	54.50	0.72	0.60	0.55

The results in *Table 7.5* show that the lag of 7 performs significantly better in all classifiers. This supports the results from the previous study in *Chapter 6* where we concluded that lag of 7 is the amount of time that the features take to impact the WTI crude oil price. The results also reveal the SVM is the best classifier with accuracy of 74.29%. Naïve Bayes and MLP have achieved overall accuracies of 67.14% and 61.43%. These classification accuracies (CA) are based on an unseen data set with 70 instances (SET 2). SVM performed slightly better in classifying *Increase* instances than *Decrease* instances with 83.80% and 63.60% accuracy respectively. In contrast, Naïve Bayes observed a steep gap between its *Increase* CA and *Decease* CA with values of 40.50% and 97.00% respectively. MLP model was better at classifying *Increase* instances.

The precision and recall of SVM model in the lag of 7 reveal that it strikes an adequate balance between the number of correct instances and the number of instances classified for both *Increase* and *Decrease* class (Increase {P = 0.72, R = 0.78}, Decrease {P = 0.78, R = 0.70}). In contract, Naïve Bayes achieves precision of 0.94 for the *Increase* class but with only 0.41 recall, suggesting that it classifies a large number of *Decrease* instances as *Increase.*

$$precision = \frac{true\ positive}{true\ postive + false\ positive}$$

$$recall = \frac{true\ positive}{true\ postive + false\ negative}$$

The results achieved in *Table 7.5* were a result of parameter optimization using SET 1 as a training and evaluation data set. *Table 7.6* to *7.8* list the optimized parameters used for the building the classification models. *Table 7.6* is a list of optimized parameters for SVM. The option to build a logistical model was set to *false*. However, in order to obtain estimated class probabilities for each instance, *buildLogisticmodels* can be changed to true. The data represented in *Figure 7.3* and *7.4* are based on an SVM model where *buildLogisticModels* is set to true (refer to the note in *Figure 7.3 and 7.4*). The c parameter was set to 1.0 and the optimal kernel used to build this model was the Puk kernel.

Table 7.6 - *A table listing the optimized parameters for the SVM model*

Parameter/Configurations	Optimized Value
buildLogisticModels	False
C parameter	1.0
Kernel	Puk Kernel

The Naïve Bayes classifier was modelled using supervised discretization. This is where the numerical attribute values are converted to nominal values using a process referred to as binning. Selecting supervised discretization improved classification accuracy by almost 12%. However the classification still remained below SVM with 67.14%.

Table 7.7 - *A table listing the optimized parameters for the NB model*

Parameter/Configurations	Optimized Value
useSupervisedDiscretization	True

The MLP model achieved the lowest classification accuracy out of all three classifiers with a classification accuracy of 61.43%. The neural network model was built using two hidden layers with 9 units in the first hidden layer and 3 units in the second hidden layer. The learning rate was set at 2.0 and the momentum was configured at 1.0. *Figure 7.2* provide a graphical representation of the MLP model showing the number of input units, hidden layers and units, and output units.

Table 7.8 - *A table listing the optimized parameters for the MLP model*

Parameter/Configurations	Optimized Value
hiddenLayers/units	9,3
learningRate	2.0
Momentum	1.0

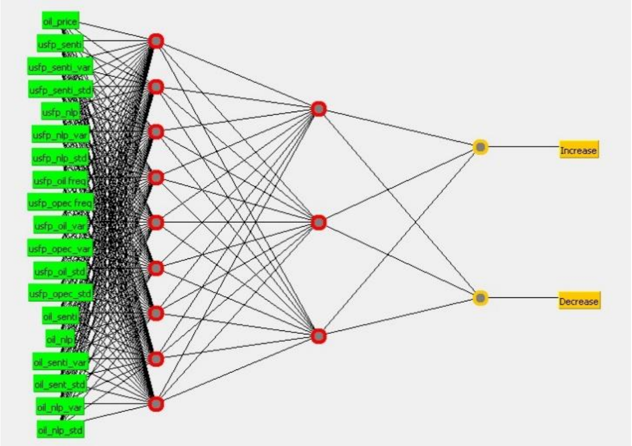


Figure 7.2 - A visualization of the inputs, number of hidden layers/units, and outputs of the MLP model

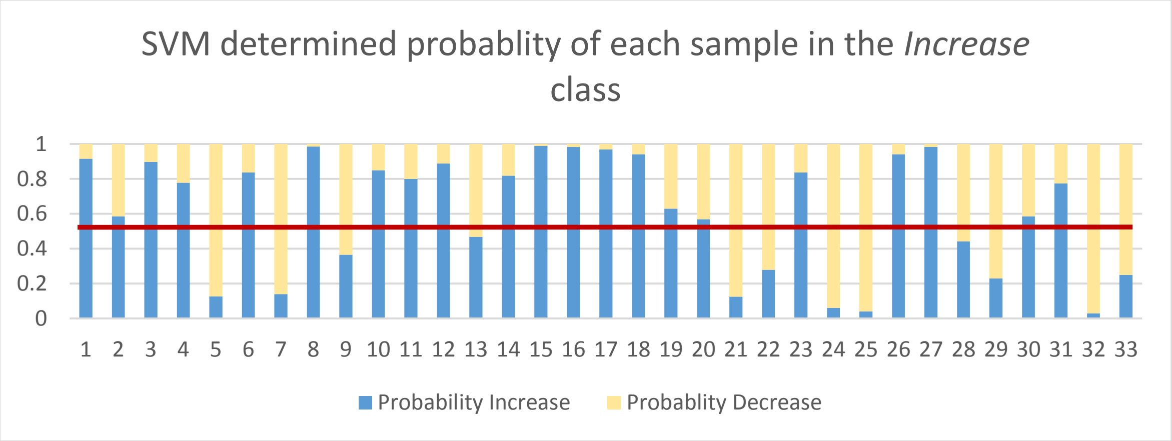


Figure 7.3 - *A chart showing the classification probability of each Increase instance in SET 2. Note: These probabilities were obtained by modifying the SVM parameters to build a logistic model. The classification accuracy experienced a drop from 74.29% to 72.86%.*

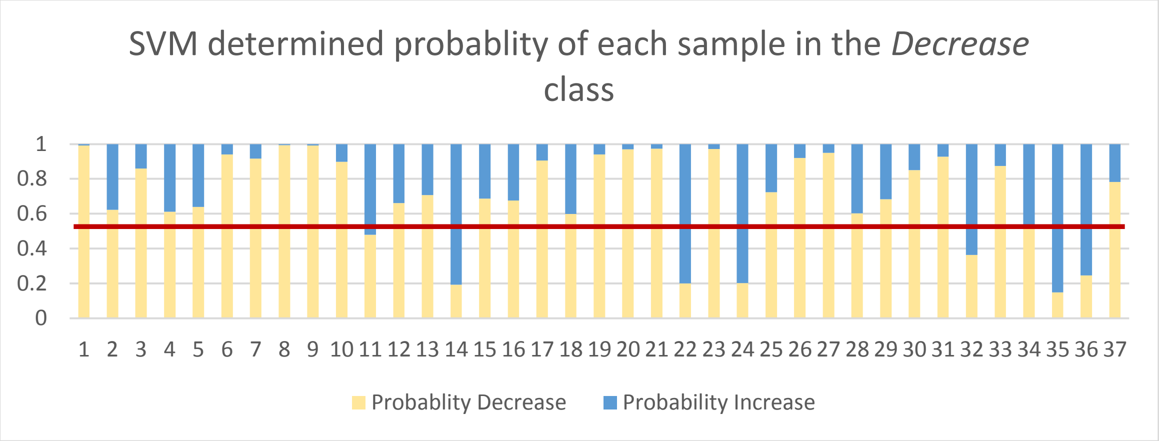


Figure 7.4 - *A chart showing the classification probability of each Decrease instance in SET 2. Note: These probabilities were obtained by modifying the SVM parameters to build a logistic model. The classification accuracy experienced a drop from 74.29% to 72.86%.*

Figure 7.3 and *Figure 7.4* show the class probability determined by SVM for each instance in SET 2. The yellow bars represent the probability that this instance is part of the *Decrease* class according to SVM. The blue bar represents the probability that the instance is part of the *Increase* class. In *Figure 7.3* any instance where the blue bar is above the red line, the instance has been correctly classified. The same is true for *Figure 7.4* and yellow bars.

7.5 Discussion

The assessment of the results in this study allows us to accept the hypothesis formulated at the beginning of the experiment, that the direction of weekly WTI crude oil prices **can** be *effectively* (better than 50-50 coin toss) predicted using **machine learning techniques** and modelling weekly WTI crude oil price as a function of USFP and OC sentiment and frequency of "oil" and OPEC member references. As "effectively" was defined as better than a coin toss, we have achieved our goal across all three classification models. The findings from *Chapter* 7 have also been support through this investigation. The models built with a lag of 7 performed significantly better than the models with a lag of 1. SVM achieved the highest classification accuracy out of all models. This could be due to the fact that SVM relies on only the support vectors to ensure optimality whilst the other two models use all data points. Naïve Bayes performed quite well as well with a classification accuracy of 67.14 %. Using supervised discretization significantly improved the results with Naïve Bayes.

7.6 Conclusion

Comparing our model with the models currently in literature (*Table 7.6)*, we can observe that our SVM model has achieved the highest accuracy. From the results we can conclude that the novel inputs extracted from Twitter (USFP and OC sentiment, frequency of "oil" and OPEC references) are better predictors of future prices WTI crude oil prices than the previous inputs mentioned in literature. Furthermore, this study supports the claim that non-linear models have superior forecasting abilities when it comes to crude oil predictions. The results from this study also support the findings in the previous Chapter that the price of WTI crude oil is not immediately impacted by change in supply and demand. Rather, these factors have a lagging impact of 7 weeks.

Table 7.9 - *A comparison of directional crude oil price predictive models in literature*

Contributors	Approach	Frequency	Period	Accuracy
Morana (2001)	Semi parametric approach	Daily	11/21/1998 to 1/21/1999	46.67%
Gori et al. (2007)	Adaptive Neuro Fuzzy Inference System (ANFIS)	Monthly	2/1999 to 12/2003	45.76%
Fan et al (2006)	Genetic algorithm	Daily	6/27/2005 to 7/26/2005	54.54%
Ghaffari and Zare (2009)	Adaptive Neuro Fuzzy Inference System (ANFIS)	Daily	5/1/2007 to 5/31/2007	68.18%
Li et al. (2014)	Least squre support vector regression (LSSVR)	Monthly	1/2/2002 to 3/20/2009	52.52%
Chen (2014)	Linear Regression	Monthly	1/1991 to 8/2012	65.00%
Shambora and Rossiter (2007)	ANNs	Daily	1/1/1998 to 12/31/2003	53.10%
Our Model	Support Vector Machine (SMO)	Weekly	1/3/2011 to 3/2/2015	74.29%*

Chapter 8

Conclusion and Future Works

This investigation allowed us to identify a gap in the existing methodologies of predicting crude oil. The literature showed that the current models relied on historical oil prices, financial instruments, oil variables (e.g. inventory levels), and irregular and infrequent events (i.e. elections, natural disasters) as inputs for their forecasting models. However, non-traditional methods but increasingly popular methods such as social media data analysis, word frequencies had not been explored as inputs to forecast crude oil, thus justifying further exploration.

Referring back to the objectives set out in the introduction, the investigation has successfully achieved all of its goals. Through the use of Twitter API, SQL, a methodology to extra and store US foreign policy and oil companies' sentiment was achieved. The statistical study supported our selection of attributes and led to a novel discovery of the seven week lag. Using the encouraging results of the statistical study, the attributes with statistically significant correlation were used as inputs to the forecasting model. Building the model using three different types of classifiers (Naïve Bayes, SVM, and ANN), it was found that SVM returned the highest classification accuracy of 74%. Through, the use of unexplored inputs, we were able to achieve and overall classification accuracy superior to that of existing models.

Despite the superior performance, there are several **limitations** of our predictive model.

- The model is relies on a limited number of Twitter accounts to retrieve the data. In the long term, these Twitter accounts may become more or less active thus potentially impacting the performance of the model.
- The model relies on sentiment, which is still only 80.7% accurate in the case of Stanford NLP and 64% in the case of SentiStrength. There is a large margin for error in sentiment which can ultimately impact or skew the accuracy of the model.

The model uses a variety of different features to forecast the price of oil. However, are all of those features equally important? The relative contribution of each attribute to the overall classification accuracy can be explored as a next step for this investigation. Furthermore, can this similar methodology be applied to other grades of oil as well such as Brent or Dubai? Referring back to the original aim of the investigation on attempting to improve on the existing methods for forecasting the directional shift of crude oil price by introducing **Twitter** sentiment of **US foreign policy (USFP)** and **oil companies' (OC)** as an input into the machine learning forecasting model for crude oil prices. The study has clearly achieved this purpose as the results show the USFP and OC sentiment as inputs of a machine learning model provide superior performance results to existing models in literature. Nonetheless, additional investigation still needs to take place to test the robustness of the model.

References

1. CHEN, S.-S., 2014. FORECASTING CRUDE OIL PRICE MOVEMENTS WITH OIL-SENSITIVE STOCKS. *Western Economic Association International,* 52(2), pp. 830-844.
2. Chernenko, S., Schwarz, K. & Wright, J., 2004. The Information Content of Forward and Futures Prices: Market Expectations and the Price of Risk. *FRB International Finance Discussion Paper.*
3. Chin, M., Leblanch, M. & Coibion, O., 2005. The Predictive Content of Energy Futures: An Update on Petroluem, Natural Gas, Heating Oil and Gasoline. *HBER.*
4. Diakopoulos, N. & Shamma, D., 2010. Characterizing debate performance via aggregated Twitter sentiment. New York, ACM, pp. 1195-1198.
5. Fan, Y., Zhang, Y.-J., Tsai, H. & Wei, Y., 2008. Estimating Value at risk of curde oil price and its spillover effect using the GED-GARCH approach. *Energy Economics,* Volume 30, pp. 3156-3171.
6. Ghaffari, A. & Zare, S., 2009. A novel algorithm for Prediction of Crude Oil Price Variation Based on Soft Computing. *Energy Economics,* Volume 31, pp. 531-536.
7. Gori, F., Ludovisi, D. & Cerritelli, P., 2007. Forecast of Oil Price and Consumption in the SHort0term Under Thre Scenarios: Parabolic, Linear and Chaotic behavior. *Energy Economics,* Volume 32, pp. 1291-1296.
8. Mirmirani, S. & Li, H., 2004. A Comparison of VAR and Neural Networks with Genetic Algorithm in Forecasting Price of Oil. *Advances in Econometrics,* Volume 19, pp. 203-223.
9. Mohammadi, H. & Su., L., 2010. Applications of ARIMA-GARCH Models. *Energy Economics,* Volume 32, pp. 1001-1008.
10. Morana, C., 2001. A Semiparametric Approach to Short-term Oil Price Forecasting. *Energy Economics,* Volume 23, pp. 325-338.
11. OpenCV, 2015. *Introduction to Support Vector Machines.* [Online]
    Available at:
    http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
    [Accessed 1 April 2015].

12. Pindyck, R., 1999. The Long-run Evolution of Energy Prices. *The Energy Journal,* Volume 20, pp. 1-27.
13. Rachenko, S., 2005. The Long-run Forecasting of Energy Prices Using the Model of Shifting Trend. *University of North Carolina at Charlotte.*
14. Russell, M., 2014. Lecture 17: Introduction to Artificial Neural Networks. *EE3J2.*
15. Saedsayad.com, 2015. *Artificial Neural Network.* [Online]
    Available at: http://www.saedsayad.com/artificial_neural_network.htm
    [Accessed 1 April 2015].
16. Shambora, W. & Rossiter, R., 2007. Are There Exploitable Inefficiencies in the Futures Market for Oil?. *Energy Econoimcs,* Volume 29, pp. 18-27.
17. Silva, E. G. S., Legey, L., L. F. & Silva, E. A. S., 2010. Forecasting oil price trends using wavelets and hidden Markov models. *Energy Economics,* Volume 32, pp. 1507-1519.
18. Socher, R., Perelygin, A., Wu, J. Y. & Jason Chuang, 2013. *Recursive Deep Models for Semantic Compositionality.* s.l., ENNLP.
19. Stanford, 2013. *An Introduction to Convolutional Neural Networks.* [Online]
    [Accessed 1 April 2014].
20. Tehrani, R. & Khodayar, F., 2011. A hybrid optimized artificial intelligent model to forecast crude oil using genetic algorithm. *African Journal of Business Management.*
21. Wang, S., Yu, L. & Lai, K., 2004. A Novvel Hybrid AI System Framework for Crude Oil Prie Forecasting. *Lecture Notes in Computer Science,* Volume 3327, pp. 233-242.
22. Xie, W., Yu, L., Xu, S. & Wang, S., 2006. A new method for Crude Oil Price Forecasting Based on Support Vector Machines. *International Conference on Computational Science,* pp. 444-451.
23. Ye, M., J, Z. & J, S., n.d. Forecasting Short-run Crue Oil Price Using High and Low Inventory Variables. *Energy Policy,* Volume 34, pp. 2736-2743.
24. Ye, M., Zyren, J. & Sore, J., 2002. Forecasting Crude Oil Spot Price Using OECD Petrolem Invetory Levels. *International Advances in Ecnomic Research,* Volume 8, pp. 324-334.
25. Yu, L., Wang, S. & Lai, K. K., 2008. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics.*

Appendix A

List of United States Foreign Policy Think Tanks and Associated Twitter Accounts

Name	**Twitter Account**
AEI Foreign Policy	@AEIfdp
American-Iranian Council	@US_Iran
Aspen Institute	@AspenInstitute
AtlanticCouncil	@AtlanticCouncil
BASIC	@basic_int
Brookings	@BrookingsInst
Carnegie Council	@carnegiecouncil
Cato Institute	@CatoInstitute
Center for a New American Security	@CNASdc
American Progress	@amprog
Center for Global Development	@CGDev
Center for International Maritime Security	@CIMSEC
Center for International Policy	@CIPonline
Center for Security Policy	@securefreedom
Center for Strategic and Budgetary Assessments	@CSBA_
The Center for Strategic & Int'l Studies	@CSIS
The Center for the National Interest	@CFTNI
The Center on International Cooperation	@nyuCIC
The Chicago Council on Global Affairs	@ChicagoCouncil
Claremont Institute	@ClaremontInst
The Combating Terrorism Center	@CTCWP
The Council for the National Interest	@fixMidEpolicy
The Council on Foreign Relations	@CFR_org
Council on Hemispheric Affairs	@cohastaff
Eisenhower Institute	@eigbc
Foreign Policy In Focus	@FPIF
The Foreign Policy Initiative	@ForeignPolicyI
Foreign Policy Research Institute	@FPRInews
Gatestone Institute	@GatestoneInst
Global Financial Integrity	@GFI_Tweets
Halifax International Security Forum	@HFXforum
Heritage Foundation	@Heritage
Hoover Institution	@HooverInst
Hudson Institute	@HudsonInstitute
The Institute for National Security and Counterterrorism	@INSCT

Institute for the Study of War	@TheStudyofWar
Inter-American Dialogue	@The_Dialogue
The Jamestown Foundation	@JamestownTweets
Jewish Institute for National Security Affairs	@jinsadc
Kennan Institute	@kennaninstitute
Kissinger Associates	@KissingerAssoc
The McCain Institute	@McCainInstitute
Middle East Forum	@meforum
The Middle East Policy Council	@MidEastPolicy
Miller Center for Public Affairs	@Miller_Center
The Minnesota International Center	@MICglobe
The National Bureau of Asian Research	@NBRnews
National Center for Policy Analysis	@NCPA
NCAFP	@NATLCOMMITTEE
National Security Network	@natsecnet
Nautilus Institute	@Nautilus_Inst
New America Foundation	@NewAmerica
Nuclear Threat Initiative	@NTI_WMD
Pacific Council	@PacCouncil
Pacific Forum CSIS	@PacificForum
Project 2049 Inst	@Project2049
The Project on Middle East Democracy	@POMEDwire
RAND Corporation	@RANDCorporation
Stratfor	@Stratfor
The Streit Council	@StreitCouncil
The U.S.-China Policy Foundation	@USCPF
U.S. Inst. of Peace	@USIP
Washington Institute	@WashInstitute
Watson Institute	@WatsonInstitute
The Wilson Center	@TheWilsonCenter
WorldAffairsCouncils	@WACAmerica
World Policy	@WorldPolicy
Foreign Policy	@ForeignPolicy
Department of State	@StateDept
StratPost	@StratPost
Brookings FP	@BrookingsFP
Foreign Affairs	@ForeignAffairs
RealClearWorld	@RealClearWorld
Global Security News	@NTI_GSN
GlobalPost	@GlobalPost

Appendix B

List of Oil Companies and Associations with Twitter Accounts

Name	Twitter Account
Oil and Gas News	@oilandgasnews
oil and gas industry	@oil_and_gas
ExxonMobil UK	@ExxonMobil_UK
Halliburton	@Halliburton
Baker Hughes	@BHInc
Oil & Gas Technology	@OGTCavendish
Maersk Oil	@maerskoil
Oil & Gas IQ News	@OilandGasIQ
GE Oil & Gas	@ge_oilandgas
Rigzone	@Rigzone
UpstreamOnline	@UpstreamOnline
Platts Gas	@PlattsGas
Platts Oil	@PlattsOil
EnergyUpdate	@EnergyUpdate
Petroleum Economist	@PetroleumEcon
Offshore	@offshoremgzn
OPEC News	@OPECnews
OECD Statistics	@OECD_Stat
OECD	@OECD
FuelFix	@fuelfixblog
Anjli Raval	@AnjliRaval
Jennifer A. Dlouhy	@jendlouhyhc
Crude Oil Prices	@CrudeOilPrices
Oil&Gas Investments	@OilandGasInvest
US CRUDE OIL	@USCRUDEOIL
Oil & Gas Journal	@OGJOnline
Energy Department	@ENERGY
Ernest Moniz	@ErnestMoniz
IEA	@IEA
FT Energy	@ftenergy
World Oil Online	@WorldOil
Enel Group	@EnelGroup
EIA	@EIAgov
ConocoPhillips	@conocophillips
Statoil ASA	@statoilasa
eni.com	@eni
LUKOIL	@lukoilengl

Qatar Petroleum	@qatarpetroleum
Petrobras Global	@petrobrasglobal
Total	@Total
Chevron	@Chevron
Shell Oil Company	@Shell_US
Shell	@Shell
BP America	@BP_America
BP	@BP_plc
PetroChina News	@PetroChinaBRK
PetroChina	@chinapetro
ExxonMobil	@exxonmobil
ExxonMobil Europe	@ExxonMobil_EU
Gazprom	@GazpromNewsEN
Saudi Aramco	@Saudi_Aramco
ICIS	@ICISOfficial

Appendix C

List of OPEC Countries

Country

Algeria

Angola

Ecuador

Iran

Iraq

Kuwait

Libya

Nigeria

Qatar

Saudi Arabia

United Arab Emirates

Venezuela

Appendix D

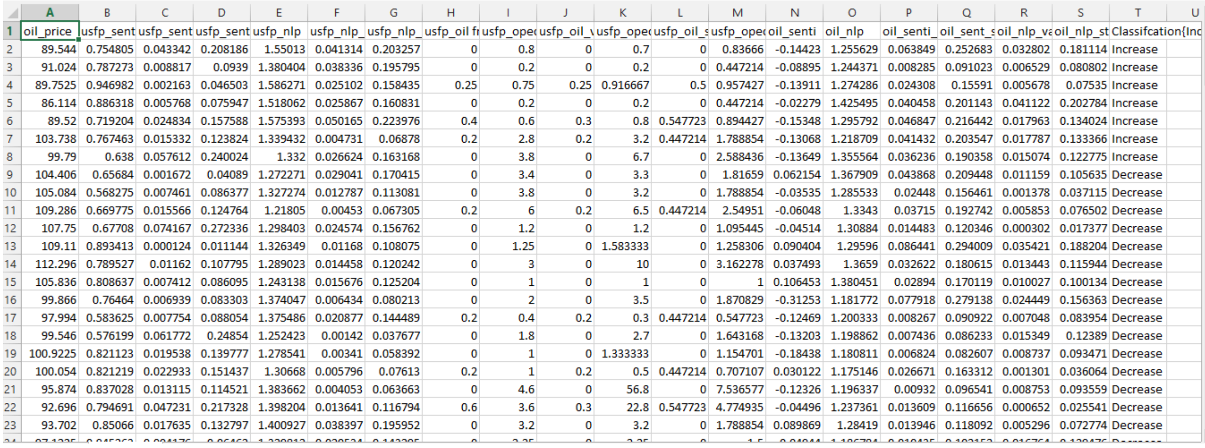
List of Weekly WTI Crude Oil Prices

Date	$/bar	Date	$/bar	Date	$/bar	Date	$/bar
1/3/2011	89.544	9/12/2011	88.934	5/21/2012	90.882	1/28/2013	97.332
1/10/2011	91.024	9/19/2011	83.652	5/28/2012	87.0575	2/4/2013	96.176
1/17/2011	89.7525	9/26/2011	81.178	6/4/2012	84.434	2/11/2013	96.954
1/24/2011	86.114	10/3/2011	79.434	6/11/2012	83.27	2/18/2013	94.38
1/31/2011	89.52	10/10/2011	85.348	6/18/2012	81.11	2/25/2013	92.19
2/7/2011	85.514	10/17/2011	86.818	6/25/2012	80.226	3/4/2013	91.004
2/14/2011	84.134	10/24/2011	92.316	7/2/2012	85.735	3/11/2013	92.7
2/21/2011	95.26	10/31/2011	93.244	7/9/2012	85.78	3/18/2013	93.046
2/28/2011	101.052	11/7/2011	96.966	7/16/2012	90.34	3/25/2013	96.0775
3/7/2011	103.738	11/14/2011	99.318	7/23/2012	88.876	4/1/2013	95.074
3/14/2011	99.79	11/21/2011	96.89	7/30/2012	89.098	4/8/2013	93.36
3/21/2011	104.406	11/28/2011	99.906	8/6/2012	93.14	4/15/2013	88
3/28/2011	105.084	12/5/2011	100.078	8/13/2012	94.434	4/22/2013	90.998
4/4/2011	109.286	12/12/2011	96.064	8/20/2012	96.224	4/29/2013	93.4
4/11/2011	107.75	12/19/2011	97.74	8/27/2012	95.684	5/6/2013	95.844
4/18/2011	109.11	12/26/2011	99.81	9/3/2012	95.675	5/13/2013	94.648
4/25/2011	112.296	1/2/2012	102.3875	9/10/2012	97.562	5/20/2013	94.756
5/2/2011	105.836	1/9/2012	100.432	9/17/2012	93.702	5/27/2013	93.32
5/9/2011	99.866	1/16/2012	99.945	9/24/2012	91.348	6/3/2013	94.25
5/16/2011	97.994	1/23/2012	99.354	10/1/2012	90.814	6/10/2013	96.358
5/23/2011	99.546	1/30/2012	97.8	10/8/2012	91.422	6/17/2013	96.652
5/30/2011	100.9225	2/6/2012	98.56	10/15/2012	91.59	6/24/2013	95.83
6/6/2011	100.054	2/13/2012	101.726	10/22/2012	86.354	7/1/2013	100.65
6/13/2011	95.874	2/20/2012	107.175	10/29/2012	85.87	7/8/2013	104.704
6/20/2011	92.696	2/27/2012	107.52	11/5/2012	85.982	7/15/2013	106.882
6/27/2011	93.702	3/5/2012	106.324	11/12/2012	85.866	7/22/2013	105.876
7/4/2011	97.1225	3/12/2012	106.15	11/19/2012	87.4	7/29/2013	105.544
7/11/2011	96.72	3/19/2012	106.41	11/26/2012	87.274	8/5/2013	105.166
7/18/2011	98.006	3/26/2012	105.122	12/3/2012	87.002	8/12/2013	106.974
7/25/2011	97.828	4/2/2012	103.5225	12/10/2012	85.712	8/19/2013	105.476
8/1/2011	90.854	4/9/2012	102.552	12/17/2012	88.244	8/26/2013	108.33
8/8/2011	82.862	4/16/2012	103.152	12/24/2012	90.1425	9/2/2013	108.77
8/15/2011	85.364	4/23/2012	103.784	12/31/2012	92.765	9/9/2013	108.356
8/22/2011	85.056	4/30/2012	103.472	1/7/2013	93.38	9/16/2013	106.218
8/29/2011	88.072	5/7/2012	96.984	1/14/2013	94.582	9/23/2013	103.096
9/5/2011	87.905	5/14/2012	93.108	1/21/2013	95.4125	9/30/2013	103.144

Date	$/bar	Date	$/bar	Date	$/bar
10/7/2013	102.698	3/24/2014	100.66	9/8/2014	92.43
10/14/2013	101.508	3/31/2014	100.462	9/15/2014	93.52
10/21/2013	97.572	4/7/2014	102.72	9/22/2014	93.15
10/28/2013	96.938	4/14/2014	103.9475	9/29/2014	91.444
11/4/2013	94.306	4/21/2014	102.112	10/6/2014	87.628
11/11/2013	93.944	4/28/2014	100.508	10/13/2014	82.88
11/18/2013	93.92	5/5/2014	100.29	10/20/2014	82.122
11/25/2013	92.9675	5/12/2014	101.916	10/27/2014	81.292
12/2/2013	96.206	5/19/2014	103.82	11/3/2014	78.242
12/9/2013	97.23	5/26/2014	103.9525	11/10/2014	76.496
12/16/2013	97.854	6/2/2014	103.234	11/17/2014	75.378
12/23/2013	99.1525	6/9/2014	105.968	11/24/2014	72.355
12/30/2013	96.4675	6/16/2014	107.228	12/1/2014	67.178
1/6/2014	92.416	6/23/2014	106.692	12/8/2014	61.136
1/13/2014	92.976	6/30/2014	105.5175	12/15/2014	55.89
1/20/2014	96.1875	7/7/2014	103.254	12/22/2014	55.58
1/27/2014	97.29	7/14/2014	102.368	12/29/2014	53.4425
2/3/2014	97.78	7/21/2014	104.346	1/5/2015	48.774
2/10/2014	100.208	7/28/2014	102.194	1/12/2015	47.066
2/17/2014	102.9325	8/4/2014	97.496	1/19/2015	46.4575
2/24/2014	102.772	8/11/2014	97.172	1/26/2015	45.326
3/3/2014	103.074	8/18/2014	94.954	2/2/2015	50.576
3/10/2014	99.554	8/25/2014	96.258	2/9/2015	51.136
3/17/2014	99.774	9/1/2014	94.0625	2/16/2015	51.69
				2/23/2015	49.156
				3/2/2015	49.59

Appendix E

Screenshot Featureset.csv



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	oil_price	usfp_sent	usfp_sent	usfp_sent	usfp_nlp	usfp_nlp_	usfp_nlp_	usfp_oil fr	usfp_opec	usfp_oil_v	usfp_opec	usfp_oil_s	usfp_opec	oil_senti	oil_nlp	oil_senti_	oil_sent_s	oil_nlp_va	oil_nlp_st	Classifcation{Inc	
2	89.544	0.754805	0.043342	0.208186	1.55013	0.041314	0.203257	0	0.8	0	0.7	0	0.83666	-0.14423	1.255629	0.063849	0.252683	0.032802	0.181114	Increase	
3	91.024	0.787273	0.008817	0.0939	1.380404	0.038336	0.195795	0	0.2	0	0.2	0	0.447214	-0.08895	1.244371	0.008285	0.091023	0.006529	0.080802	Increase	
4	89.7525	0.946982	0.002163	0.046503	1.586271	0.025102	0.158435	0.25	0.75	0.25	0.916667	0.5	0.957427	-0.13911	1.274286	0.024308	0.15591	0.005678	0.07535	Increase	
5	86.114	0.886318	0.005768	0.075947	1.518062	0.025867	0.160831	0	0.2	0	0.2	0	0.447214	-0.02279	1.425495	0.040458	0.201143	0.041122	0.202784	Increase	
6	89.52	0.719204	0.024834	0.157588	1.575393	0.050165	0.223976	0.4	0.6	0.3	0.8	0.547723	0.894427	-0.15348	1.295792	0.046847	0.216442	0.017963	0.134024	Increase	
7	103.738	0.767463	0.015332	0.123824	1.339432	0.004731	0.06878	0.2	2.8	0.2	3.2	0.447214	1.788854	-0.13068	1.218709	0.041432	0.203547	0.017787	0.133366	Increase	
8	99.79	0.638	0.057612	0.240024	1.332	0.026624	0.163168	0	3.8	0	6.7	0	2.588436	-0.13649	1.355564	0.036236	0.190358	0.015074	0.122775	Increase	
9	104.406	0.65684	0.001672	0.04089	1.272271	0.029041	0.170415	0	3.4	0	3.3	0	1.81659	0.062154	1.367909	0.043868	0.209448	0.011159	0.105635	Decrease	
10	105.084	0.568275	0.007461	0.086377	1.327274	0.012787	0.113081	0	3.8	0	3.2	0	1.788854	-0.03535	1.285533	0.02448	0.156461	0.001378	0.037115	Decrease	
11	109.286	0.669775	0.015566	0.124764	1.21805	0.00453	0.067305	0.2	6	0.2	6.5	0.447214	2.54951	-0.06048	1.3343	0.03715	0.192742	0.005853	0.076502	Decrease	
12	107.75	0.67708	0.074167	0.272336	1.298403	0.024574	0.156762	0	1.2	0	1.2	0	1.095445	-0.04514	1.30884	0.014483	0.120346	0.000302	0.017377	Decrease	
13	109.11	0.893413	0.000124	0.011144	1.326349	0.01168	0.108075	0	1.25	0	1.583333	0	1.258306	0.090404	1.29596	0.086441	0.294009	0.035421	0.188204	Decrease	
14	112.296	0.789527	0.01162	0.107795	1.289023	0.014458	0.120242	0	3	0	10	0	3.162278	0.037493	1.3659	0.032622	0.180615	0.013443	0.115944	Decrease	
15	105.836	0.808637	0.007412	0.086095	1.243138	0.015676	0.125204	0	1	0	1	0	1	0.106453	1.380451	0.02894	0.170119	0.010027	0.100134	Decrease	
16	99.866	0.76464	0.006939	0.083303	1.374047	0.006434	0.080213	0	2	0	3.5	0	1.870829	-0.31253	1.181772	0.077918	0.279138	0.024449	0.156363	Decrease	
17	97.994	0.583625	0.007754	0.088054	1.375486	0.020877	0.144489	0.2	0.4	0.2	0.3	0.447214	0.547723	-0.12469	1.200333	0.008267	0.090922	0.007048	0.083954	Decrease	
18	99.546	0.576199	0.061772	0.24854	1.252423	0.00142	0.037677	0	1.8	0	2.7	0	1.643168	-0.13203	1.198862	0.007436	0.086233	0.015349	0.12389	Decrease	
19	100.9225	0.821123	0.019538	0.139777	1.278541	0.00341	0.058392	0	1	0	1.333333	0	1.154701	-0.18438	1.180811	0.006824	0.082607	0.008737	0.093471	Decrease	
20	100.054	0.821219	0.022933	0.151437	1.30668	0.005796	0.07613	0.2	1	0.2	0.5	0.447214	0.707107	0.030122	1.175146	0.026671	0.163312	0.001301	0.036064	Decrease	
21	95.874	0.837028	0.013115	0.114521	1.383662	0.004053	0.063663	0	4.6	0	56.8	0	7.536577	-0.12326	1.196337	0.00932	0.096541	0.008753	0.093559	Decrease	
22	92.696	0.794691	0.047231	0.217328	1.398204	0.013641	0.116794	0.6	3.6	0.3	22.8	0.547723	4.774935	-0.04496	1.237361	0.013609	0.116656	0.000652	0.025541	Decrease	
23	93.702	0.85066	0.017635	0.132797	1.400927	0.038397	0.195952	0	3.2	0	3.2	0	1.788854	0.089869	1.28419	0.013946	0.118092	0.005296	0.072774	Decrease	
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	3.25	0	3.25	0	1.5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Decrease	

Appendix F

MATLAB Granger Causality Test

```
function [F,c_v] = granger_cause(x,y,alpha,max_lag)
% [F,c_v] = granger_cause(x,y,alpha,max_lag)
% Granger Causality test
% Does Y Granger Cause X?
%
% User-Specified Inputs:
%   x -- A column vector of data
%   y -- A column vector of data
%   alpha -- the significance level specified by the user
%   max_lag -- the maximum number of lags to be considered
% User-requested Output:
%   F -- The value of the F-statistic
%   c_v -- The critical value from the F-distribution
%
% The lag length selection is chosen using the Bayesian information
% Criterion
% Note that if F > c_v we reject the null hypothesis that y does not
% Granger Cause x

% Chandler Lutz, UCR 2009
% Questions/Comments: chandler.lutz@email.ucr.edu
% $Revision: 1.0.0 $  $Date: 09/30/2009 $
% $Revision: 1.0.1 $  $Date: 10/20/2009 $
% $Revision: 1.0.2 $  $Date: 03/18/2009 $

% References:
% [1] Granger, C.W.J., 1969. "Investigating causal relations by econometric
%     models and cross-spectral methods". Econometrica 37 (3), 424-438.

% Acknowledgements:
%   I would like to thank Mads Dyrholm for his helpful comments and
%   suggestions

%Make sure x & y are the same length
if (length(x) ~= length(y))
    error('x and y must be the same length');
end

%Make sure x is a column vector
[a,b] = size(x);
if (b>a)
    %x is a row vector -- fix this
    x = x';
end

%Make sure y is a column vector
[a,b] = size(y);
if (b>a)
    %y is a row vector -- fix this
    y = y';
end
```

```
%Make sure max_lag is >= 1
if max_lag < 1
    error('max_lag must be greater than or equal to one');
end

%First find the proper model specification using the Bayesian Information
%Criterion for the number of lags of x

T = length(x);

BIC = zeros(max_lag,1);

%Specify a matrix for the restricted RSS
RSS_R = zeros(max_lag,1);

i = 1;
while i <= max_lag
    ystar = x(i+1:T,:);
    xstar = [ones(T-i,1) zeros(T-i,i)];
    %Populate the xstar matrix with the corresponding vectors of lags
    j = 1;
    while j <= i
        xstar(:,j+1) = x(i+1-j:T-j);
        j = j+1;
    end
    %Apply the regress function. b = betahat, bint corresponds to the 95%
    %confidence intervals for the regression coefficients and r = residuals
    [b,bint,r] = regress(ystar,xstar);

    %Find the bayesian information criterion
    BIC(i,:) = T*log(r'*r/T) + (i+1)*log(T);

    %Put the restricted residual sum of squares in the RSS_R vector
    RSS_R(i,:) = r'*r;

    i = i+1;

end

[dummy,x_lag] = min(BIC);

%First find the proper model specification using the Bayesian Information
%Criterion for the number of lags of y

BIC = zeros(max_lag,1);

%Specify a matrix for the unrestricted RSS
RSS_U = zeros(max_lag,1);

i = 1;
while i <= max_lag

    ystar = x(i+x_lag+1:T,:);
    xstar = [ones(T-(i+x_lag),1) zeros(T-(i+x_lag),x_lag+i)];
    %Populate the xstar matrix with the corresponding vectors of lags of x
    j = 1;
    while j <= x_lag
        xstar(:,j+1) = x(i+x_lag+1-j:T-j,:);
        j = j+1;
    end
    %Populate the xstar matrix with the corresponding vectors of lags of y
    j = 1;
    while j <= i
        xstar(:,x_lag+j+1) = y(i+x_lag+1-j:T-j,:);
        j = j+1;
    end
    %Apply the regress function. b = betahat, bint corresponds to the 95%
    %confidence intervals for the regression coefficients and r = residuals
    [b,bint,r] = regress(ystar,xstar);

    %Find the bayesian information criterion
    BIC(i,:) = T*log(r'*r/T) + (i+1)*log(T);

    RSS_U(i,:) = r'*r;

    i = i+1;

end

[dummy,y_lag] =min(BIC);

%The numerator of the F-statistic
F_num = ((RSS_R(x_lag,:) - RSS_U(y_lag,:))/y_lag);

%The denominator of the F-statistic
F_den = RSS_U(y_lag,:)/(T-(x_lag+y_lag+1));

%The F-Statistic
F = F_num/F_den;

c_v = finv(1-alpha,y_lag,(T-(x_lag+y_lag+1)));
```

School of Electronic, Electrical and Systems Engineering

UNIVERSITY OF BIRMINGHAM

GENERAL ETHICAL QUESTIONNAIRE FOR ALL STUDENTS

Name of student	Ahmed Zaidi
Email address of student	AXZ109@bham.ac.uk
Name of supervisor	Mourad Oussalah

Title of Research Project:	Forecasting Weekly WTI Crude Oil Using Twitter Sentiment of US

Will the research project involve humans as participants of the research (with or without their knowledge or consent at the time)? This will include any survey, interview or questionnaire that human participants may be asked to complete at any stage as the main part of the project or in the evaluation of the results/deliverables of the project. It will also include any testing of devices, software and other deliverables, which may arise as a result of a project and involves human participants other than the student undertaking the project. Analysis of images or other recordings of human participants or their property and personal possessions is also included.	No
Are the results of the research project likely to expose any person to physical or psychological harm? (Note, before starting the project you will need to complete a risk assessment in all cases)	No
Will you have access to personal information that allows you to identify individuals, or to corporate or company confidential information (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)?	No
Does the research project present a significant risk to the environment or society?	No
Are there any ethical issues raised by this research project that in the opinion of your supervisor require further ethical review?	No

You have answered NO to all of the above questions. Further ethical review is not necessary. You should have this form available at the bench inspections and include it in your final report.