DEPARTMENT OF BUSINESS ECONOMICS

EXPECTATION MEASUREMENT USING SENTIMENT ANALYSIS: AN APPLICATION TO FORECASTING OIL PRICES PROJECT WORK

by

PUNEET JAIN

2014-2016

submitted to

DR. ANANYA G DASTIDAR

CERTIFICATE OF DECLARATION

This is to certify that the Report entitled **EXPECTATION MEASUREMENT USING SENTIMENT ANALYSIS: AN APPLICATION TO FORECASTING OIL PRICES** which is submitted by me in partial fulfillment of the requirement for the award of degree of MBA (Business Economics) to the Department of Business Economics, South Campus, University of Delhi, comprises only my original work and due acknowledgment has been made in the text to all other materials used. This work has not been submitted/published anywhere else.

Name and Signature of the Supervisor

Name and signature of the Candidate

Roll No.:

Contents

INTRODUCTION

There is unanimous agreement that unforeseen large and insistent variation in the price of crude oil are unfavorable to the well-being of both oil-dependent and oil-exporting economies. Consistent & efficient prediction of the price of crude oil are required for a variety of applications. For example, central banks and private sector players factor in price of crude oil for developing macroeconomic trends and in evaluating the implied risks. Of particular importance is the relation of movement of oil price in predicting recession. For example, Edelstein and Kilian (2009), provides evidence that the financial crisis of 2008 was amplified and preceded by an economic slowdown in the automobile industry and a decline in consumer sentiment. Thus, more precise predictions of the crude oil have the potential of improving accuracy for a wide range of macroeconomic outcomes and of improving policy responses.

Apart from its utility in estimating the forecasts for the macro-economy crude oil price movement also influences user's and utility producer's behavior and decision making. For example, crude oil price directly affects the price of kerosene which is heavily used for airline industry for setting base fare, the price of diesel (another derivative of crude oil) is used by car-manufacturers to estimate product sale and decide the new product specification and finally sector's such has power and construction also reply heavily on forecast to decide whether to expand the storage facilities or build new innovative capabilities.

Consumer use forecast to make buying preference for automobile and other equipment which run on derivative of crude oil products such as diesel, kerosene, petrol and CNG. This demand is then internally used by utility companies to predict demand and create capacities accordingly also this consumer behavior is used to predict the amount of consumption hence estimate the amount of carbon added to the environment, hence aids in the development of environmental policies and estimation of carbon credits, taxes etc.

Since the advent of globalization oil prices movement has been greatly influenced by consumer/market sentiment as famously quoted by the economist Keynes and his "animal spirit”. The “expectation” of the consumer has been shown to greatly affect the volatility in the various securities market (T Rao ,Bollen et al). But measuring sentiment is in itself an uphill task and provides a lot of problems given its subject nature. First the issue of availability of data regarding any individual’s perception or sentiment on a particular subject in real time basis. Second given a person’s response the issue of scaling his response on a given metric keeping in mind the highly subject nature of the written text, and last but not least the issue of selection of respondent and the scale of sample size for optimum efficiency and consistency of estimator.

The leapfrog advancement in computer processing and technology and the advent of social media has made it possible to circumvent the problems faced in measuring sentiment on a real-time basis.

The processing speed of computers and the advent of sophisticated software has made it possible to analyze gigabytes of data in the fraction of the amount it used to take before (Moore’s Law). Also development in statistical methods and interpretation has led to development of niche software such as SAS, SPSS, R and STATA which help capture the sentiment and quantify it on real time basis. Also every day new and new methods are being developed in this regards

Development in NLP or Natural Linguistic Programming has helped computers understand the dynamics of human languages and develop niche tools such as parsing, NER tokenization etc. NLP with the help of machine learning algorithm really exploded which lead to the development of sentiment analysis and is still being actively researched and developed by leading universities like Stanford and MIT. hence NLP based sentiment classifier can help in development of customer sentiment model by processing the required text in the source material to extract the information required for analysis

Another benefit in the boom of computer and information technology is the rise of social media which provides a platform to people from around the globe to present their views on a topic in an unbiased and relatively free manner. Given its universal reach and easy to use interface makes it optimum to gather sentiment about a particular topic. In the fore-front of the social media race are websites turned corporation like Facebook, twitter, google plus, Myspace etc. Access to the constant stream will provide raw and unbiased data from nearly every corner of the globe and will help generate the customer sentiment index which the be used for developing the forecast of crude oil price with better prediction and out of -sample forecast capability.

With near globe presence and wide recognition social media has become the most commonly used medium of discussion by the financial pundits and investors alike and provide an optimum platform for data analyst and scholars to track and mine mounds of data to map complex pattern and dynamic associated with public sentiment for various commodity which leads to boom and busts. Apps like Stock Twits and Money control have completely socialized the dynamics of financial trade, economic Gurus give their prediction and insight on the go, financial pundits share their strategies, tips and last investor share their emotion such as rejoice for profit and fear and sorrow for loss for various securities and commodities at an impressive rate of 350 million messages and counting every day (Techcrunch October 2014). These tweets can be accessed through simple keyword search.

So this Project deals with the issue of forecasting oil price by developing a sentiment classifier based on tweets collected from twitter and using it to develop a forecasting model and trace its performance with the existing model.

So the objective of the project are as follows

- **Develop a sentiment classifier for oil price expectation using NLP based process available**
- **Develop a "word-cloud" for real time sentiment gathering**
- **Develop a Multi-Variate Forecasting Model using Sentiment Classifier and other Macro-Economic variables**

- **Check causality of Sentiment model and get its directionality.**
- **Check the accuracy of the Model Forecast with other Forecasting models and also**
- **check the out-of-sample forecast accuracy using RMSE and MAPE**

The plan of project goes from literature review to framework of analysis, after the methodology is presented using which empirical analysis is conducted and results are discussed and finally the conclusion is presented with future scope of work. Appendix contains the codes used in the project and a list of figures and tables is also provided.

LITERATURE REVIEW

Xue Zhang (2010) created a simple classifier based on occurrence of words like “happy”, “sad “and analyzed the tweets correlation with the movement in the securities market, they found significant statistical correlation for DJIA, NASDAQ and hence concluded that sentiment can be used to predict the outcome of the stock market as any developed can be tracked in the form of emotional outburst. **However, their model was relatively simple and did not involve advanced sentiment classifier techniques used in this project**

Bollen et al (2010) cited behavioral economics to study public mood and its impact on securities market namely Dow Jones Industrial Average index. They used daily tweets and used google 6-dimension profile of moods to create a classifier and using the classifier created a forecasting model to predict DJIA closing prices. A significant increase of accuracy namely 87.6% was achieved and 6% reduction in MAPE was also observed. **But their model did not establish any granger causality to check the impact of change in sentiment.**

T Rao (2012) based their research on Behavioral finance which incorporates public mood swings in its forecasting model by focusing on message posted on social media websites. They used tweets from twitter and search volume index for a period of one year to create and index and used it to model securities market DJIA, NIKIE. They investigated granger causative nature and included lagged value of sentiment in the forecast model. The model showed high correlation of the lagged sentiment values with the price movement and improved accuracy up to 94% for DJIA and significant reduction in MAPE model as well**. However, their sentiment model was primitive and based on a word bank which gave binary classification and did not extend it to commodity markets**

SERDA SELIN (2014) applied the sentiment analysis for prediction of Euro/Dollar Exchange rate movement and created a 3 tiered sentiment classifier as buy sell neutral.

These suggest that there exists a relationship with the number of tweets and change in exchange rate movement. **But the directionality of the nature was not defined.**

José R. Pires Manso (2012) Did a comprehensive review of the currently practiced trends in forecasting of crude oil price. They divided the techniques primarily into two types quantitative and qualitative where quantitative is further sub-divided into Time-series, financial, structural and machine learning/non-parametric approaches. The qualitative approach deals with Delphi method, text mining/sentiment analysis and other knowledge based techniques. According to the literature the most commonly followed technique is time-series followed by econometric model then by structural methods followed by new machine learning algorithm and qualitative based text mining come in the bottom of the pyramid search.

StanfordNLP Program provides software for Sentiment classification provides statistical learning, deep learning algorithm based NLP for classification into 3 sub categories it has a team comprising of the computer science department, linguistics department and computational social sciences. Software incorporates automatic parser, tokenizer and can be incorporated in leading software such as R and python. **The Stanford Algorithm provided the basis for development of sentiment classifier on the 3-point scale used in the project for forecasting**.

Saif Mohammed (2014) NRC Based Lexicon is a sentiment classifier approach which has a data bank of 18000 words and classifies these words as positive and negative based of association and rules. The source material can also be classified in range of emotional states as well. The lexicon is also available 20+ language and is manually done by crowd sourcing. **The NRC algorithm provided the basis for development of sentiment classifier on the 5-point scale used in the project for forecasting**.

Due the opposing nature of how both the Standard and NRC Classifier works both on them were used to the develop the sentiment classifier for the concerned project.

FRAMEWORK OF ANALYSIS

The role of **Expectation** is a central theme in price setting since the Keynesian era to the behavioral economics advocates. Emotional response is inherent in human nature and turns the investor into impulse animals (“animal Spirit”). In the present world the oil price is determined by a number of macro-economic indicator’s like OPEC meeting, total production, political climate in the Arab region but these event influence the expectation which leads to these price movements.**hence any index which measure the sentiment expectation of the market should be correlated with the price movement.**

Hence development of such and index would require the data about the particular commodity from a wide variety of market participants including investor’s, producers, analysts, financial pundits. This is where twitter comes into picture.

Twitter being in the forefront of the social media race is one of the top ten most visited websites in the world which has 500 million user base of which 325 million at active user. It handles around 1.8 billion search queries everyday about nearly every diverse topic possible. It’s universal interface and feed methodology makes it easy to search the latest buzz of a trending topic. Hence was used to collect data for the building the sentiment classifier model.

Apart from twitter data daily close price for gold was also taken while building the model. The rationale behind the use of gold was that it acts as a proxy for economic stability, as when their turmoil investor generally shifts their investment into safe haven and hence spikes up the prices of gold. Likewise, in a booming economy people tend to invest in assets which give a higher rate of return as they become more risk taking in nature which leads to a decline in the price of oil.

Hence A prior the relation between oil price and consumer sentiment must be positive in nature and that between in gold price and price of oil must be negative in nature. The long run relationship between these gold price and price of crude oil has been thoroughly researched but not definitive conclusion has been presented.

The directiosn of causality is expected flow from expectation/sentiment to movement in price market and should not have significant lag effect with values way in the past

METHODOLOGY

The project consisted of the following steps:

- **Collection of data from Twitter Regarding Oil Price Movement for one year 2015-2016**
- **Using daily tweets obtain the sentiment classifier using StanfordNLP and NRC-Lexicon methodology**
- **Development of Word-Cloud**
- **Developing a forecasting Model to Check Causality**
- **Making other forecasting models and comparing accuracy.**

Collection of data from Twitter Regarding Oil Price Movement

Twitter has an API which lets us access tweets from a specific key word but has the limitation of giving tweets only for the past 10 days. But since Data for one year needed to be calculated manual collection of tweets was done using MS Excel Macro programming and use of twitter advance search page

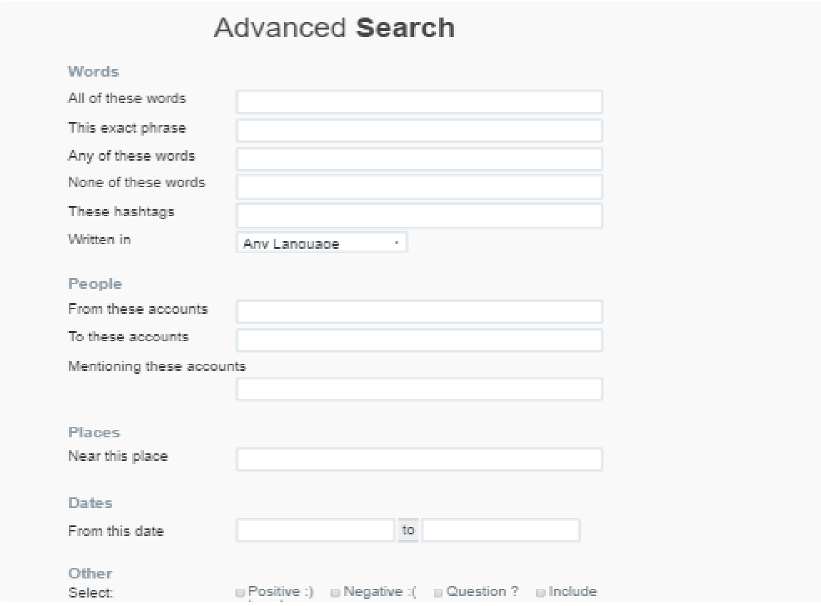
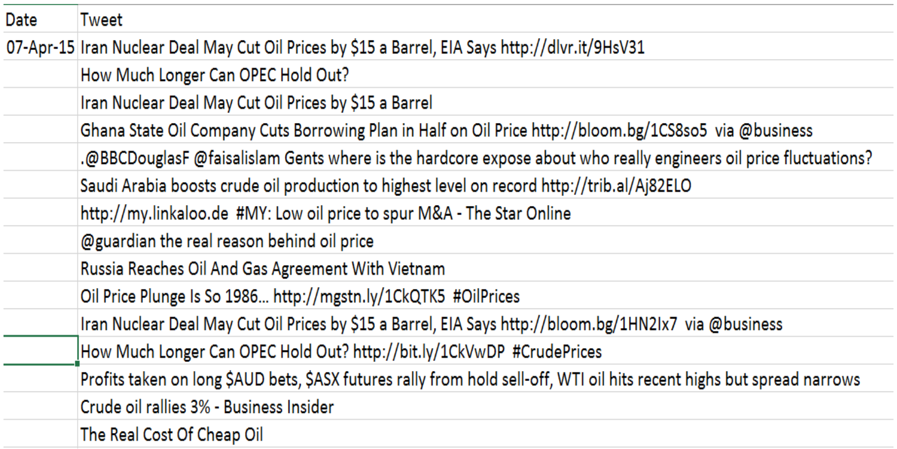


Figure 1 Twitter Advance Page showing options for customized search

Twitter provides the functionality to manually search Tweets based on key words, hashtags, username location and time-line. hence for our project we used keywords such as “oil price”,” crude oil price” and similar such keyword search string.

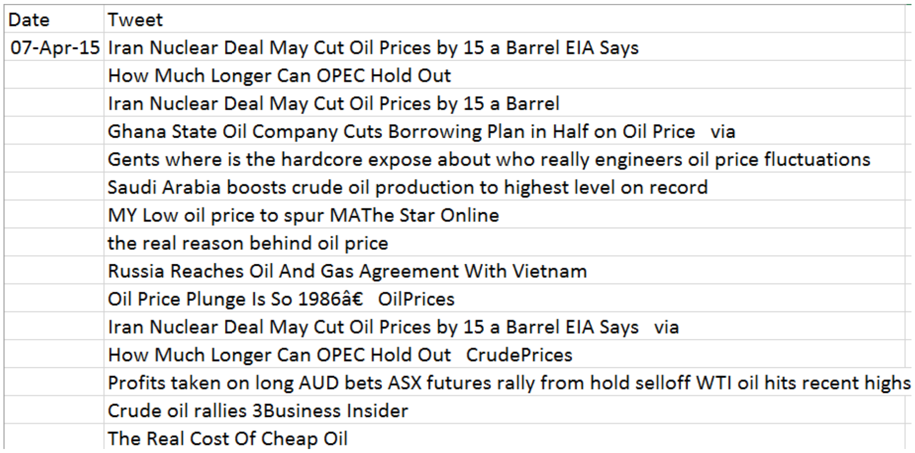
After gathering the data was cleaned in MS word using Macros programming the data for a particular day was cleaned and stored automatically.



Date	Tweet
07-Apr-15	Iran Nuclear Deal May Cut Oil Prices by $15 a Barrel, EIA Says http://dlvr.it/9HsV31
	How Much Longer Can OPEC Hold Out?
	Iran Nuclear Deal May Cut Oil Prices by $15 a Barrel
	Ghana State Oil Company Cuts Borrowing Plan in Half on Oil Price http://bloom.bg/1CS8so5 via @business
	.@BBCDouglasF @faisalislam Gents where is the hardcore expose about who really engineers oil price fluctuations?
	Saudi Arabia boosts crude oil production to highest level on record http://trib.al/Aj82ELO
	http://my.linkaloo.de #MY: Low oil price to spur M&A - The Star Online
	@guardian the real reason behind oil price
	Russia Reaches Oil And Gas Agreement With Vietnam
	Oil Price Plunge Is So 1986... http://mgstn.ly/1CkQTK5 #OilPrices
	Iran Nuclear Deal May Cut Oil Prices by $15 a Barrel, EIA Says http://bloom.bg/1HN2Ix7 via @business
	How Much Longer Can OPEC Hold Out? http://bit.ly/1CkVwDP #CrudePrices
	Profits taken on long $AUD bets, $ASX futures rally from hold sell-off, WTI oil hits recent highs but spread narrows
	Crude oil rallies 3% - Business Insider
	The Real Cost Of Cheap Oil

Figure 3 Data Cleaned from Twitter but contains unwanted elements

After Gathering the tweets it's cleaning is required which include removing hyperlinks, hashtags unwanted comma and punctuation the above data cleaning was done in R



Date	Tweet
07-Apr-15	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel EIA Says
	How Much Longer Can OPEC Hold Out
	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel
	Ghana State Oil Company Cuts Borrowing Plan in Half on Oil Price via
	Gents where is the hardcore expose about who really engineers oil price fluctuations
	Saudi Arabia boosts crude oil production to highest level on record
	MY Low oil price to spur MAThe Star Online
	the real reason behind oil price
	Russia Reaches Oil And Gas Agreement With Vietnam
	Oil Price Plunge Is So 1986â€ OilPrices
	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel EIA Says via
	How Much Longer Can OPEC Hold Out CrudePrices
	Profits taken on long AUD bets ASX futures rally from hold selloff WTI oil hits recent highs
	Crude oil rallies 3Business Insider
	The Real Cost Of Cheap Oil

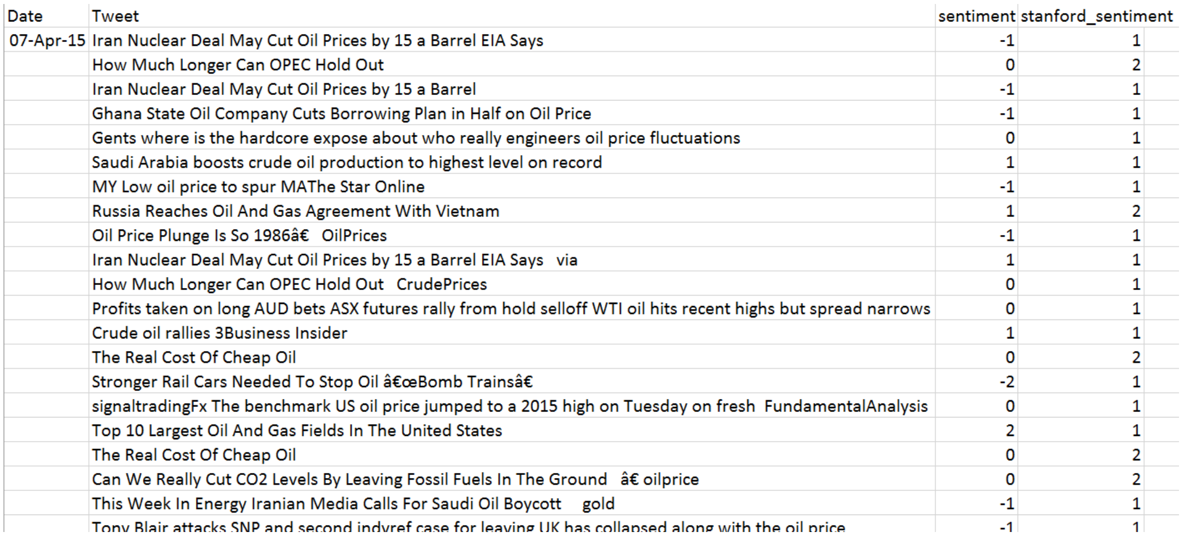
Figure 5 Data after Cleaning

Collection of Tweets ranged from 300 to 400 per day for the same keyword from **January 2016 – to January 2015** but weekends were excluded as oil price movement happens only in the week-days.

Using daily tweets obtain the sentiment classifier using StanfordNLP and NRC-Lexicon methodology

Next step was to get the sentiment for all the tweets for each day. This was done using two sentiment classifier the stanfordNLP classifier and the NRC Lexicon developed by Mohammed saif.

Both the classifier provide sentiment on different metrics the Stanford classifier provide sentiment on 3-point scale where 1 being negative 3 being positive and 2 being neutral whereas the NRC classifier give the sentiment on a 5-point rating scale from -2 to 2 where -2 being very negative to 2 being very positive and 0 being a neutral statement.



Date	Tweet	sentiment	stanford_sentiment
07-Apr-15	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel EIA Says	-1	1
	How Much Longer Can OPEC Hold Out	0	2
	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel	-1	1
	Ghana State Oil Company Cuts Borrowing Plan in Half on Oil Price	-1	1
	Gents where is the hardcore expose about who really engineers oil price fluctuations	0	1
	Saudi Arabia boosts crude oil production to highest level on record	1	1
	MY Low oil price to spur MAThe Star Online	-1	1
	Russia Reaches Oil And Gas Agreement With Vietnam	1	2
	Oil Price Plunge Is So 1986â€ OilPrices	-1	1
	Iran Nuclear Deal May Cut Oil Prices by 15 a Barrel EIA Says via	1	1
	How Much Longer Can OPEC Hold Out CrudePrices	0	1
	Profits taken on long AUD bets ASX futures rally from hold selloff WTI oil hits recent highs but spread narrows	0	1
	Crude oil rallies 3Business Insider	1	1
	The Real Cost Of Cheap Oil	0	2
	Stronger Rail Cars Needed To Stop Oil â€œBomb Trainsâ€	-2	1
	signaltradingFx The benchmark US oil price jumped to a 2015 high on Tuesday on fresh FundamentalAnalysis	0	1
	Top 10 Largest Oil And Gas Fields In The United States	2	1
	The Real Cost Of Cheap Oil	0	2
	Can We Really Cut CO2 Levels By Leaving Fossil Fuels In The Ground â€ oilprice	0	2
	This Week In Energy Iranian Media Calls For Saudi Oil Boycott gold	-1	1
	Tony Blair attacks SNP and second indyref case for leaving UK has collapsed along with the oil price	-1	1

Figure 6 Sentiment Along with the tweet

The average values of the daily sentiment were taken into consideration for all the days from January 2016 -January 2015

Development of Word cloud

Word cloud is the development of a graphic of the most commonly occurring word from any source material. The word cloud is used extensively in creation of the sentiment tracker to detect consumer sentiment or change in the sentiment on a real time basis. In a word cloud the size of word gives its relative importance or frequency of occurrence in the source material

The steps involved in development include: -

- Gathering the tweets
- Cleaning the data which involves removing commas, hashtags and removing any other commonly occurring words
- Convert all letter into lower case
- Removing the basic words like if, an, the etc.
- Measuring the frequency of the words occurring
- Use Geo-corpus package in R to create word-cloud

Developing Forecasting Model to Determine Ganger Causality

After getting the sentiment classifier a Forecasting model was developed to check. It included checking for stationarity in order to check for integration order and then checking for co-integration to establish whether a long run relation exists between the data or not and finally depending on the co-integration developing a vector auto-regressive or vector error correction model. After development of the model, granger causality was used to empirically detect the impact of sentiment on the price movement of oil price

Post development stability

Developing other Model to detect Accuracy.

To check the performance of the model developed using sentiment other model were also developed using Vector auto regressive moving average, ARIMA and performance of all the three models was checked by using the Root Mean Square Error and Mean Absolute Percentage Error method.

Empirical Analysis

Developing Multi-Variate Forecast Model.

After running the algorithm in R software the daily average score of the sentiment for both the sentiment classifier were tabulated and shown in figure7.

Table 1 Summary of Sentiment Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Sentiment_NRC	257	.2212435	.2105064	-1.052288	.3913043
Sentimet_stanford	257	1.279395	.1426786	1.065217	1.967014

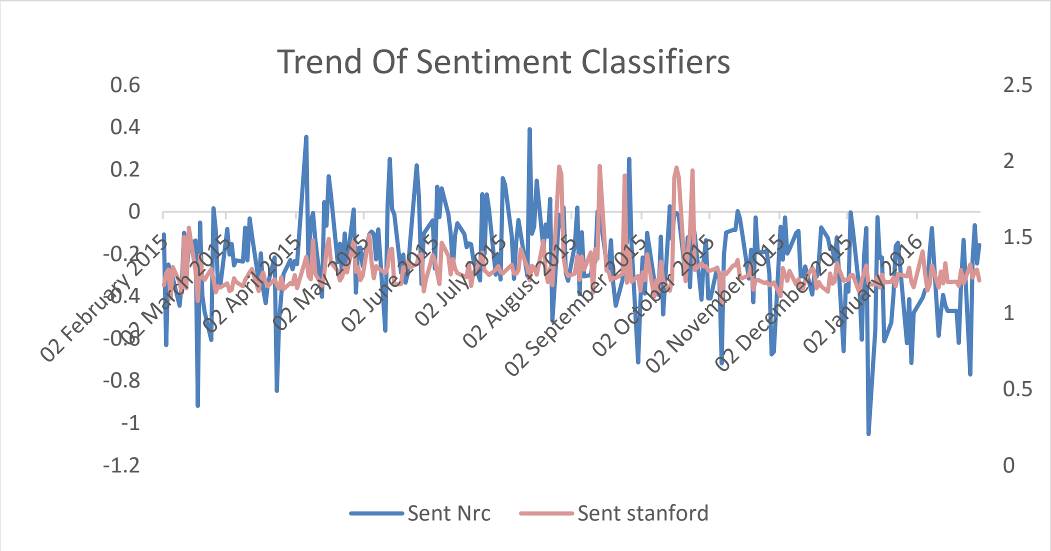


Figure 7 Time Graph of Sentiment models

Next step was to check for the stationarity in the series for the Sentiment Stanford. The method to check was to first visually inspect the ACF function to determine the lags to be incorporated and also by looking at the AIC criterion. Next the Augmented dickey fuller test was conducted to check for stationarity. The time graph clearly shows a constant but does not show a clear trend hence we run the ADF test with specification of no trend and for 9 lags.

Augmented Dickey-Fuller test for unit root			**Number of obs = 247**	
	Test Statistic	1% Value	5% Critical Value	10% critical Value
Sentiment Classifier	**-0.290**	**-2.580**	**-1.950**	**-1.620**
MacKinnon approximate p-value for Z(t) = 1.56				

Table 2 ADF Results for Sentiment classifier

The ADF test cannot clearly reject at 5% level of significance and hence does possess unit roots.

Next the difference in done and tested again using ADF test and the result clearly can be reject at 5 % level of significance hence the series only has one-unit root and hence can be said to be **integrated of the order (1)**

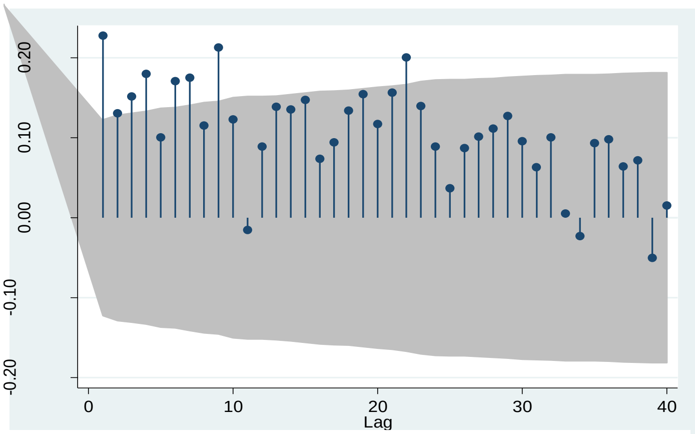


Figure 8 ACF plot of sentiment classifier

Table 2 ADF for differenced Sentiment Classifier

Augmented Dickey-Fuller test for unit root			**Number of obs = 247**	
	Test Statistic	1% Value	5% Critical Value	10% critical Value
Difference of classifier	-7.086	-3.461	-2.880	-2.570
MacKinnon approximate p-value for Z(t) = 0.000				

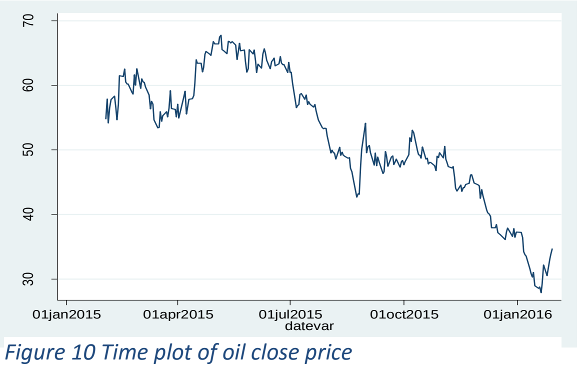


Figure 10 Time plot of oil close price

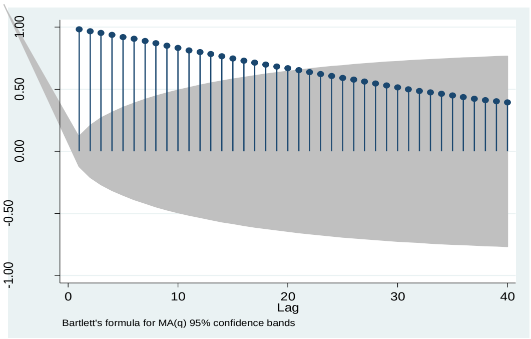


Figure 9 Time plot of ACF of oil close price

Now we analyze the daily close price of crude oil and gold price and check for stationarity. For oil price we can clearly see a negative trend being followed and ACF curves 16 significant lags. So running the Augmented dickey fuller with trend and lags of 15 we cannot reject the null hypothesis of no conclude the series is integrated order (1)

Augmented Dickey-Fuller test for unit root		**Number of obs = 247**		
	Test Statistic	1% Value	5% Critical Value	10% critical Value
Price of crude oil	-0.518	-3.460	-2.880	-2.570
MacKinnon approximate p-value for Z(t) = 0.884				
Difference in price of oil	-18.322	-3.460	-2.880	-2.570
MacKinnon approximate p-value for Z(t) = 0.00				

Table 3 ADF Test Result for oil price

Similarly, for gold price the analysis shows that the series is non-stationary with negative

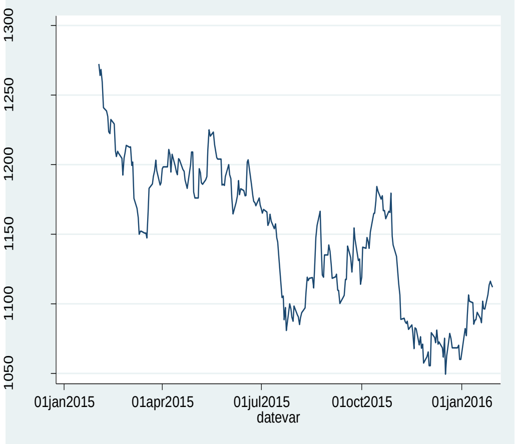


Figure 11 Time Plot of Gold Price

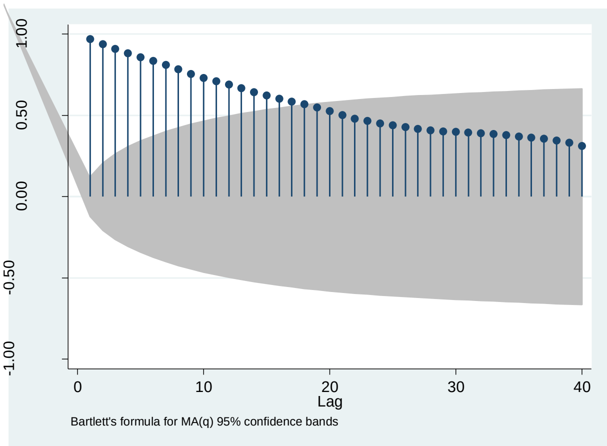


Figure 12 ACF for Gold price

Augmented Dickey-Fuller test for unit root			**Number of obs = 247**	
	Test Statistic	1% Value	5% Critical Value	10% critical Value
Price of gold	-2.459	-3.460	-2.880	-2.570
MacKinnon approximate p-value for Z(t) = 0.1259				
Difference in price gold	-15.609	-3.460	-2.880	-2.570
MacKinnon approximate p-value for Z(t) = 0.00				

Table 4 ACF test for Gold Price

As the Augmented Dickey Fuller test at 5% significance level cannot be rejected and running the test for differenced value we find the series to be integrated for order (1)

Now in order to Determine where there is a long run relationship between the three variable we have to test for co-integration and find the number of co-integrating vector

But before running the test we need to find the lags that need to be included in the model STATA gives a command by which the lags are selected based on AIC, BIC and Q. hence running the model we find the number **of lags significant to be 1 gives the minimum AIC , HQIC and SBIC**

Sample: 5 - 257					Number of obs = 253			
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	2025.57				1849.24	16.0362	16.053	16.0781
1	**1226.51**	**1598.1**	**9**	**0.000**	**3.58582***	**9.79061***	**9.85804***	**9.9582***
2	1219.14	14.739	9	0.098	3.63246	9.8035	9.9215	10.0968
2	1219.14	14.739	9	0.098	3.63246	9.8035	9.9215	10.0968
4	1207.29	17.722*	9	0.039	3.81435	9.85213	10.0713	10.3968

Table 5 Optimum Lag Selection for Co integration

Next we test for co-integration. there are two methods for checking the co-integration the ARDL method is used when the variable has both integration order 1 and 0 and the Johannsen test for co-integration is used when all the variables are of the same integration order since all the variables are order one we use **Johannsen method**

Sample: 3-257				**Lags = 2**	
Maximum rank	parms	LL	eigenvalue	Trace statistic	5% critical value
0	12	1277.8308	.	90.9486	29.6
1	17	-1240.734	0.25245	16.7550	15.41
2	**20**	**1232.9836**	**0.05898**	**1.2541***	**3.76**

Table 6 Result for Johannsen Test for Co-integration

The Null hypothesis of greater than 2 co-integration equation can be rejected at 5% level of significance hence the result clearly shows that there exist at least 2 co-integrating equation. Hence some long run relationship does exist between the said variables

Due to the presence of co-integration the best method for developing the model chosen will be Vector Error Correction Model which provides both Short run and long run estimates.

Developing Uni-variate Forecast Model

In order to check the efficiency of the said model an ARIMA as well as VARMAX (Vector Auto-regressive Moving Average X) were also developed for comparison.

The lag selection for AR and MA part was done by looking and the Partial Auto-correlation and auto-correlation function graph respectively of the differenced variable (as the series is integration order 1)

From visual inspection we can see that that only AR component seems to consist of 1 lag and MA component also consist of lag 1

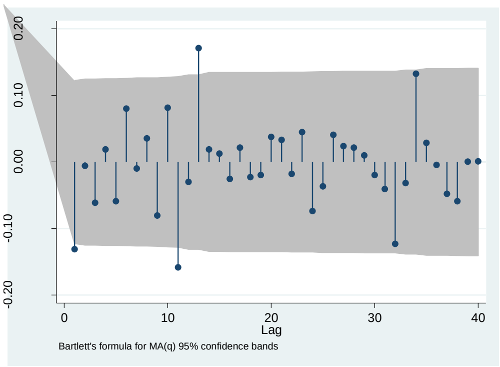


Figure 14 ACF for Differenced oil price

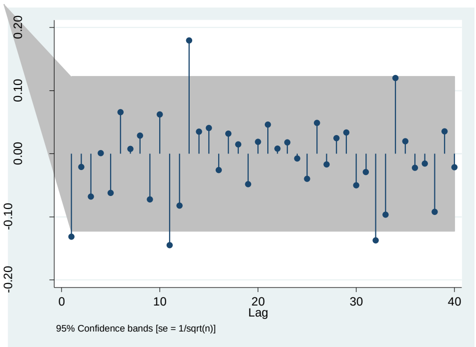


Figure 13 PACF for Differenced oil price

Variables	**Arima_00**	**Arima_11**	**Arima_10**	**Arima_01**	**Arima_12**
Price	-.07247	-.07613	-.749	**-.750**	-.763
AR L1		.449	-.1431		.542
AR L2					
MA L2		-.598		**-.155292****	-.696
MA L2					.029
AIC	847.024	844.90	844.2805	**843.90**	846.60
BIC	853.935	858.2682	854.55705	**854.268**	863.882

Table 7 Selection from Various Combination of ARIMA

Next we compare the various model's of ARIMA and use AIC and BIC to select the model. The model which minimizes AIC and BIC is choosen as the best model.

From the above table 10 we select ARIMA (0,1,1) as this gives the least AIC and BIC criterion. Following a similar process, we develop the VARMAX process using the same principle we find the VARMAX component as AR =1, MA =1 and include gold price as an independent variable.

RESULTS

The VECM Results are as follows: -

Equation	Parms	RMSE	R-sq	chi2	P>chi2
D_priceoil	5	1.4289	0.0384	9.974277	0.0760
D_sentstanford	5	.136188	0.3584	139.6741	0.0000
D_goldprice	5	9.88859	0.0218	5.572989	0.3500

Table 8 Overall statistics of VECM model

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Oil Price						
Ce L1.	-.0058517	.0038509	-1.52	0.129	-.0133995	.001696
priceoil LD.	-.1239961	.062072	-2.00	0.046	-.2456549	-.0023373
Sentiment LD.	-1.356431	.6604307	-2.05	**0.040**	-2.650852	-.062011
Goldprice LD.	-.0026124	.0091638	-0.29	0.776	-.0205731	.0153483
Sentiment Stanford						
Ce L1.	.0032213	.000367	8.78	0.000	.0025019	.0039407
Priceoil LD.	-.0058086	.0059161	-0.98	0.326	-.0174039	.0057867
Sentiment LD.	-.0586689	.0629454	-0.93	**0.351**	-.1820396	.0647018
Goldprice LD.	.000237	.0008734	0.27	0.786	-.0014748	.0019489

Table 9 Coefficient of VECM model

The table 12 gives us the short run as well long run coefficient estimates for the VECM model we see that both the lagged value of oil as well sentiment are significant at 5% level of significant whereas the long run coefficient in the case of oil price comes out to be statistically significant

In case of equation of sentiment, we find that the lagged value of price oil and sentiment come out to be insignificant at 5% level of significance the VECM can further be used to check for granger causality.

Prior to causality some test need to run the check the stability of the model

First we the check the stability of the estimates by using Eigen value stability condition to check for misspecification bias if any in the model.

The Eigen value are all less than 1 which indicates that's there is no miss-specification bias in the model

Eigenvalue	Modulus
1	1
1	1
3738504	.37385
-.2381011	.238101
.0146765 +. 0346412i	.03722
0146765 - .0346412i	.03722

Table 10 Stability test of VECM Estimates

Next we carried out LM test to check for auto-correlation in the estimates. Ho: of no autocorrelation could not rejected at 5% level of significance indicating absence of any auto-correlation in the model

lag	chi2	df	Prob
1	7.9245	9	0.54177
2	3.8479	9	0.92113

Table 11 Lm test for auto-correlation

Checking for granger causality

The Concept of granger causality implies that if the lag value of independent variables are jointly significant then the variable is said to granger cause the dependent variable and similarly another equation is created with the dependent variable becoming the independent and vice-versa. If only one set of variables are significant the causative nature is said to be uni-directional and if both are jointly significant then there is unidirectional causality and more analysis needs to be done.

So granger causality can be detected for the VECM model by analyzing the lagged value in both the equation to establish a causality direction. Hence the table gives the direction as follows.

Oil Price	**z**	**P>z**
Overall	1.4289	0.0384
Sentiment LD.	-2.05	**0.040**
Sentiment Stanford		
Overall	.136188	0.3584
Priceoil LD.	-0.98	0.326
Sentiment LD.	-0.93	**0.351**

Table 12 Granger Causality Analysis

Here we see that the sentiment does granger cause oil price and oil price does not granger cause sentiment hence we see that there is by directional causalit

Word Cloud

The Adjoining image gives the word cloud for the “improvement”, “rise”, “confidence”, “rebound” being given relative importance which shows that the market was optimistic and also reflected in the rise of from previous close price. This provide an indirect method to

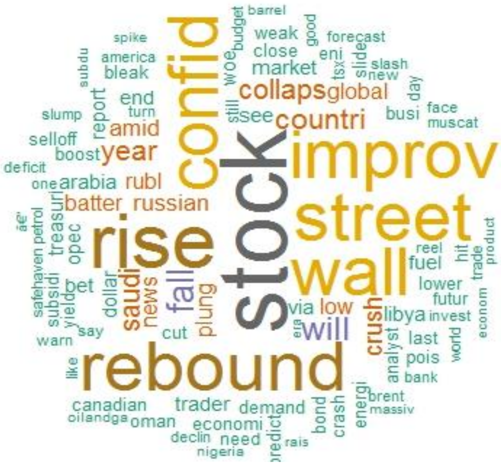


Figure 15 word cloud for 31st December

To view and analyze the sentiment at any time of the Day. similarly, the word cloud for 11 December gives relative importance to “risk” ,“low” ,”fall”,”tumble” ,”slump” etc which clearly show a negative sentiment and was also reflected in the days close which closed at a significantly lower price.

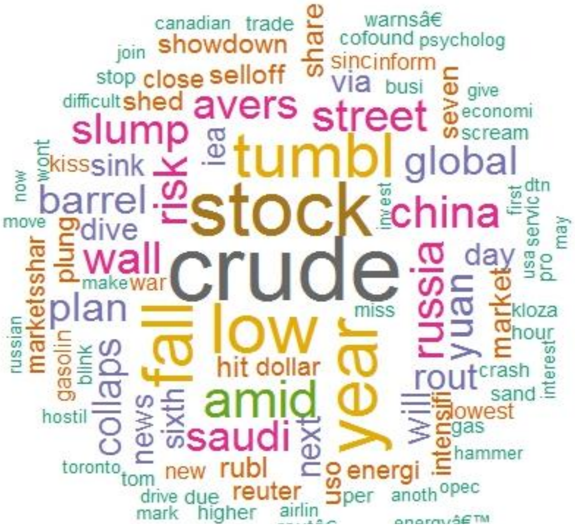


Figure 16 word -cloud for 11 December

Forecasting from Model

The forecasting of the model was done using the out-of-sample 22-day one-step forecast and the accuracy was compared using the following metrics: -

RMSE = Square root of {summation of (actual value -predicted value)^2/Number of Forecast}

MAPE = Summation of {|actual value -predicted value|/actual value}

Two metrics have been used to nullify the short-comings present in each such as Mape limitations is that it gives more weightage to negative error while the RMSE gives more emphasis on larger error as compared to smaller error’s

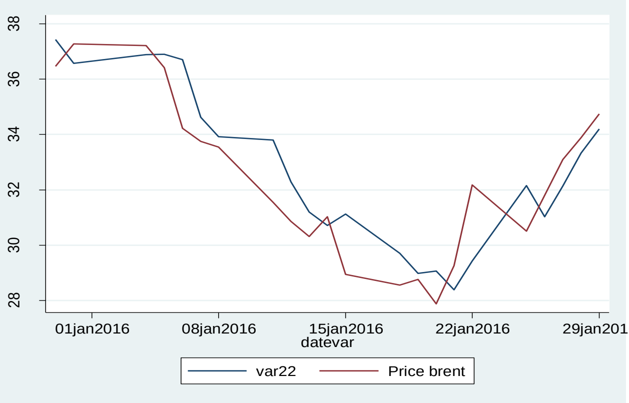


Figure 17 Forecast 22 day using VEC model

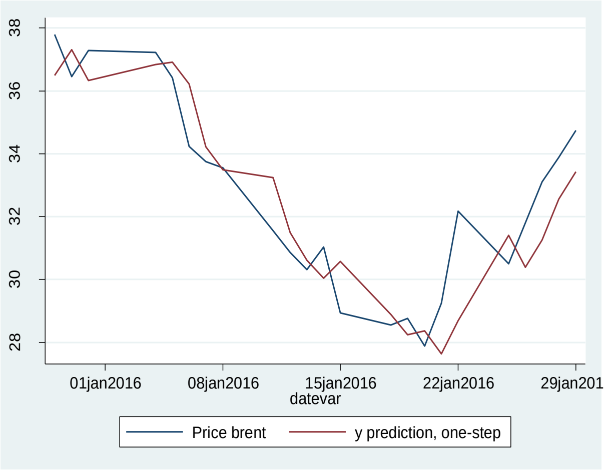


Figure 18 22-day forecast using ARIMA (0,1,1)

Hence forecast for all the three models were made for the period 1 January 2016 to 29th January 2016The results in table show the MAPE and RMSE scores for all the three forecast and we see that the VEC model scores the least on the both the metrics

\

MODEL	**MAPE**	**MSE**
ARIMA	.7824196	2.102205
VARMA	.7689778	2.0756647
VEC	.521535	1.79

Table 13 Result of out of sample forecasts accuracy

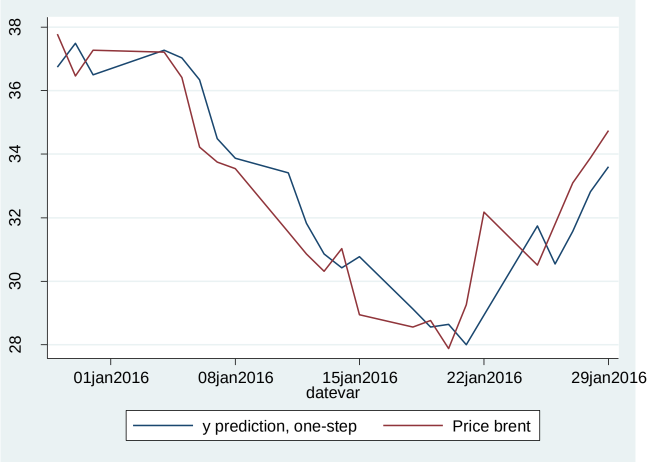


Figure 19 22-day forecast using VARMAX Model

ANALYSIS OF THE RESULTS

The time series plot Figure 5 of sentiment using NRC lexicon shows mainly negative response except for few spikes which clearly a general negative sentiment appeared throughout the year which corroborates with the trend followed by oil price as it depreciated to almost half its value.

The VEC model was developed and the optimum lag for the sentiment variable came out to be one. Which was according to theory as the sentiment should effect the price in a long duration as expectation tend to change quickly and so does people's behavior and reaction towards.

Also the co-integration within the 3 variables suggest a long run relationship which can attributed that theoretically the price of oil must always correct itself to reflect the sentiment/expectation of the people and which is being empirically demonstrated.

The results of the granger causality test were also in line with expectation that sentiment causes change in price movement but vice versa is not true hence there Is only uni-direction causality as expected. Also the long run estimates of oil price came out to be insignificant.

The accuracy for forecast compared using MAPE and RMSE showed the VEC model giving the minimum score on both metrics. Hence gave the bet prediction within these models

CONCLUSION

The project aims was to measure expectation by building a sentiment classifier using sentiment Analysis on the data available from twitter and use it to forecast oil price. The project involved building the sentiment classifier on a scale matrix using the methodology adopted by the Stanford program and NRC Lexicon to build the sentiment classifier for daily tweets from 29th January 2016 to 1st January 2015. Further a word-cloud was developed which is a visual representation of the sentiment on a real basis for crude sentiment analysis. Apart from building the classifier the average sentiment variable was used to develop a multivariate forecasting model including the price of gold. The model was then used to test the causality between sentiment and price movement of oil and detect the direction. Further the model's forecasting performance was also checked with other uni-variate model using RMSE and MAPE Score.

The sentiment classifier based on a rating scale of 5 ranging from -2 to 2 gave mostly negative results which corroborated with the near half depreciation in the value of oil. Further granger causality was uni-directional with sentiment classifier causing the change in price of oil in the short run. Further the inclusion of the sentiment classifier also helped reduce its MAPE and RMSE score thereby increasing its accuracy.

Hence the model shows great promise and can be improved further by incorporating more macro-economic variables in the proces

FUTURE SCOPE OF WORK

The Analysis in the present context was done for around 1 year which can be expanded to generate more precise estimates. Also the classifier developed can be improved further by incorporating deep learning algorithm for better prediction of sentiment. Also new innovative statistical and machine learning techniques such as Kalman filters, Artificial Neural nets, support Vector Machine can be used make better predicting model and test their accuracy with the current contemporary models.

BIBLOGRAPHY

1. Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
2. Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 3859–3864, New York, NY, USA. ACM.
3. Edelstein, P., and L. Kilian (2009), "How Sensitive Are Consumers to Retail Energy Prices?" Journal of Monetary Economics, 56, 766-779.
4. J Kumar, T Rao, S Srivastava - Big Data Analytics, 2012 – Springer
5. Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood Detects the stock market. Journal of Computational Science, 2(1), March 2011, Pages 1-8,
6. Niaz Bashiri Behmiri and Manso José R. Pires. Working paper Alternative Investment Analyst Review
7. SERDA SELIN OZTURK A Sentiment Analysis of Twitter Content as a Predictor of Exchange Rate Movements Review of Economic Analysis 6 (2014) 132-140
8. Saif M. Mohammad, Emotion Measurement Sentiment Analysis: Detecting Valence, Reactions, and Other Affectual States from Text., 2016.
9. Zhang X , H Fuehres, PA Gloor - Advances in Collective Intelligence 2011, 2012 – Springer

Website and software reference

10. http://nlp.stanford.edu/
11. R packages and manual on CRAN Repositories
12. Help file of stata 11
13. www.wikipedia.com/Moore's%law.php
14. www.r-bloggers.com

LIST OF FIGURES AND TABLES

APPENDIX:

Code for Extraction of twitter data from website written in VBA

```
Sub Macro1()
FTW = Worksheets("Sheet1").Cells(2, "A").Value
While Index < 2000
Index = Index + 1
If InStr(1, ActiveCell, "View summary", 1) > 0 Then
ActiveCell.Rows("1:13").EntireRow.Select
Selection.Delete Shift:=xlUp
ActiveCell.Offset(0, 1).Range("A1").Select

ElseIf InStr(1, ActiveCell, "View conversation", 1) > 0 Then
ActiveCell.Rows("1:12").EntireRow.Select
Selection.Delete Shift:=xlUp
ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf InStr(1, ActiveCell, "retweets", 1) > 0 Then
ActiveCell.Rows("1:12").EntireRow.Select
Selection.Delete Shift:=xlUp
ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf InStr(1, ActiveCell, "retweet", 1) > 0 Then
   ActiveCell.Rows("1:13").EntireRow.Select
   Selection.Delete Shift:=xlUp
   ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf InStr(1, ActiveCell, FTW, 1) > 0 Then
    ActiveCell.Rows("1:1").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf InStr(1, ActiveCell, "added,", 1) > 0 Then
    ActiveCell.Rows("1:2").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf InStr(1, ActiveCell, "In reply to", 1) > 0 Then
    ActiveCell.Rows("1:1").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf IsEmpty(ActiveCell) Then
    ActiveCell.Rows("1:1").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf ActiveCell = "More" Then
    ActiveCell.Rows("1:1").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
ElseIf ActiveCell = "View photo" Then
    ActiveCell.Rows("1:1").EntireRow.Select
    Selection.Delete Shift:=xlUp
    ActiveCell.Offset(0, 1).Range("A1").Select
Else
   ActiveCell.Offset(1, 0).Range("A1").Select
   End If
   Wend
End Sub

Sub Macro5()
'
' Macro5 Macro
FTW = Cells(2, "A").Value
Final = "C:\Users\Puneet\Desktop\Data New\" & FTW & ".xlsx"
    Application.Run "'15 sepxls.xlsm'!Macro1"
    Range("B2").Select
    Application.Run "'15 sepxls.xlsm'!Macro2"
    ActiveWorkbook.SaveAs Filename:=Final, FileFormat:=xlOpenXMLWorkbook _
        , CreateBackup:=False
    'ActiveWorkbook.SaveAs Filename:= _
  Cells(3, "A").Value = FTW
Range("F2").Select
    ActiveCell.FormulaR1C1 = "=RIGHT(R[52]C[-4],11)"
    Range("E2").Select
    ActiveCell.FormulaR1C1 = "=RIGHT(R[51]C[-3],11)"
End Sub
```

Code for Getting Sentiment Score written in R

```
library(coreNLP)
library(xlsx)
library(plyr)
library(wordcloud)
library(RColorBrewer)
library(SnowballC)
data<-data.frame(date=character(),sentiment=numeric(),stanford_sentiment=numeric())
dummy5<- character(0)
dummy6<-numeric(0)

init
setwd("C:/Users/Puneet/Documents/Sentiment Analysis Project/New")
files=list.files()
#getting nrc sentiment
for(j in 1:length(files))
{
  dum<-paste(filrd,".xlsx",sep="")
  Dummy<- read.xlsx(dum,1)
  dummy2<- as.character(Dummy[,2])
  dummy3<- as.character(Dummy[,1])
  dummy4<- gsub("@\\w+","",dummy2)
  dummy4<- gsub("[[:punct:]]","",dummy4)
  dummy4<- gsub("http\\w+","",dummy4)
  dummy4<- gsub("[ \t]{2,}"," ",dummy4)
  dummy4<- gsub("^\\s+|\\s+$"," ",dummy4)
  Dummy$sentiment<-get_sentiment(char_v=dummy4,"nrc",dummy4)
  dummy5[j]<-dummy3[2]
  data[j,2]<-mean(Dummy$sentiment)
}
data$sentiment<-dummy5
#sentiment from stanford sentiment
initCoreNLP()
for(j in 1:68)
{
  dum<-paste(j,".xlsx",sep="")
  Dummy<- read.xlsx(dum,1)
  dummy2<- as.character(Dummy[,2])
  dummy3<- as.character(Dummy[,1])
  dummy4<- gsub("@\\w+","",dummy2)
  dummy4<- gsub("[[:punct:]]","",dummy4)
  dummy4<- gsub("http\\w+","",dummy4)
  dummy4<- gsub("[ \t]{2,}"," ",dummy4)
  dummy4<- gsub("^\\s+|\\s+$"," ",dummy4)

  for(i in 1:length(dummy4))
  {

    abc<-annotateString(dummy4[i])
    a<-getSentiment(abc)
   dummy6[i]<-a$sentimentValue[1]

  }
  Dummy$stanford_sentiment<-dummy6
  dummy5[j]<-dummy3[2]
  data[j,2]<-mean(Dummy$stanford_sentiment)
}
```

Code for Generating Word- Cloud written in R

```
library(syuzhet)
library(xlsx)
library(plyr)
setwd("C:/Users/Puneet/Documents/Sentiment Analysis Project/Data")
month<-c("Jan","Nov","Dec")

data<-data.frame(date=character(),sentiment=numeric())
dummy5<- character(0)
for(j in 1:68)
{
  dum<-paste(j,".xlsx",sep="")
  Dummy<- read.xlsx(dum,1)
  dummy2<- as.character(Dummy[,2])
  dummy3<- as.character(Dummy[,1])
  dummy4<- gsub("@\\w+","",dummy2)
  dummy4<- gsub("[[:punct:]]","",dummy4)
  dummy4<- gsub("[[:digit:]]","",dummy4)
  dummy4<- gsub("http\\w+","",dummy4)
  dummy4<- gsub("[ \t]{2,}"," ",dummy4)
  dummy4<- gsub("^\\s+|\\s+$"," ",dummy4)
  Dummy$sentiment<-get_sentiment(char_v=dummy4,"nrc",dummy4)
  dummy5[j]<-dummy3[1]
  data[j,2]<-mean(Dummy$sentiment)
}
data[,1]<-dummy5
a<-VectorSource(dummy4)
jeopCorpus <- Corpus(a)
jeopCorpus <- tm_map(jeopCorpus, PlainTextDocument)
jeopCorpus  <- tm_map(jeopCorpus, content_transformer(tolower))
jeopCorpus <- tm_map(jeopCorpus, removeWords, c("price", "oil","oilprice","prices"
                                                ,"drop","oilpric", stopwords('english')))
#Eliminate extra white spaces
jeopCorpus <- tm_map(jeopCorpus, stripWhitespace)
jeopCorpus <- tm_map(jeopCorpus, stemDocument)

dtm <- TermDocumentMatrix(jeopCorpus)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 30)

wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```