News Sentiment Analysis on Stock Prices

Group 2

Meghana Kasula

Pruthvi Nagaraju

Shaloma Ghosh

Vinit Gupta

Weining Zheng

Contents

Executive Summary

- Stock Prices are often used as an important indication of overall strength and health of a company. In general if the stock prices continues to climb, the company is said to be growing.
- We analyze the stock prices and observe that if everyone is happy, especially the customers, there is a likelihood that the stock prices will increase or at least not decrease.
- This is where the integration of fundamental and technical analysis seems inevitable. Fundamental analysis takes the underlying causes in price shift into consideration and technical causes taking previous stock prices into consideration.

Project Objective

- Public mood on news/twitter about a company is representative of overall sentiment about the respective company.
- The overall sentiment about a company can have a direct or indirect impact on its stock prices. As a result, analyzing sentiments along with the traditional stock prices forecasting is a very important aspect.
- One of the biggest source of extracting public reaction to a company or its product or any stock price related variables are Twitter, Google trends, Reddit, etc.
- Time series analysis is the most common and fundamental method used to perform this task.
- Our project aims to combine the conventional time series analysis with news from Reddit to predict weekly changes in stock price.

Data Set Details

- Our Data set was taken from the website: https://www.kaggle.com/aaron7sun/stocknews
- There are two channels of data provided in this dataset:
  1. News data: It is the historical news headlines crawled from Reddit World News Channel . They are ranked by Reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)
  2. Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01)
- The Stock data has 6 variable namely, open price, close price, high price, low price, volume and adjusted close price.
- The News Data has the top 25 news for that day as ranked by Reddit users.

Data Preprocessing – Stock prices

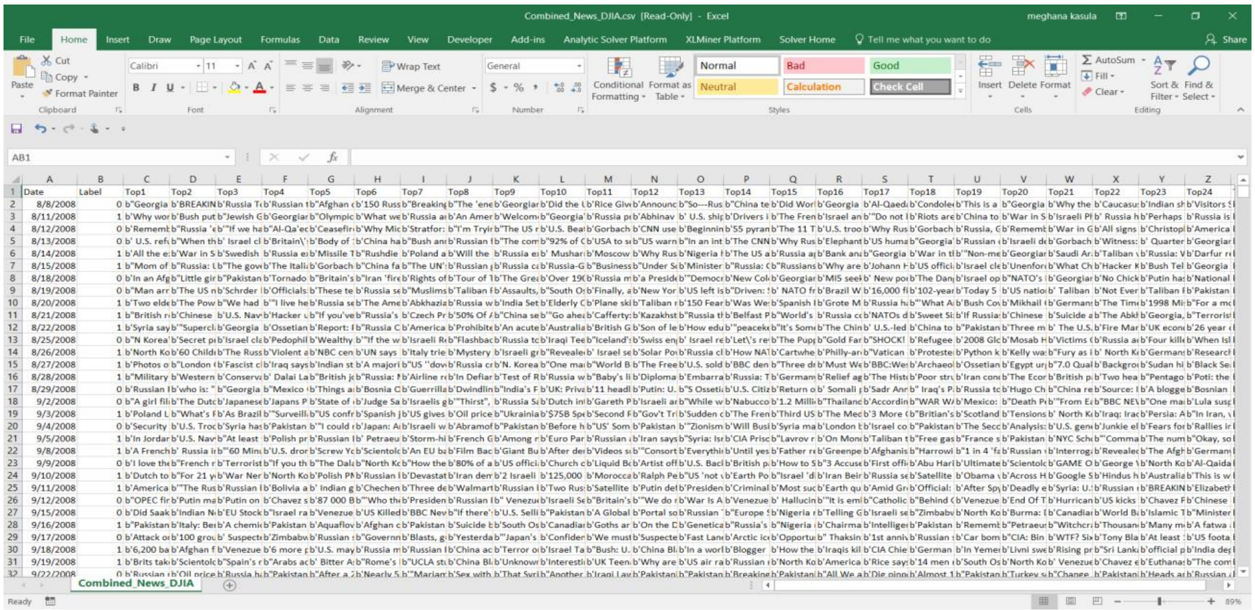
In data consisting of stock volume traded and the monthly DJIA trend, since the numerical value of the volume data was very high, we did log transformation of it to easy analysis.

Below is the screenshot of the data after taking log transformation

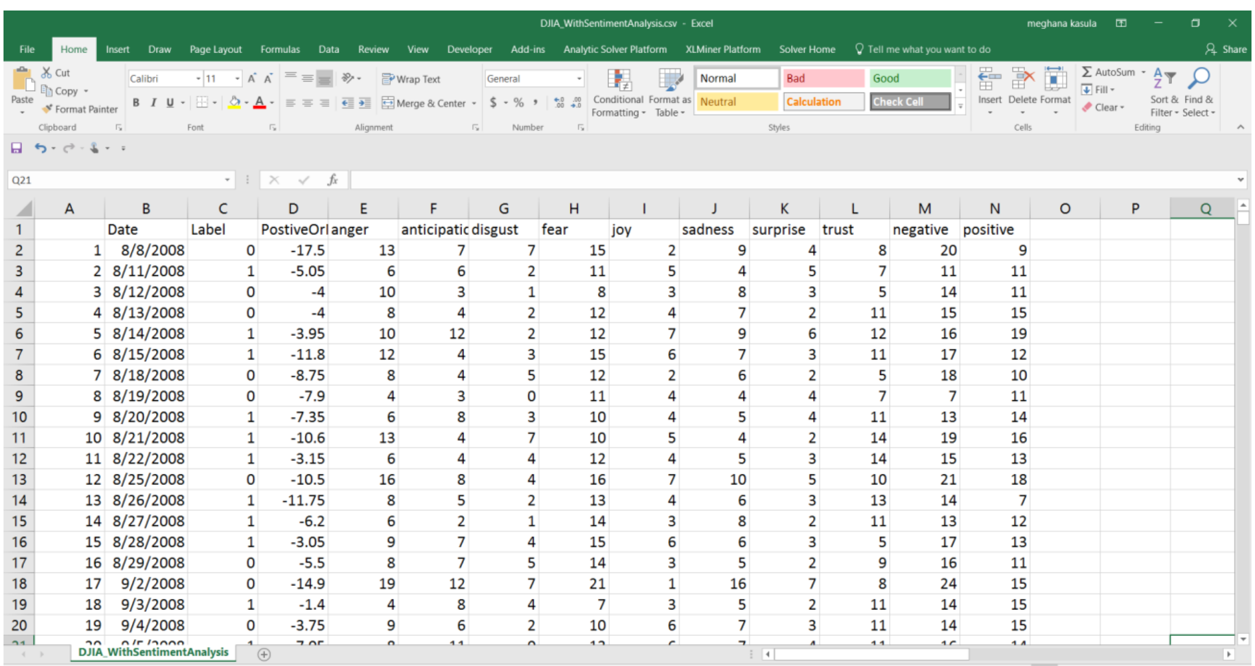
Date	Log-volume	DJIA-Trend
1/1/2009	169.6276785	57.5
2/1/2009	162.6223986	74.5
3/1/2009	190.8132166	100
4/1/2009	179.7340123	72
5/1/2009	170.3791548	54.5
6/1/2009	184.4933673	51
7/1/2009	183.476448	48.5
8/1/2009	173.8289629	47
9/1/2009	174.6467819	43
10/1/2009	183.9273714	40.5
11/1/2009	165.4728293	36
12/1/2009	181.0328085	29.5
1/1/2010	158.7686308	30
2/1/2010	158.5472318	32.5
3/1/2010	190.156725	28.5
4/1/2010	174.1601127	30.5
5/1/2010	168.6200503	48.5
6/1/2010	183.428946	37.5
7/1/2010	174.2254229	36
8/1/2010	181.8424556	34.5

Sentiment Analysis (News Data)

- The News Data has a total of 27 columns with Data and label and 25 top news included.
- Following is the screenshot of the raw data.



- Following is the dataset after preprocessing (performing sentiment analysis for each day)



	Date	Label	PostiveOrI	anger	anticipatic	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	8/8/2008	0	-17.5	13	7	7	15	2	9	4	8	20	9
2	8/11/2008	1	-5.05	6	6	2	11	5	4	5	7	11	11
3	8/12/2008	0	-4	10	3	1	8	3	8	3	5	14	11
4	8/13/2008	0	-4	8	4	2	12	4	7	2	11	15	15
5	8/14/2008	1	-3.95	10	12	2	12	7	9	6	12	16	19
6	8/15/2008	1	-11.8	12	4	3	15	6	7	3	11	17	12
7	8/18/2008	0	-8.75	8	4	5	12	2	6	2	5	18	10
8	8/19/2008	0	-7.9	4	3	0	11	4	4	4	7	7	11
9	8/20/2008	1	-7.35	6	8	3	10	4	5	4	11	13	14
10	8/21/2008	1	-10.6	13	4	7	10	5	4	2	14	19	16
11	8/22/2008	1	-3.15	6	4	4	12	4	5	3	14	15	13
12	8/25/2008	0	-10.5	16	8	4	16	7	10	5	10	21	18
13	8/26/2008	1	-11.75	8	5	2	13	4	6	3	13	14	7
14	8/27/2008	1	-6.2	6	2	1	14	3	8	2	11	13	12
15	8/28/2008	1	-3.05	9	7	4	15	6	6	3	5	17	13
16	8/29/2008	0	-5.5	8	7	5	14	3	5	2	9	16	11
17	9/2/2008	0	-14.9	19	12	7	21	1	16	7	8	24	15
18	9/3/2008	1	-1.4	4	8	4	7	3	5	2	11	14	15
19	9/4/2008	0	-3.75	9	6	2	10	6	7	3	11	14	15

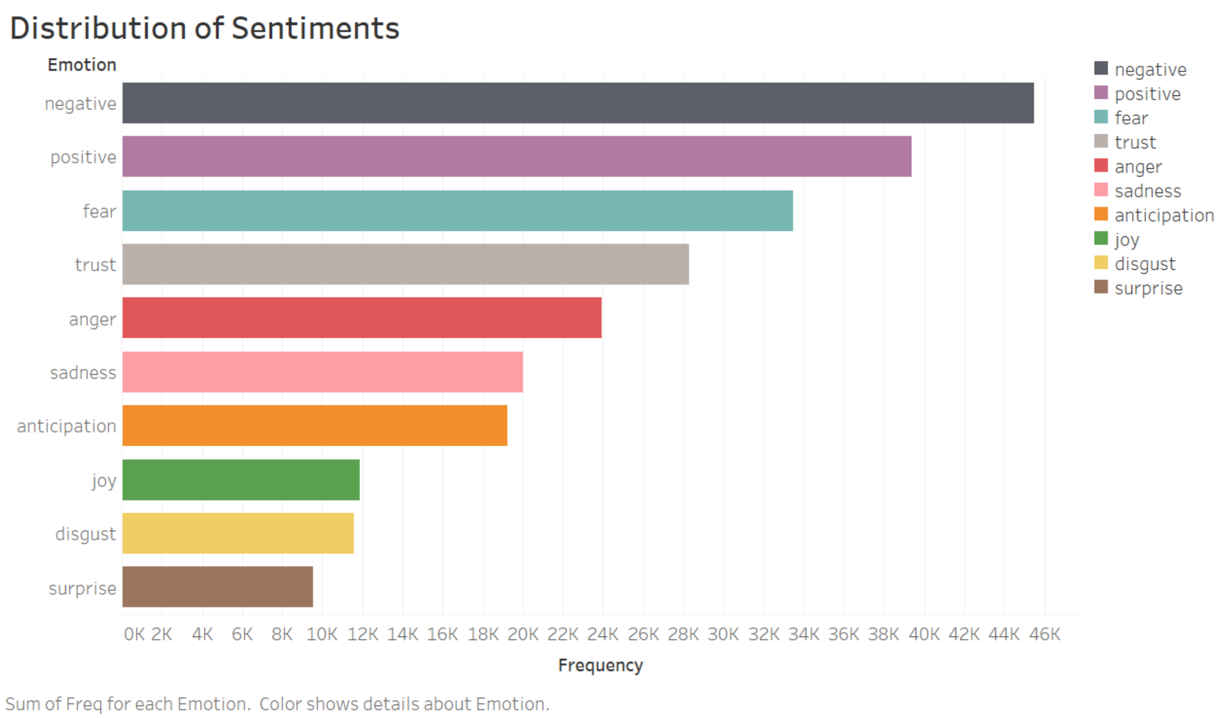
- The steps involved in processing it are as follows:
  - Combined all 25 news Data in one column.
  - Removed punctuation, control characters like \b,\t etc and digits.
  - Used a package called “Syuzhet” on R to do an sentiment analysis of the combined news
  - Used two Functions like ‘get_sentiment’ (meant for positive and negative sentiment) and ‘get_nrc_sentiment’ (meant for emotions like Trust, joy, anger, sadness etc.) to get sentiments of combines news each day.
  - We have now 11 sentiments namely - Positive or Negative (sign indicated both), Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Negative and Positive.
  - We will be using all sentiments to explore the news data on stocks
  - We will be only using Positive or Negative Dataset for forecasting since it can store information of all kinds of news.

Following is the code we used to perform sentiment analysis

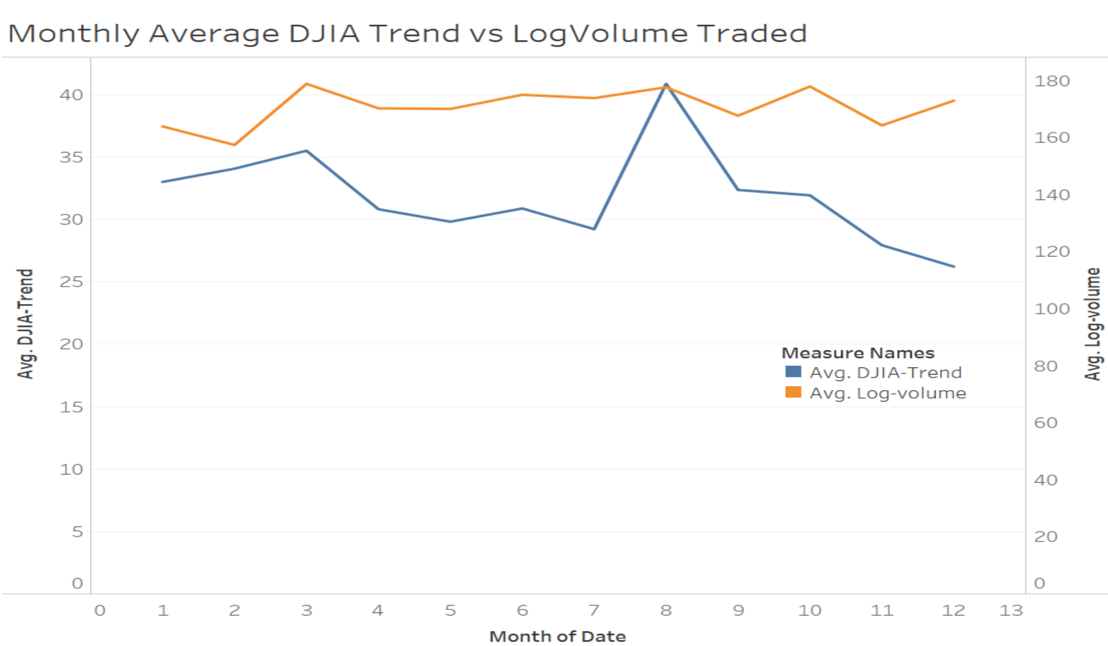
```r
library(syuzhet)
library(tidytext)
library(dplyr)
library(tidyr)

#Loading data
topNews=read.csv("C:/Users/pru/Desktop/Data mining/Stock Forecadting project/Combined_News_DJIA.csv")

#converting data.frame into character vector
s1=data.frame(lapply(topNews, as.character), stringsAsFactors=FALSE)

#combining top 25 news data
c1=c=do.call(paste, c(s1[3:27], sep = ""))

#removing punctuation using global sub
c1 = gsub('[[:punct:]]', '', c1)
#removing control characters
c1 = gsub('\\d+','',c1)
#removing digits
c1 = gsub('[[:cntrl:]]', '', c1)

#get_sentiment function gives the overall positive or negative sentiment
sentiment <- get_sentiment(c1,method="syuzhet")
sentiment

s1$PostiveOrNegative=get_sentiment(c1,method="syuzhet")

dictionary=get_sentiment_dictionary(dictionary = "syuzhet")

#get_nrc_sentiment is used to get eight emotions
#(anger, fear, anticipation, trust, surprise, sadness, joy, and disgust)
#and two sentiments (negative and positive)

nrc = get_nrc_sentiment(c1)
View(nrc)

# adding all the sentiment columns to the original dataset
f1=cbind(s1,nrc)
View(f1)

write.csv(f1, file = "DJIA_WithSentimentAnalysis.csv")
```

Data exploration

- Frequency of all the emotions was calculated and it was observed that overall, the frequency of negative emotion was the highest.

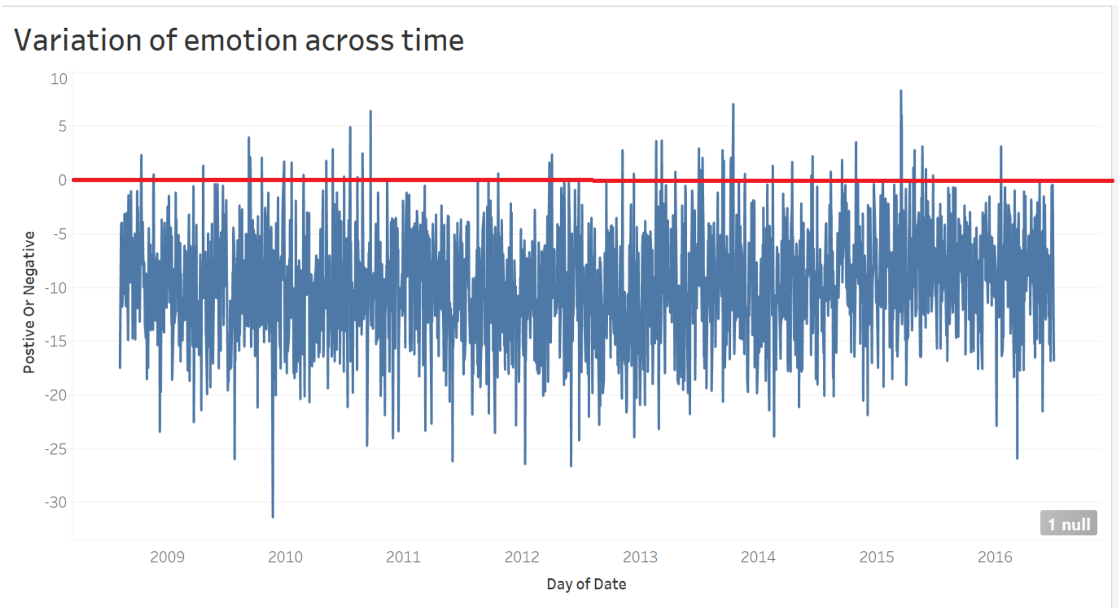


- The graph below shows the correlation between volume of the stocks and the DJIA



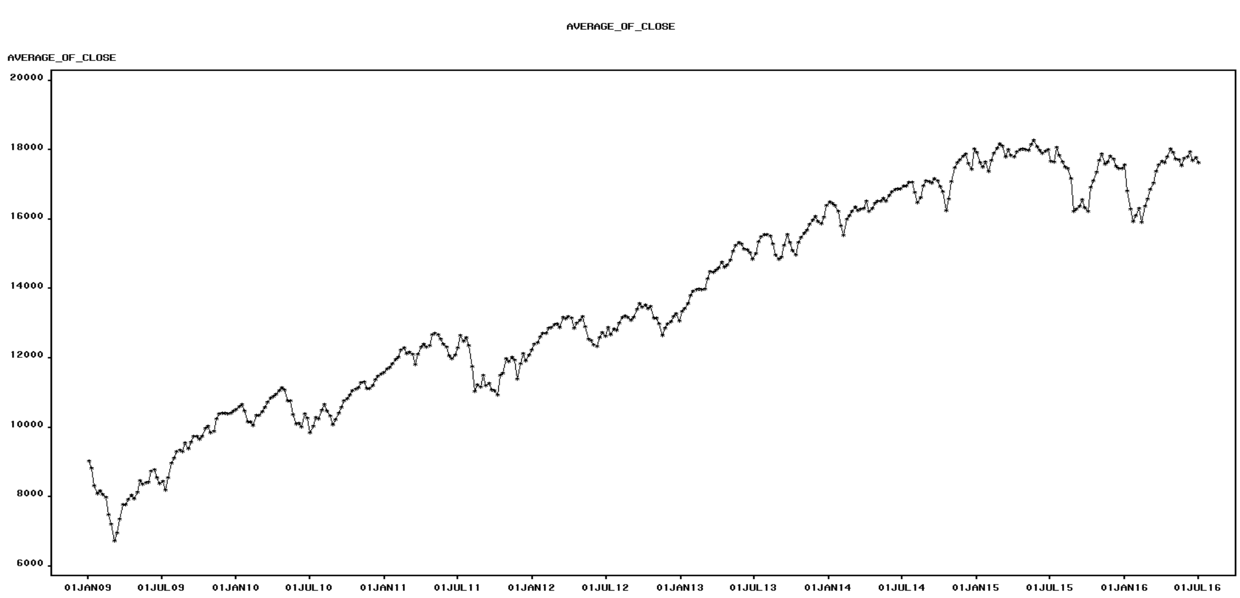
trend. Peaks in the DJIA trend coincides with the peaks in the stock volume at most places.

- Most of the sentiments are Negative, this shows that the news data could be biased.

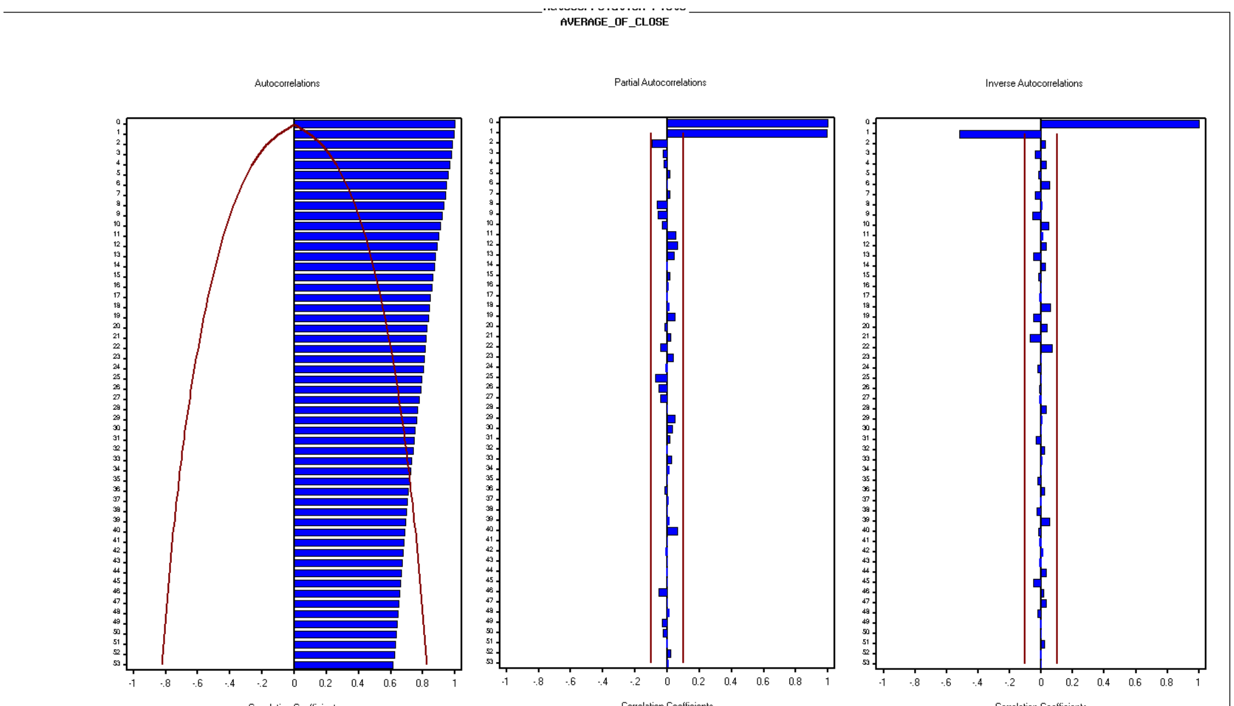


Modeling - Forecasting Without Sentiments

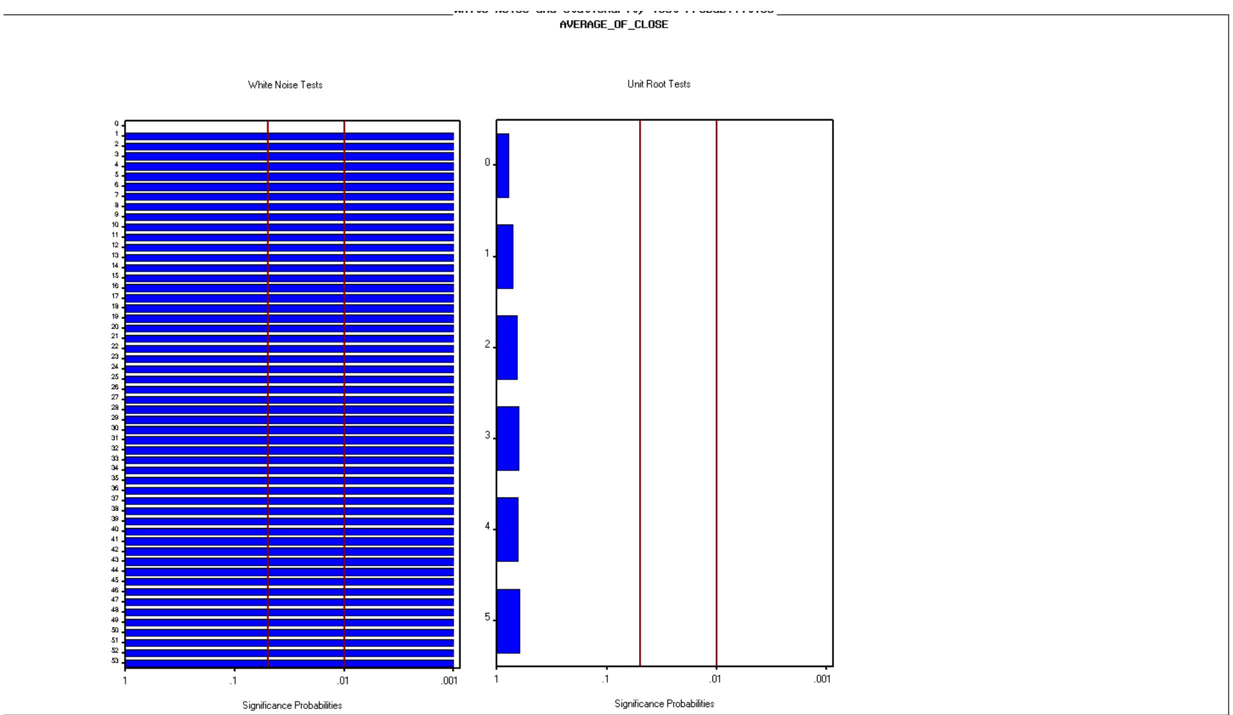
- Objective The objective of the project is to forecast the DJIA stock prices.



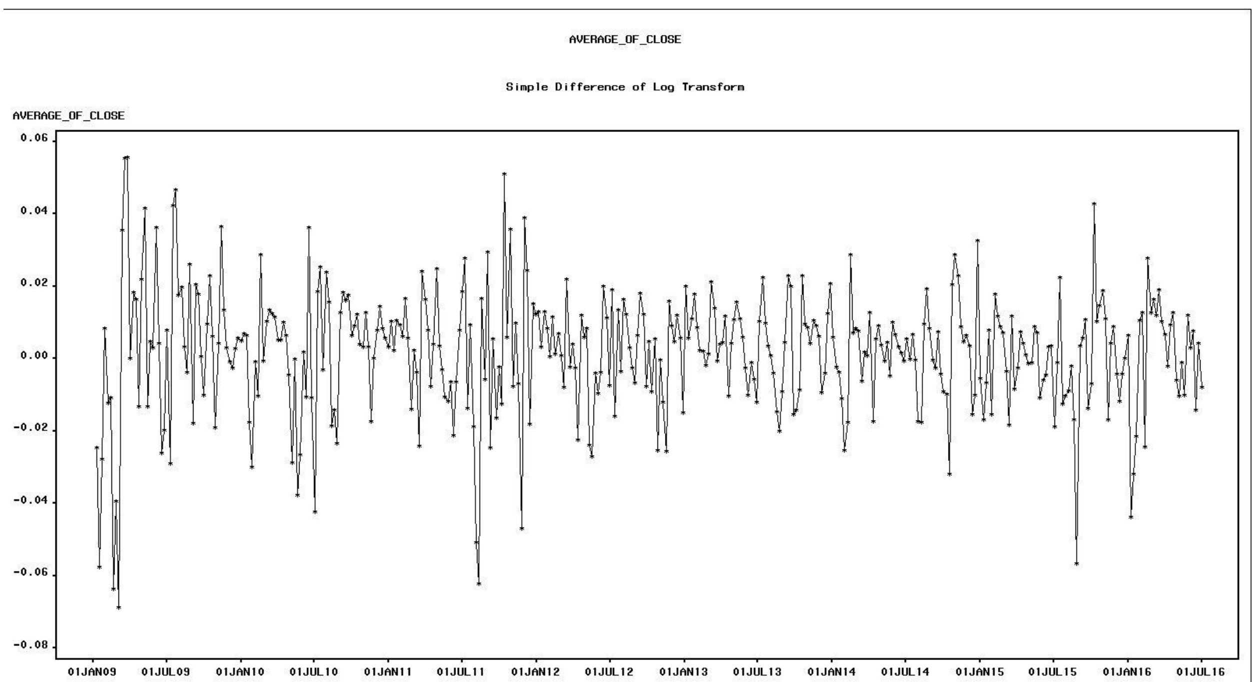
The trend indicates that the stock value has increased over time with certain dips. This is due to general improvement the market and the diminishing value of money over time.



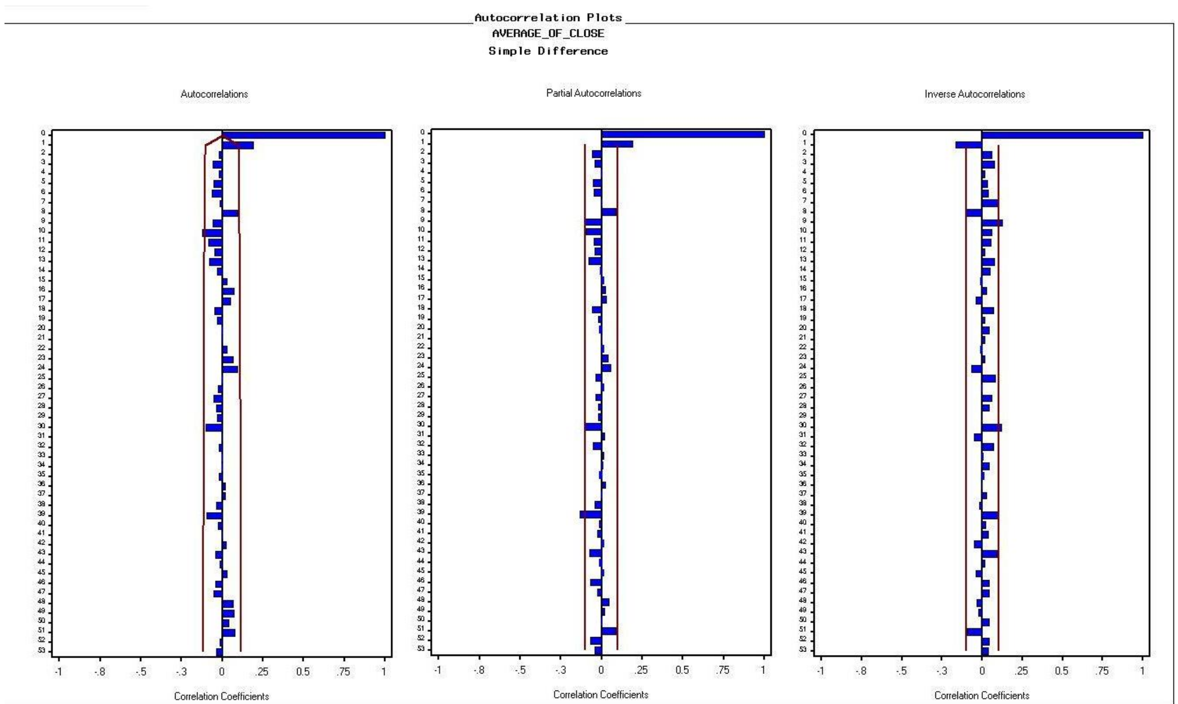
The autocorrelation plot indicates that most of the values are beyond the 95% confidence.



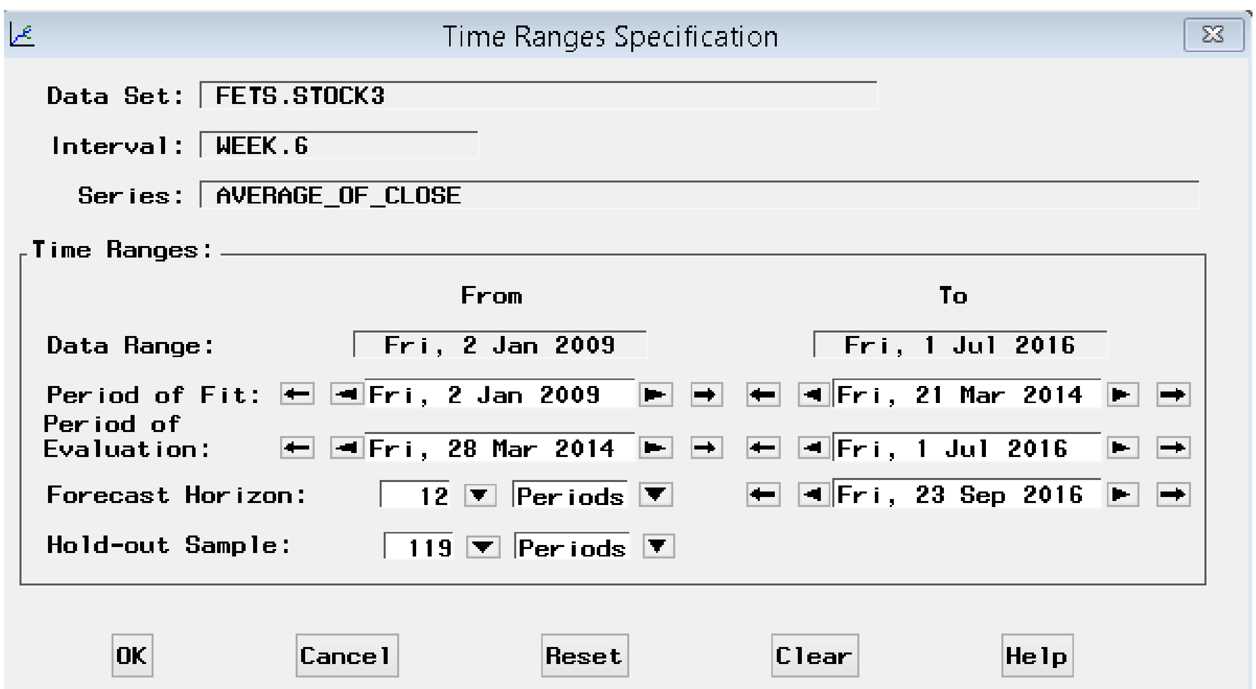
There is significant white noise.



First differences have been applied of log transfers.



The correlation and auto correlation plots look much better with spikes at p=1 and q=1


Time Ranges Specification

Data Set: FETS.STOCK3

Interval: WEEK.6

Series: AVERAGE_OF_CLOSE

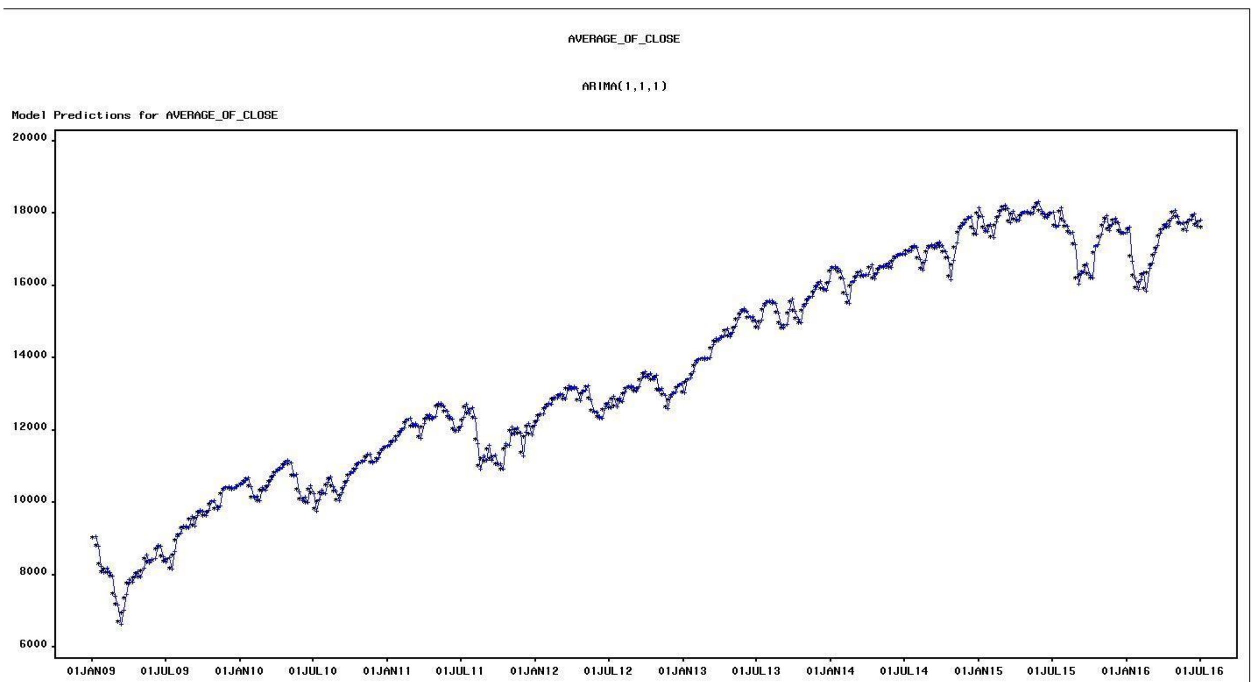
Time Ranges:

	From	To
Data Range:	Fri, 2 Jan 2009	Fri, 1 Jul 2016
Period of Fit:	Fri, 2 Jan 2009	Fri, 21 Mar 2014
Period of Evaluation:	Fri, 28 Mar 2014	Fri, 1 Jul 2016
Forecast Horizon:	12 Periods	Fri, 23 Sep 2016
Hold-out Sample:	119 Periods	

OK Cancel Reset Clear Help

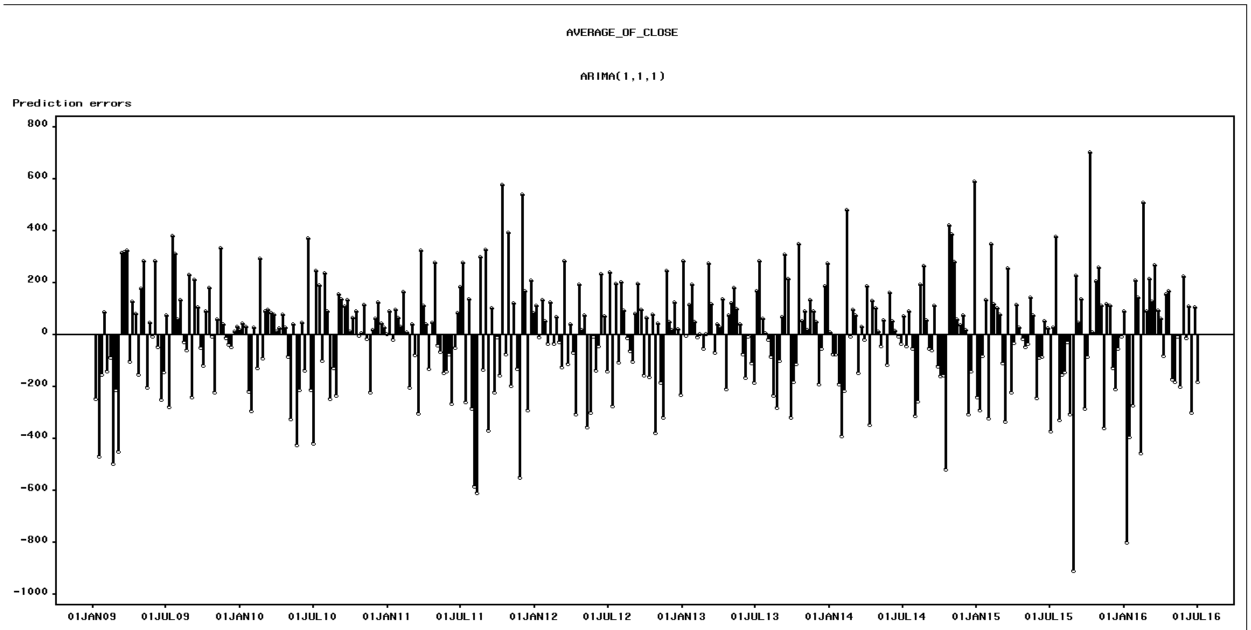
We divided the data into training and holdout data set with 70% training and 30 holdout .

There were 3 models which we developed out of which ARIMA (1, 1, 1) and Damped Trend Exponential smoothing showed least RMSE.

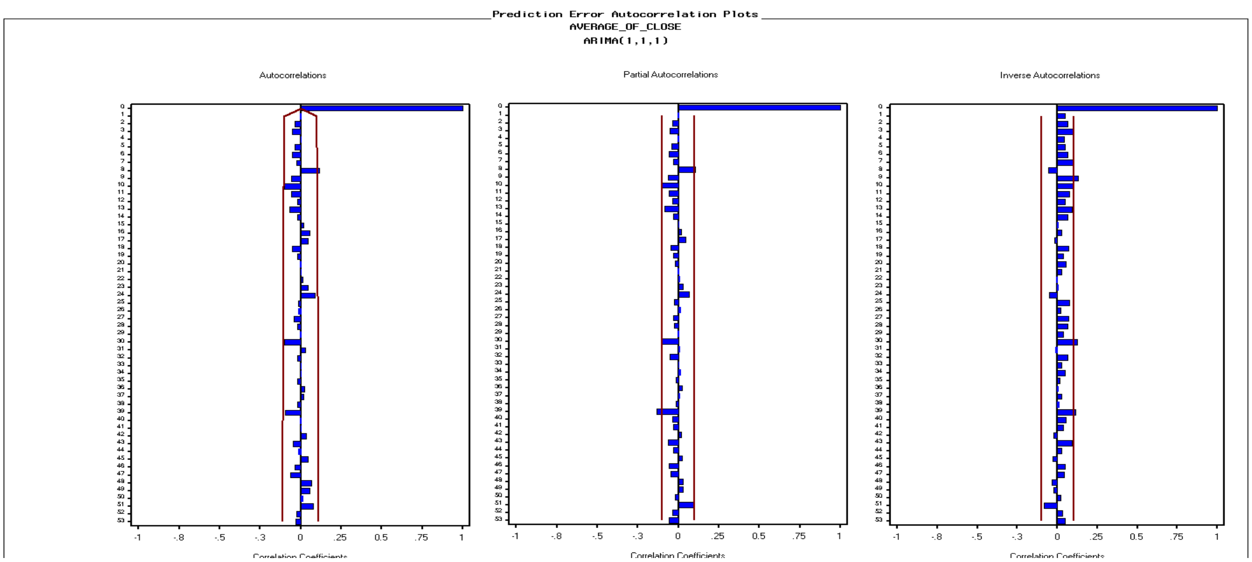


This is the trend of the ARIMA(1, 1, 1) model.

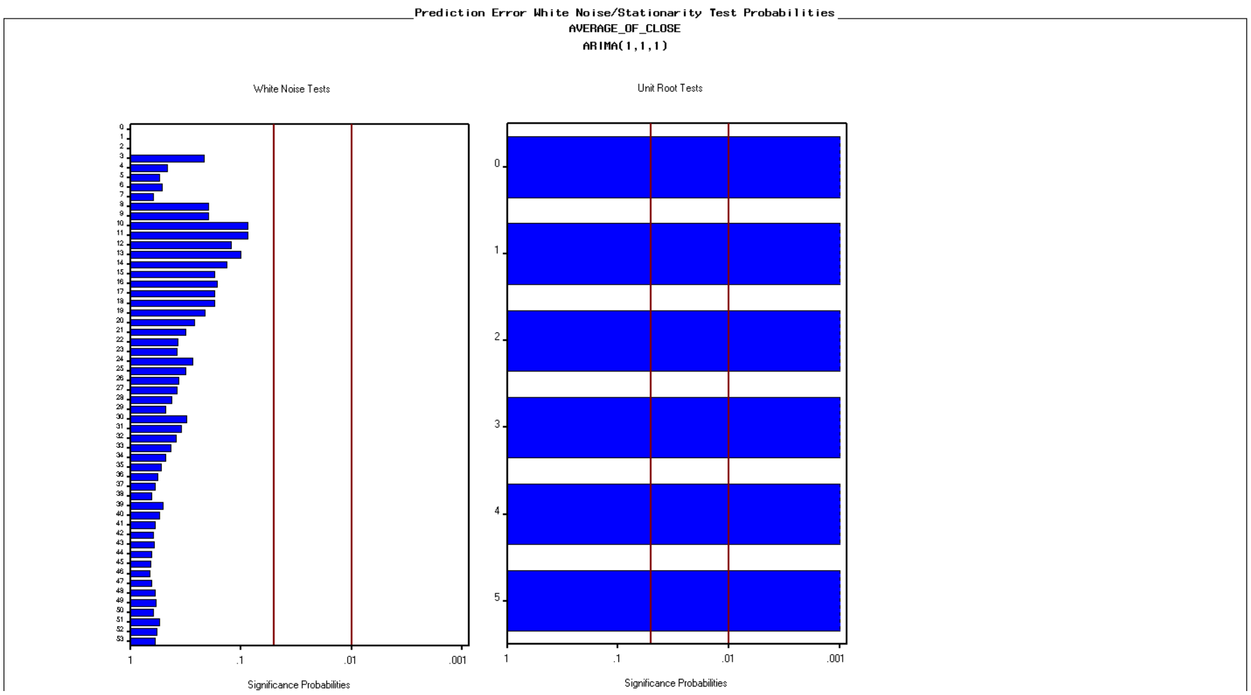
Model 1



Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event.



Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with all values within the 95% confidence interval



Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.

Parameter Estimates
AVERAGE_OF_CLOSE
ARIMA(1,1,1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	26.50307	14.5889	1.8167	0.0719
Moving Average, Lag 1	-0.08087	0.2999	-0.2697	0.7879
Autoregressive, Lag 1	0.12412	0.2985	0.4159	0.6783
Model Variance (sigma squared)	38081	.	.	.

Fit Range: Fri, 2 Jan 2009 to Fri, 21 Mar 2014

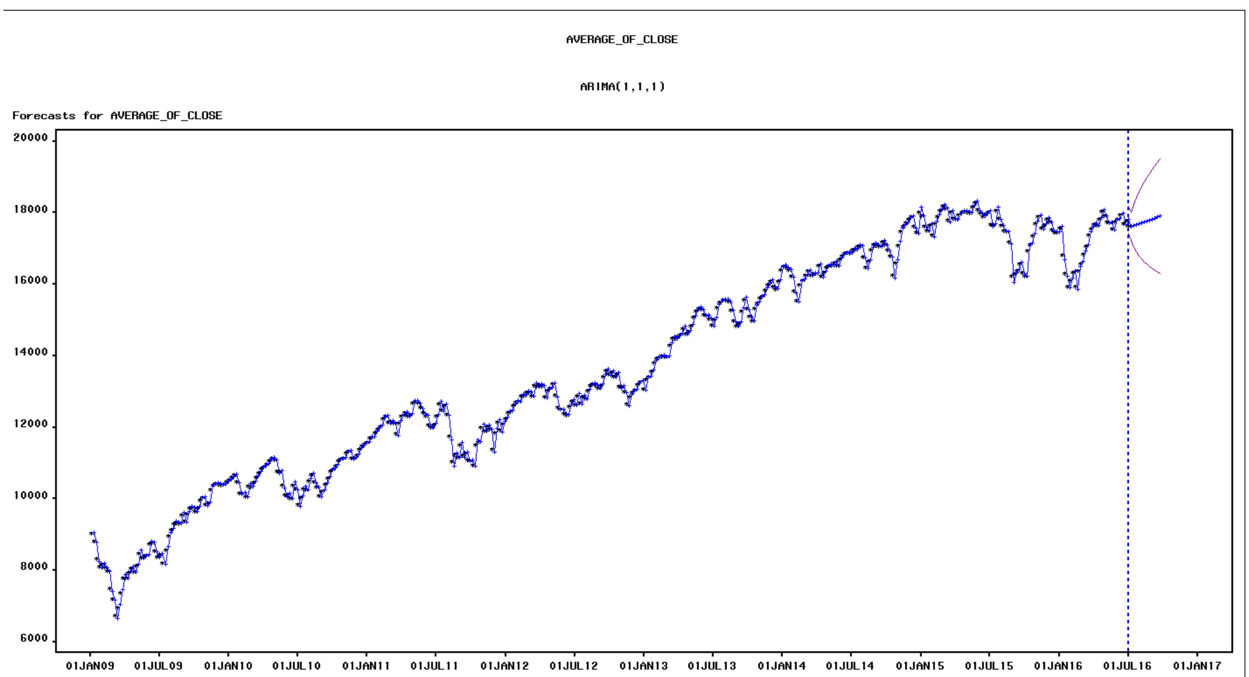
Parameter Estimates: The variable intercepts have significance in the model forecast

Statistics of Fit
AVERAGE_OF_CLOSE
ARIMA(1,1,1)

Statistic of Fit	Value
Mean Square Error	57148.7
Root Mean Square Error	239.05798
Mean Absolute Percent Error	1.03295
Mean Absolute Error	176.96037
R-Square	0.850

Evaluation Range: Fri, 28 Mar 2014 to Fri, 1 Jul 2016

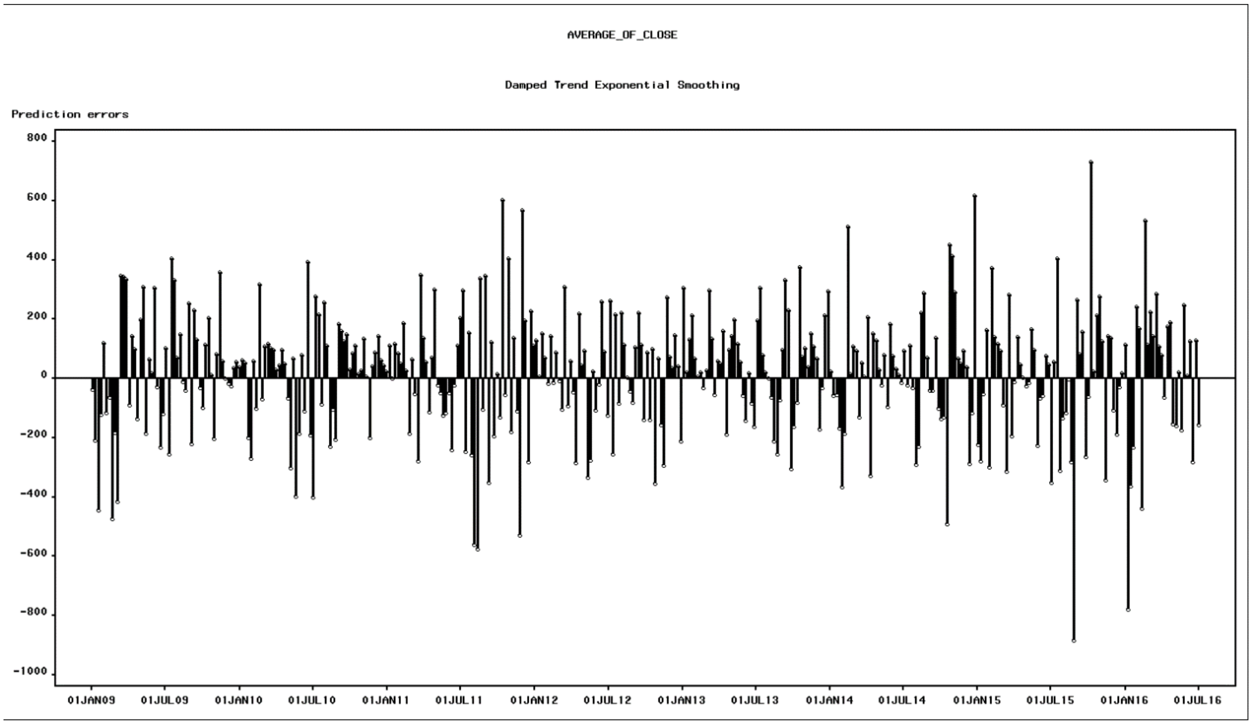
Statistics of fit: MAPE is 1.03%, which is good.



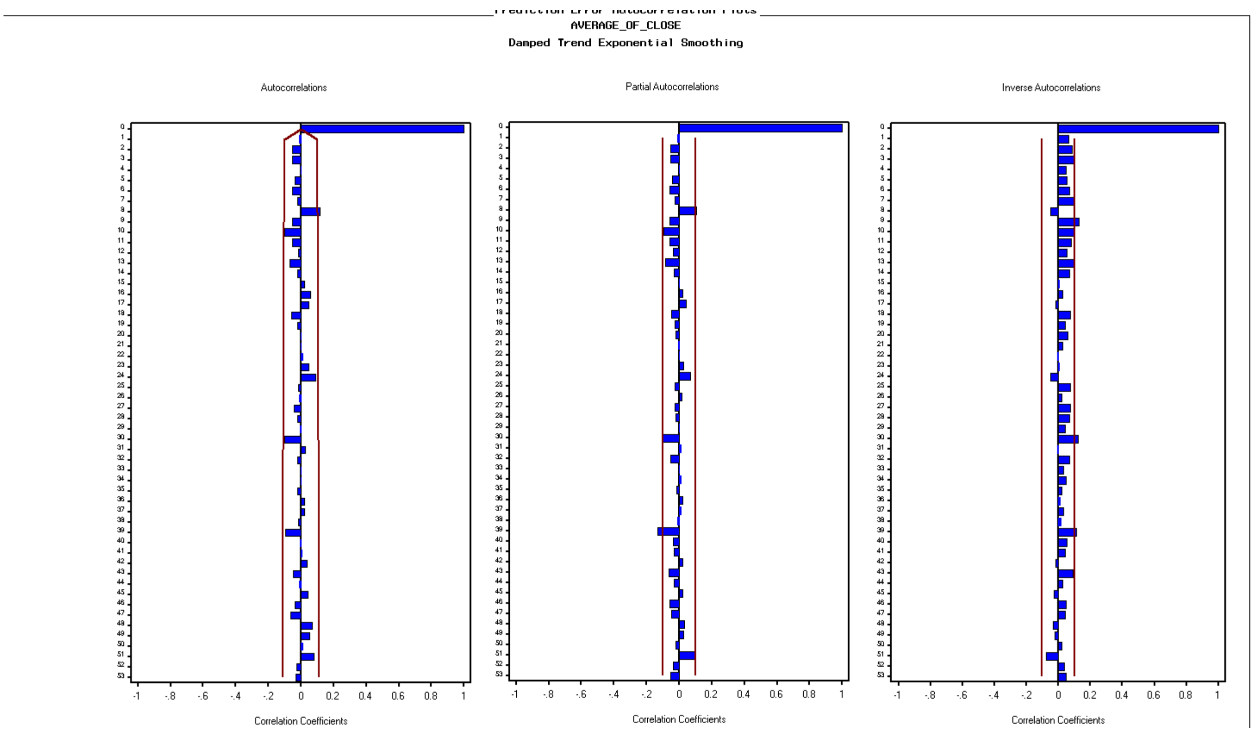
Forecast plot: The forecasted plot captures the trend well.

Model 2 : Damped Trend Exponential Smoothing

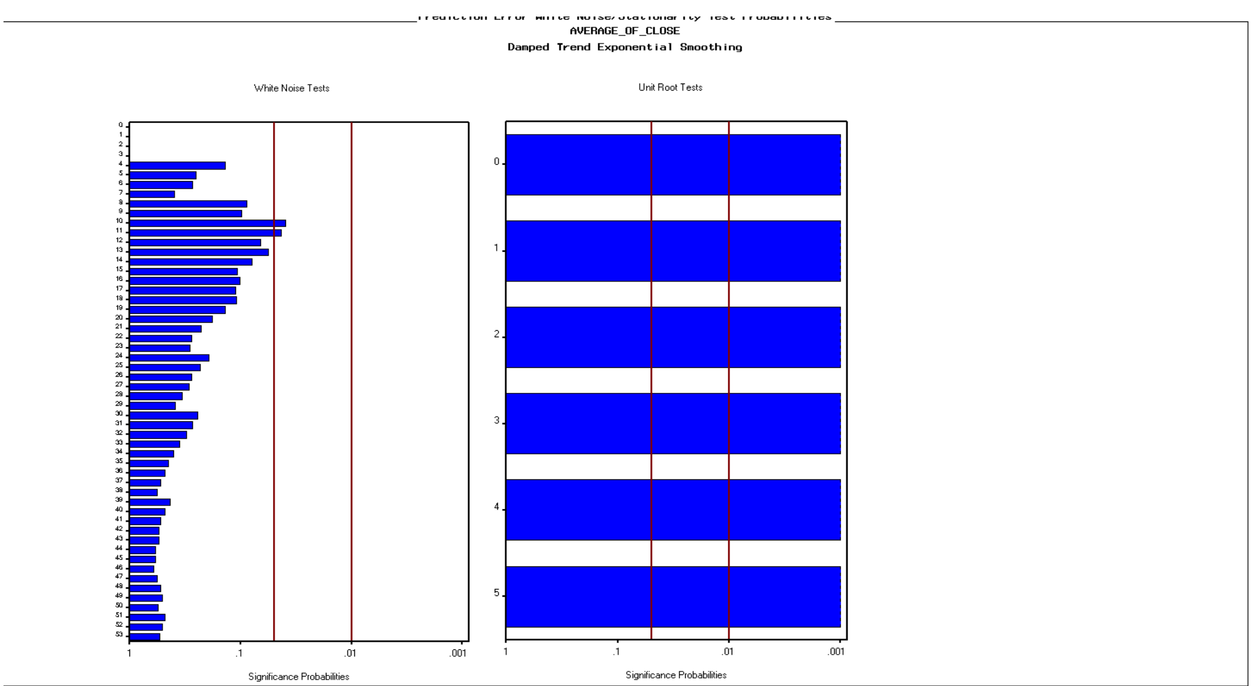
We then observed the parameters of the next best model in terms of RMSE to compare both models and finally chose one model over the other.



Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous model.



Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



Prediction error white noise/ Unit root test: There is some significant white noise, the unit test indicates that the series is stationary.

AVERAGE_OF_CLOSE
Damped Trend Exponential Smoothing

Model Parameter	Estimate	Std. Error	T	Prob>|T|
LEVEL Smoothing Weight	0.99900	0.8965	1.1144	0.2674
TREND Smoothing Weight	0.99900	8.9609	0.1115	0.9114
DAMPING Smoothing Weight	0.21835	0.8766	0.2491	0.8037
Residual Variance (sigma squared)	38397	.	.	.
Smoothed Level	16288	.	.	.
Smoothed Trend	30.72989	.	.	.

Fit Range: Fri, 2 Jan 2009 to Fri, 21 Mar 2014

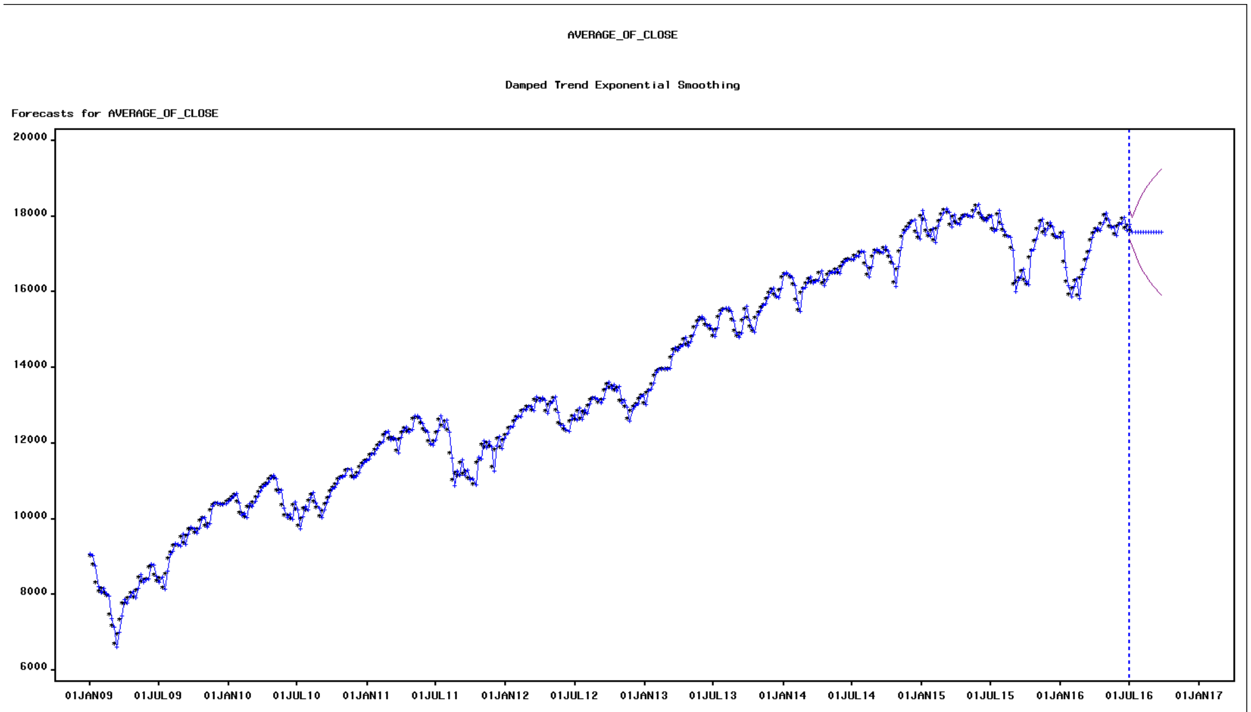
Parameter Estimates: The variables have less significance in the model forecast, however we don't reject the model based on this.

AVERAGE_OF_CLOSE
Damped Trend Exponential Smoothing

Statistic of Fit	Value
Mean Square Error	57310.7
Root Mean Square Error	239.39646
Mean Absolute Percent Error	1.04213
Mean Absolute Error	178.58214
R-Square	0.850

Evaluation Range: Fri, 28 Mar 2014 to Fri, 1 Jul 2016

The statistics of fit looks good.



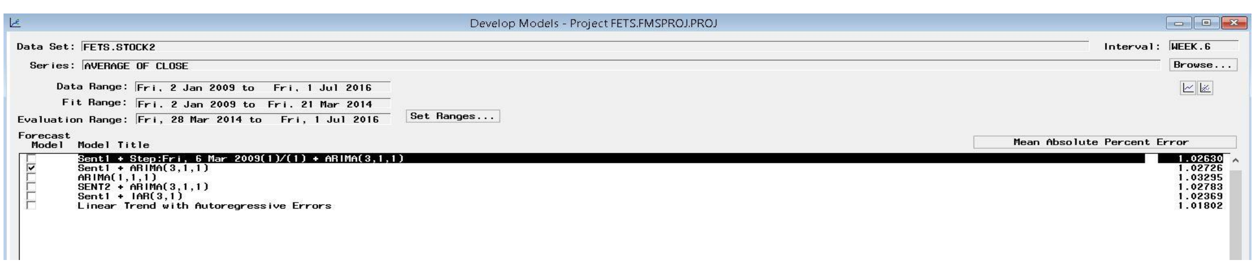
Forecast plot: The forecast shows no trend and is simply the mean. Thus we reject the model

Conclusion

Out of the two models we select the ARIMA model as it shows better forecasting plots and has better RMSE.

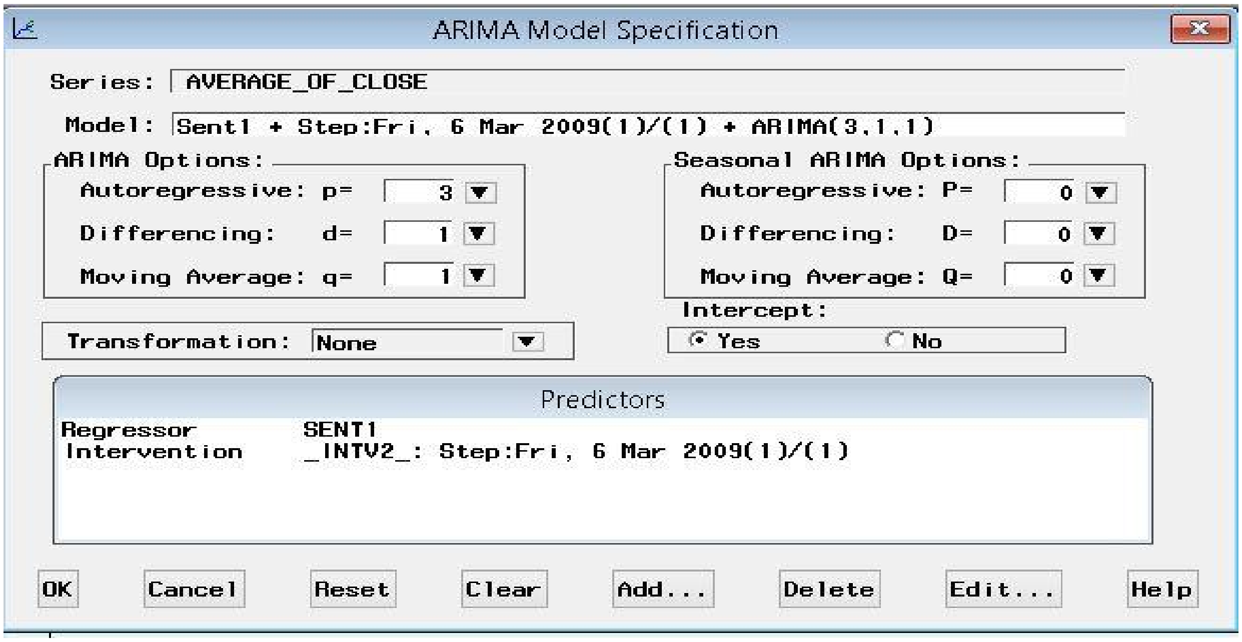
Modelling - Forecasting with Sentiments

- Objective: To predict the average closing of DJIA stock data, now incorporating sentiments from reddit.

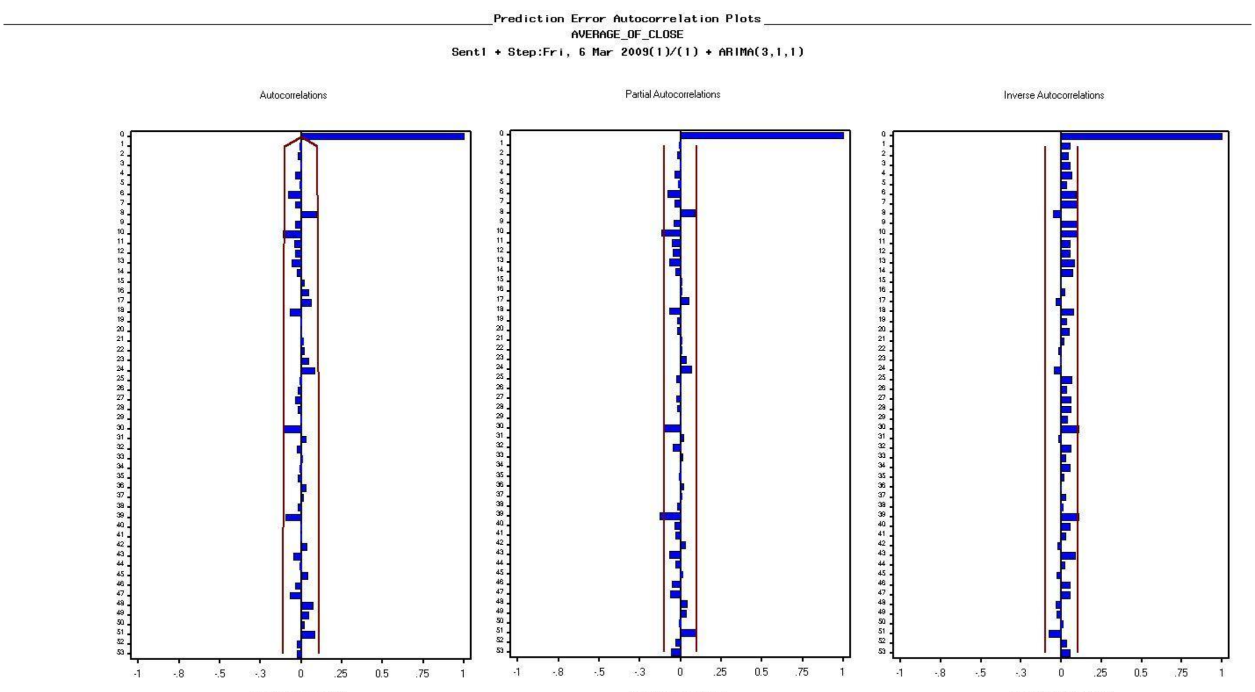


Primary 2 models stand out model with sentiment as regressor and Intervention at 6th March 2009. There was a drastic dip in the stock prices due to the recession and a model without intervention.

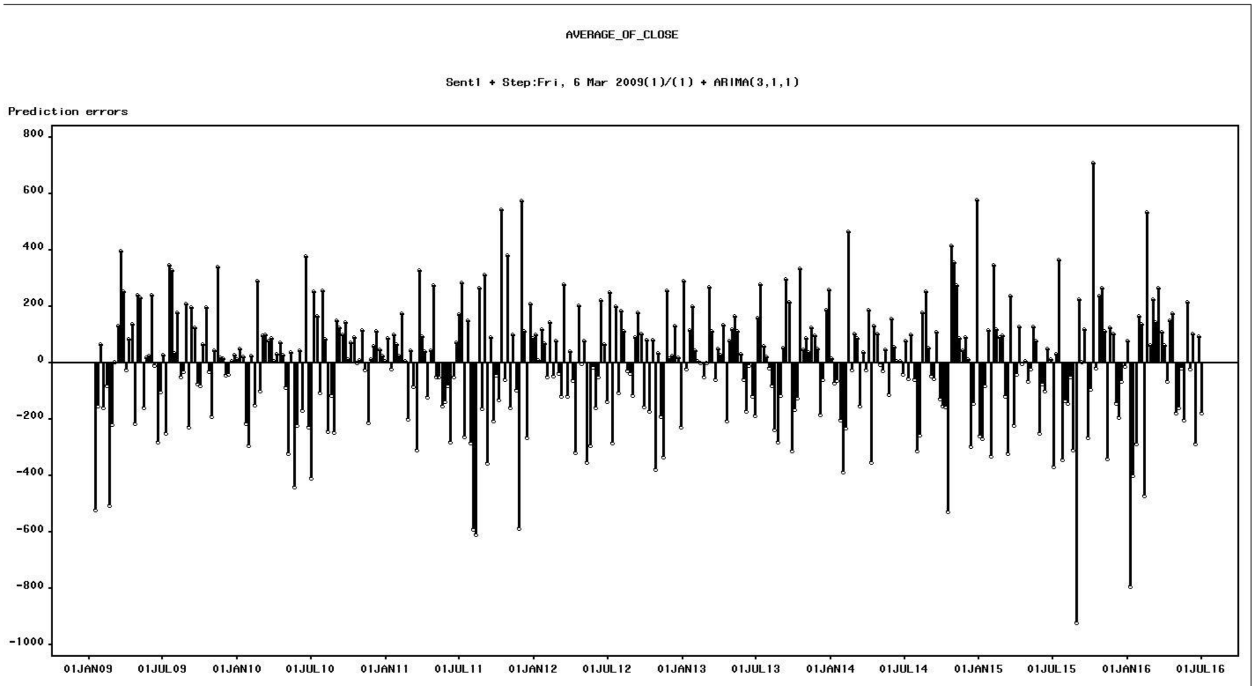
Model 1



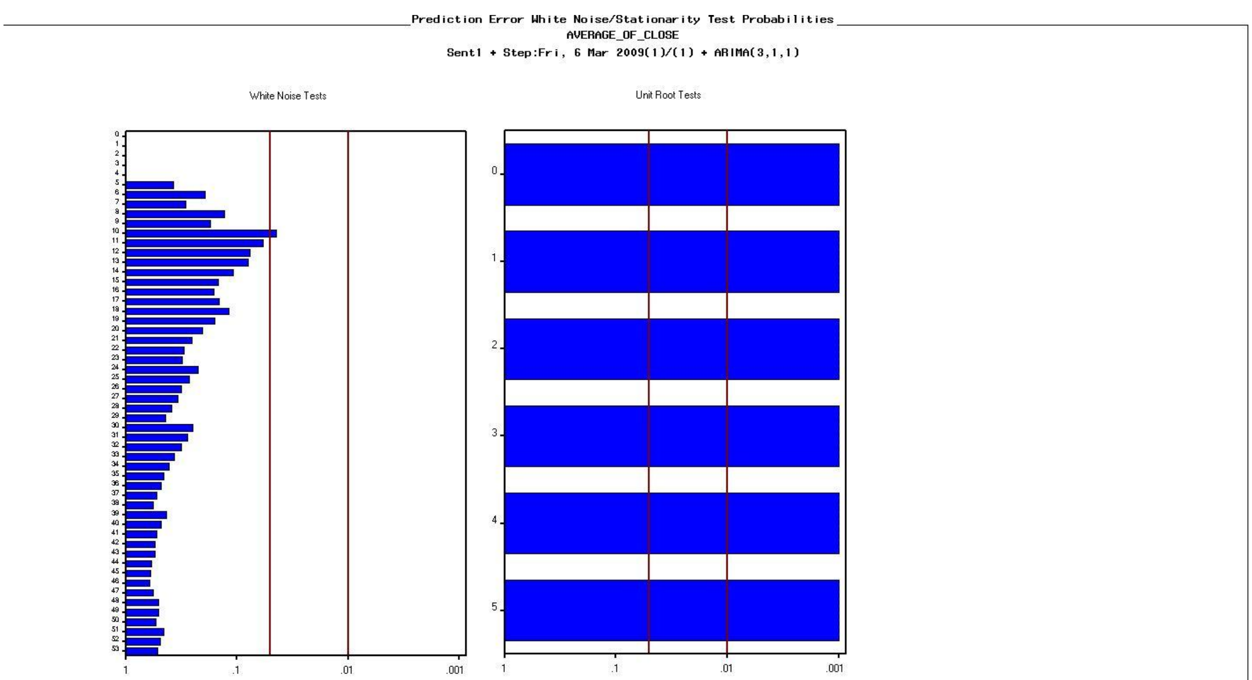
We have incorporated step intervention at 6th March 2009 and Sentiment 1 as regressor



Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous models



Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.

Parameter Estimates
AVERAGE_OF_CLOSE
Sent1 + Step:Fri, 6 Mar 2009(1)/(1) + ARIMA(3,1,1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	28.92932	13.8811	2.0841	0.0395
Moving Average, Lag 1	-0.89185	0.2275	-3.9211	0.0002
Autoregressive, Lag 1	-0.68245	0.2339	-2.9175	0.0043
Autoregressive, Lag 2	0.14221	0.0861	1.6521	0.1014
Autoregressive, Lag 3	-0.06801	0.0638	-1.0665	0.2886
Sent1	0.73159	3.0740	0.2380	0.8123
Step:Fri, 6 Mar 2009(1)/(1)	-466.06485	193.1885	-2.4125	0.0175
Step:Fri, 6 Mar 2009(1)/(1) Num1	382.18318	217.8540	1.7543	0.0822
Step:Fri, 6 Mar 2009(1)/(1) Den1	-0.94921	0.0934	-10.1625	<.0001
Model Variance (sigma squared)	37663	.	.	.

Fit Range: Fri, 2 Jan 2009 to Fri, 21 Mar 2014

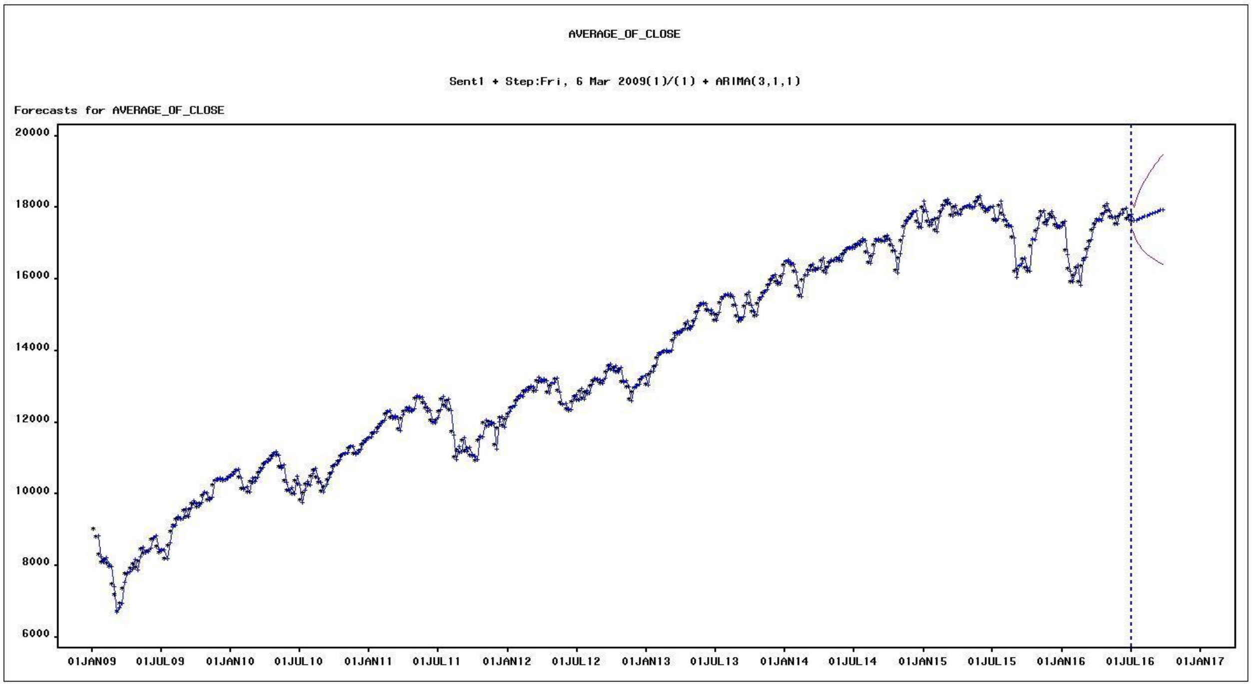
Parameter estimates: The sentiment has an impact of .73 on the model.

Statistics of Fit
AVERAGE_OF_CLOSE
Sent1 + Step:Fri, 6 Mar 2009(1)/(1) + ARIMA(3,1,1)

Statistic of Fit	Value
Mean Square Error	56939.6
Root Mean Square Error	238.62012
Mean Absolute Percent Error	1.02630
Mean Absolute Error	175.82671
R-Square	0.851

Evaluation Range: Fri, 28 Mar 2014 to Fri, 1 Jul 2016

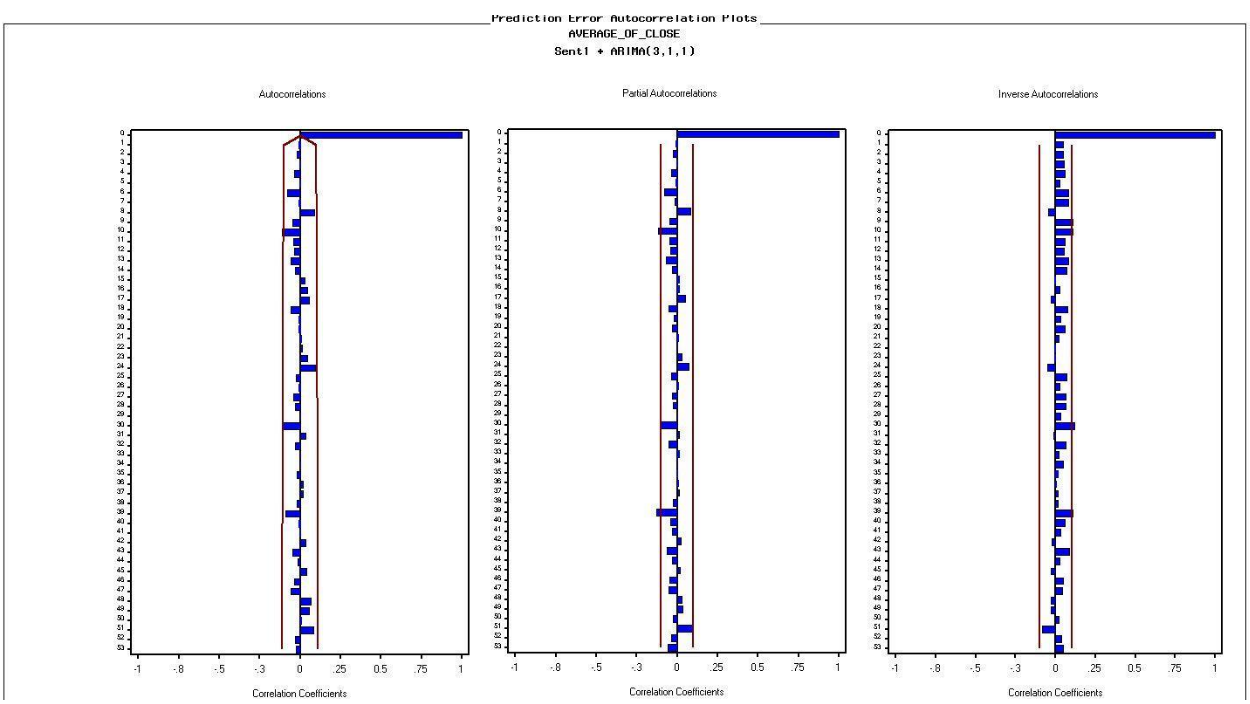
The Statistics of fit looks good



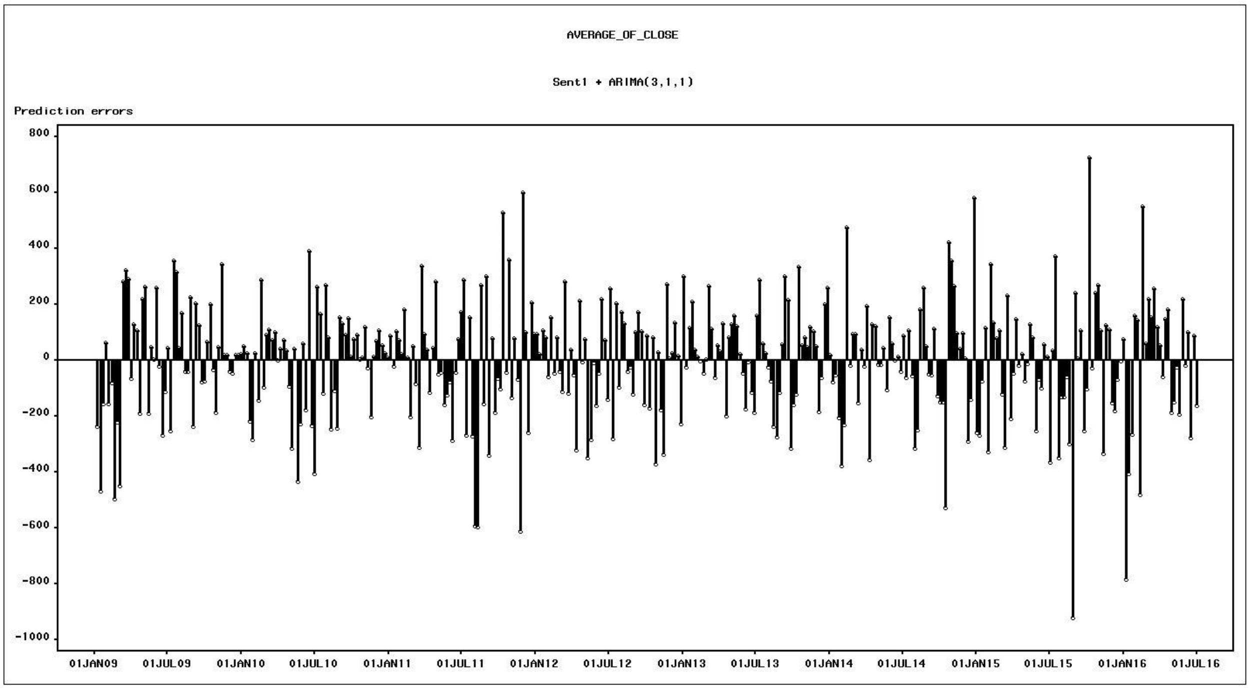
Forecast plot: The forecasted plot captures the trend well.

Model 2

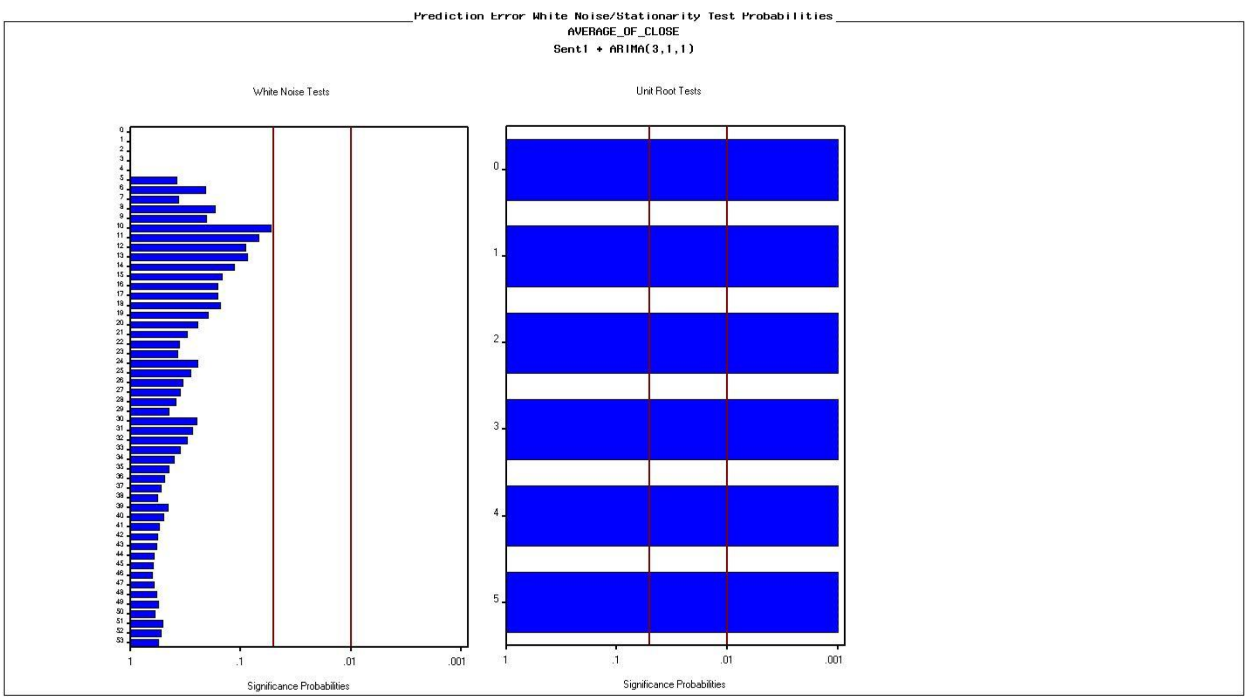
Analysing the model with sentiments as regressors however without interventions



Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous models



Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.

AVERAGE_OF_CLOSE
Sent1 + ARIMA(3,1,1)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	26.79283	14.0242	1.9105	0.0586
Moving Average, Lag 1	-0.88597	0.1375	-6.4444	<.0001
Autoregressive, Lag 1	-0.67283	0.1483	-4.5354	<.0001
Autoregressive, Lag 2	0.15705	0.0767	2.0482	0.0429
Autoregressive, Lag 3	-0.07525	0.0646	-1.1653	0.2463
Sent1	1.53930	3.0424	0.5060	0.6139
Model Variance (sigma squared)	38116	.	.	.

Fit Range: Fri, 2 Jan 2009 to Fri, 21 Mar 2014

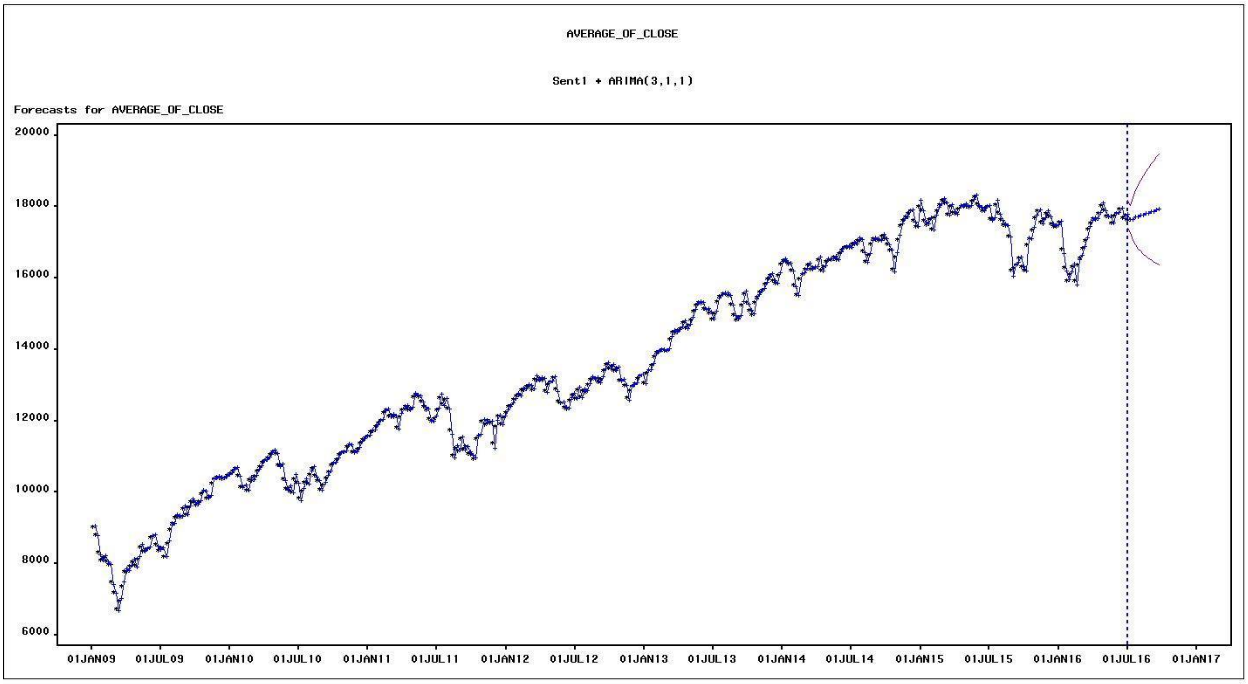
Parameter estimates: The sentiment has an impact of 1.53 on the model.

AVERAGE_OF_CLOSE
Sent1 + ARIMA(3,1,1)

Statistic of Fit	Value
Mean Square Error	56972.0
Root Mean Square Error	238.68808
Mean Absolute Percent Error	1.02726
Mean Absolute Error	175.98509
R-Square	0.851

Evaluation Range: Fri, 28 Mar 2014 to Fri, 1 Jul 2016

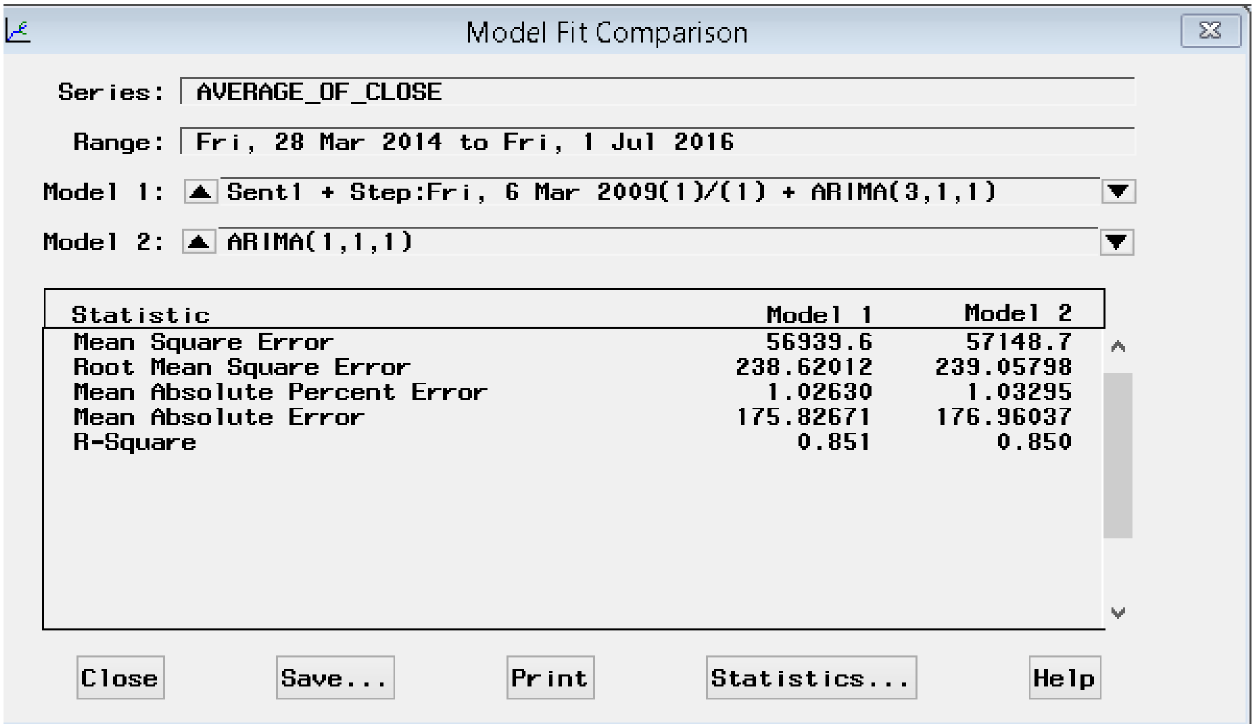
The statistics of fit looks good



Forecast plot: The forecast plot captures the trend well

Conclusion: Based on the RMSE values we conclude that the model with intervention performs better.

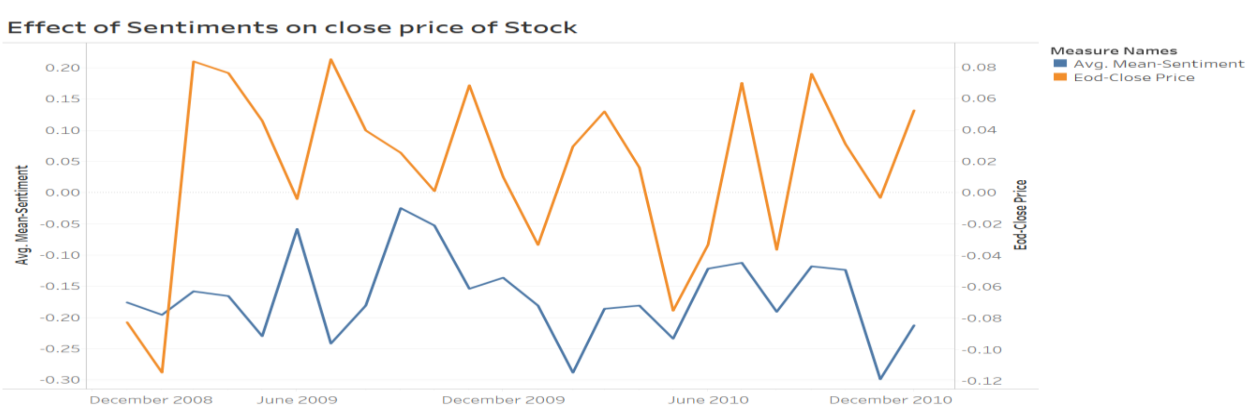
Comparison between Models in Part 1 and Part 2



Conclusion: We observe that the model with sentiments is better in terms of both RMSE and MAPE

Volume prediction

As seen from the graph below, the positive and negative sentiments have high correlation with the closing stock price. As the mean sentiment starts to decrease, the close price also tends to decrease and vice versa.



Our objective is to try and predict the volume of stock that will be traded based on the Google Trends obtained during the relevant time period.

Load the data set containing the log-volume of stocks traded and Google Trends during that time. Divide the data into train and test data set. Build a simple linear regression model.

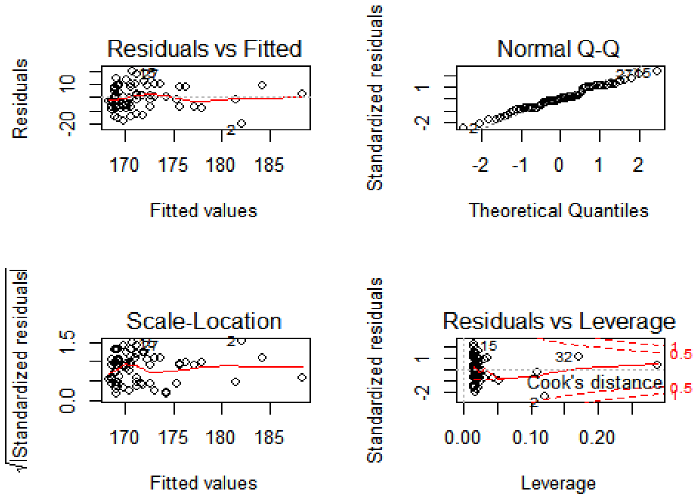
```
DJIA_Monthly_VolTrend <- read.csv("~DJIA_Monthly_VolTrend.csv")
djia.train = DJIA_Monthly_VolTrend[1:70,]
djia.test = DJIA_Monthly_VolTrend[71:90,]

reg1 = lm(Log.volume~DJIA.Trend, data=djia.train)
summary(reg1)
```

```
##
## Call:
## lm(formula = Log.volume ~ DJIA.Trend, data = djia.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4927  -6.7835  -0.1204   8.0891  19.4154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 163.69456    2.36267  69.284  < 2e-16 ***
## DJIA.Trend    0.24726    0.06643   3.722 0.000403 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.652 on 68 degrees of freedom
## Multiple R-squared:  0.1692, Adjusted R-squared:  0.157
## F-statistic: 13.85 on 1 and 68 DF,  p-value: 0.000403
```

Observe the error plots of the model to see if they are normally distributed or not

```
opar = par()
par(mfrow=c(2,2))
```

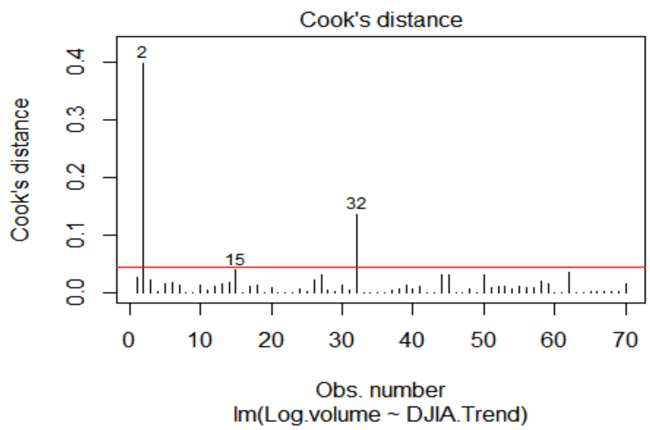
```
plot(reg1)
```



Calculate the cooks distance to find the observations skewing the model.

```
z=cooks.distance(reg1)
round(z,4)
cutoff = 4/nrow(DJIA_Monthly_VolTrend)
length(z[z>cutoff])
## [1] 2
plot(reg1,which = 4, cook.levels = cutoff)
abline(h=cutoff,col="red")
```

The plot shows the observation which are above the cooks distance cutoff. Build another model after removing the observations flagged above.




```
reg2 = lm(Log.volume~DJIA.Trend, data=djia.train[-c(2,15,32),])
summary(reg2)

##
## Call:
## lm(formula = Log.volume ~ DJIA.Trend, data = djia.train[-c(2,
##     15, 32), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1914  -6.3723  -0.2104   7.4772  17.8346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.52572    2.43230  66.820  < 2e-16 ***
## DJIA.Trend    0.28105    0.07259   3.871 0.000254 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.073 on 65 degrees of freedom
## Multiple R-squared:  0.1874, Adjusted R-squared:  0.1749
## F-statistic: 14.99 on 1 and 65 DF,  p-value: 0.0002536
```

It can be observed that the R2 clearly improved in the second model.

Let's have a look at AIC of both models.

```
AIC(reg1,reg2)

##        df     AIC
## reg1  3      504.7169
## reg2  3      473.9715
```

Model 2 definitely performs better than model 1. Let's make prediction on test data set based on our model 2. Add the predicted values as a column in our existing dataset

```
pred = predict(reg2,newdata = djia.test)
djia.test$prediction = pred
```

Thus, we can conclude that, up to a certain extent we can predict the volume of stock to be traded during a given month based on the google trends.

Business Applications

The main purpose of our project is to use text analytics results in improvement of prediction accuracy for stock prices. Both Trading Institutions and Individual Investors can be benefit from it.

1. **Trading Institutions**

   Many traders take decisions in the financial market solely based on what other people think and what they recommend. By understanding the impact of the public sentiment, Trading Institutions can prepare the trend changes of the stock price in advance. Moreover, by set the opening price to an investor satisfied level, more and more investors will be attracted to invest. In this way, they can alter their investment strategies and portfolio to make the maximum profit.

2. **Individual Investors**

   The use of text analytics provides new revenue streams for traditional publishing outlets and new sources of insight and efficiency for individual investors. The relationship between text and stock price allows information to move quicker and deeper. Common sentiments, such as happy and anger, of certain news can allow the individual investors to understand what the price trend will be in the future and allow them to change their own investment portfolios in order to gain the maximum profit.

Conclusion

We have investigated the causative relation between public mood as measured from a large scale collection of news from reddit.com and the DJIA values. Our results show that firstly public mood can indeed be captured from the large-scale news feeds by means of simple natural language processing techniques. We haven't been able to obtain high percentage result as expected, however we have obtained MAPE of about 1% using SAS Time Series Forecasting System. It is worth mentioning that our analysis doesn't take into account many factors. Firstly, our dataset doesn't really map the real public sentiment, it only considers the reddit using, english speaking people. It's possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people's mood indeed affect their investment decisions, hence the correlation. But in that case, there's no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research.

References

[1] https://pdfs.semanticscholar.org/4ecc/55e1c3ff1cee41f21e5b0a3b22c58d04c9d6.pdf

[2] http://www.kdnuggets.com/2016/01/sentiment-analysis-predictive-analytics-trading-mistake.html

[3] https://www.kaggle.com/aaron7sun/stocknews

[4] https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

[5] https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

[6] https://trends.google.com/trends/explore?q=djia

[7] https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment

[8] http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

[9] https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf

Appendix

Stock Data

- Before Data Processing

DJIA_table .csv

- After Data Processing

DJIA_table - Volume transformed.csv

News Data

- Before Data Processing

RedditNews.csv

- After Data Processing

ReditSentimentAnalysis.csv .xlsx