# 2015

## Correlation between Market Sentiment, Research & Trading /Predicting Trading Volatility



Madhavi Pilli, Tanmoy Choudhury & Vipin Singh

# Abstract

In this study, we empirically examine the relationship between market sentiments, abnormal news, volatility and returns. This study is motivated by realization that market sentiment plays a significant role on trading volatility. We then extend our examination to in house research rating and trading volumes to explore which one plays significant role on portfolio manager's trading decision. Our main empirical findings are: (1) the effects of sentiment and abnormal news to trading volatility are outstanding. We observed significant correlation exists between abnormal news, sentiment, trading volatility and day return; (2) in house day trading seems to be more oriented towards market sentiment rather than the research rating.

# Introduction

Traditional financial theory advocates that shifts in investor sentiment is not expected to affect stock prices since such shocks are supposed to be offset in no time by the action of rational arbitrageurs[1]. Basic understanding is that, mispricing is short lived and would eventually revert back to its fundamentals. However, it is apparent that sentiment may affect stock prices.

In this paper, we attempt to establish the correlation of abnormal news and sentiment while the stock fundamentals are strong. We refer to in house research rating for individual stocks to identify its fundamentals, like rating 1 means a Strong Buy, means the stock's future direction is strong. With such a strong rating even though market sentiment may be negative the overall position for such stock should not impact a long-term return outlook. Our study shows a remarkable correlation between sentiment and volatility, probing the traditional financial hypothesis; furthermore we also observe that sentiment has considerable impact on in house trading decisions, even though the long-term research outlook is contradicting at times.

## Market Sentiment

Market sentiment is important to day traders and technical analysts, who use technical indicators to attempt to measure and profit from the short-term price changes often caused by investors' attitudes toward a security. Market sentiment is also important to contrarian investors, who like to trade in the opposite direction of the prevailing sentiment. For example, if everyone is buying, a contrarian would sell.

Evidence from studies in behavioral finance suggests that investor sentiment leads to stock mispricing (e.g., Brown and Clift 2005, Baker and Wurgler 2006, Lemmon and Portniaguina 2006). In general, this literature finds investors are overly optimistic (pessimistic) during periods of high (low) investor sentiment, leading to overvaluation (undervaluation) that reverses in the future.

Relationship between stock prices and consumer confidence seem to appear in the business press whenever new consumer survey data is released. Headlines like "Rise in Consumer Sentiment Sends Share Prices Higher"[2] and "Stocks Tumble, Spurred by Dive in Consumer Confidence" [3] suggest that stock prices respond directly to measures of consumer confidence.

## Research
While quantifying the benefits is difficult, most of research's value lies in the unquantifiable benefits provided to investors:

- Comparative operating and valuation data on a company;
- Earnings estimates and target valuations based on reasonable data included in the report; and
- A reliable source of independent, third-party information on a continuous basis so that investors can track performance and evaluate an investment.

Research coverage benefits are not immediate, and the decision to invest in stock research is generally a long-term process. It takes time for investors to familiarize themselves with a stock and get comfortable with a new company and its investment potential.

Research, however, provides the market with more information and increases market efficiency, but it is hard to determine exactly when a report will convince an investor or a fund manager to buy a stock. It could be near the publication date or months later.

# Business Context and Question

Financial analysts have always indicated that market volatility impacts trading to a large extent. Access to volatility prediction will help in house fund managers in taking appropriate trading decision in the best interest of the shareholders.

We are trying to analyze market sentiment (Ravenpack), proprietary research and trade data to provide meaningful correlations that could lead to prediction of trading volatility.

In house research capabilities, which provide research advisory (Buy/Sell indicator) for various instruments that could be linked to market sentiment and historical trade information to predict trade volatility.

## QUESTION

Do active fund managers consider market sentiment, research and volatility along with other parameters before making an investment decision?

# Data Acquisition and Analysis

Data acquisition started with identifying the list of securities to be included in the analysis. To short list the securities we had to do manual data janitor work to identify securities that had relevant data in the sentiment as well as research datasets.

For the shortlisted securities we pulled out all other attributes that were needed for analysis based on discussions with subject matter experts as well as researching online to see what attributes could be considered.

## Data

- Raven Pack
  - Several individual predictor variables
    - Sentiment Strength
    - Abnormal News
    - Security Returns
    - Security Volatility Score
- Asset Management Warehouse
  - Net Trading volume per security
  - Research rating per security

A list of twenty securities was chosen for this analysis and data was gathered from the above dataset. List of securities and sample dataset are provided in appendix. Time series for this analysis is Q1 of 2014.

## Pre-processing of Data

All referred datasets used in this paper were in a relational database. Manual pre-processing of data was performed to get the data in the desired format. SQL queries were used to extract the data into coma separated files to be able to upload in to R.

The proprietary research notes are valid for a month and can be published on demand as well. To synchronize the data with trading and sentiment dataset for mashup's, research dataset was enhanced to include every day rating. Ex. say for a given security a rating was published on the beginning of the month; we copied the same rating for all the dates till next rating was published.

## Cleanse and Transform

As parts of data cleanse all 'NA' records were removed from the dataset. Simple transformation was performed to convert string to dates for all date related columns.

For research dataset a generic function was used for replacing each NA with the most recent non-NA prior to it.

```
#Loading Sentiment data in to R
sentimentdata <- read.csv("sentiment_data.csv", stringsAsFactors=FALSE)
str(sentimentdata)
sentimentdata$trade_dt <- as.Date(sentimentdata$trade_dt,'%m/%d/%y')
sentimentdata$eff_date <- as.Date(sentimentdata$eff_date,'%m/%d/%y')

#Getting Trade data in to R
tdata <- read.csv("trade_data.csv", stringsAsFactors=FALSE)
str(tdata)
tdata$trade_execution_dt <- as.Date(tdata$trade_execution_dt,'%m/%d/%y')
#Take net quantity as '000 to avoid exponential values
tdata$net_qty <- tdata$net_qty/1000

## Reading and pre-processing of research Data in to R
researchdata <- read.csv("research_data2.csv", stringsAsFactors=FALSE)

## Get the research Data for columns needed
researchdata <- researchdata[,c('Symbol','Rating_nbr','publish_date')]
```

4

# Replace each NA with the most recent non-NA prior
```
for(i in 1:20) { df[,i] <- na.locf(df[,i])}
```

## Convert the Long Format to Wide Format using the reshape
```
research.ts <- reshape(researchdata, timevar = "publish_date", idvar = c("Symbol"),direction =
"wide")
```

## Approach

We first try to explore Raven Pack dataset; establish correlation between sentiment, abnormal news and returns with volatility score. Furthermore, a mashup is performed between Raven Pack dataset and trading to observe any correlation. We then try to visualize in house trading and research dataset individually and perform a mashup between two dataset to observe a correlation between research rating and trading volume. We record our assessment based on the exploration and observation.

### Linked Dataset

Mash up of Sentiment data and Trade data is done to develop a linear model to show the correlation between volatility and the other attributes in the data set sentiment strength, abnormal news, return percent, trade net quantity. A correlation plot is also developed for this mashed up dataset.

#Link Sentiment and Trade data using sqldf
```
m <- sentimentdata
z <- tdata
str(z)
str(m)
#Library required for sqldf
library(sp)
library(gsubfn)
library(tcltk)
library(sqldf)
```

#Create merged data set (Trade and Sentiment)
```
tempdata <- sqldf("select a.volatility,b.net_qty,a.sentiment_strength,a.abnormal_news,a.return_pct
from m a, z b where a.symbol = b.symbol and a.trade_dt = b.trade_execution_dt")
```

Furthermore, a linked dataset between Trading and Research are created to establish any influence of research rating on in house trading. A correlation plot is also developed for this mashed up dataset.

```
#Link Trade and Research using sqldf
z <- tdata
r <- rdata
#Create merged data set (Trade and research)
tempdata1 <- sqldf("select Rating_nbr,b.net_qty t_vol
from r a, z b
where a.Symbol = b.symbol
and publish_date = b.trade_execution_dt")
```

## Data Visualization

Time series, QQ plots, correlation plot as well as scatter plots are developed for these datasets to be able to better visualize the data. This helped in further analysis of the data, finding correlation as well as developing of a linear model for certain attributes in this data sets.

- Sentiment Data
- Abnormal News Volume
- Return Pct. (Day)
- Trade volume
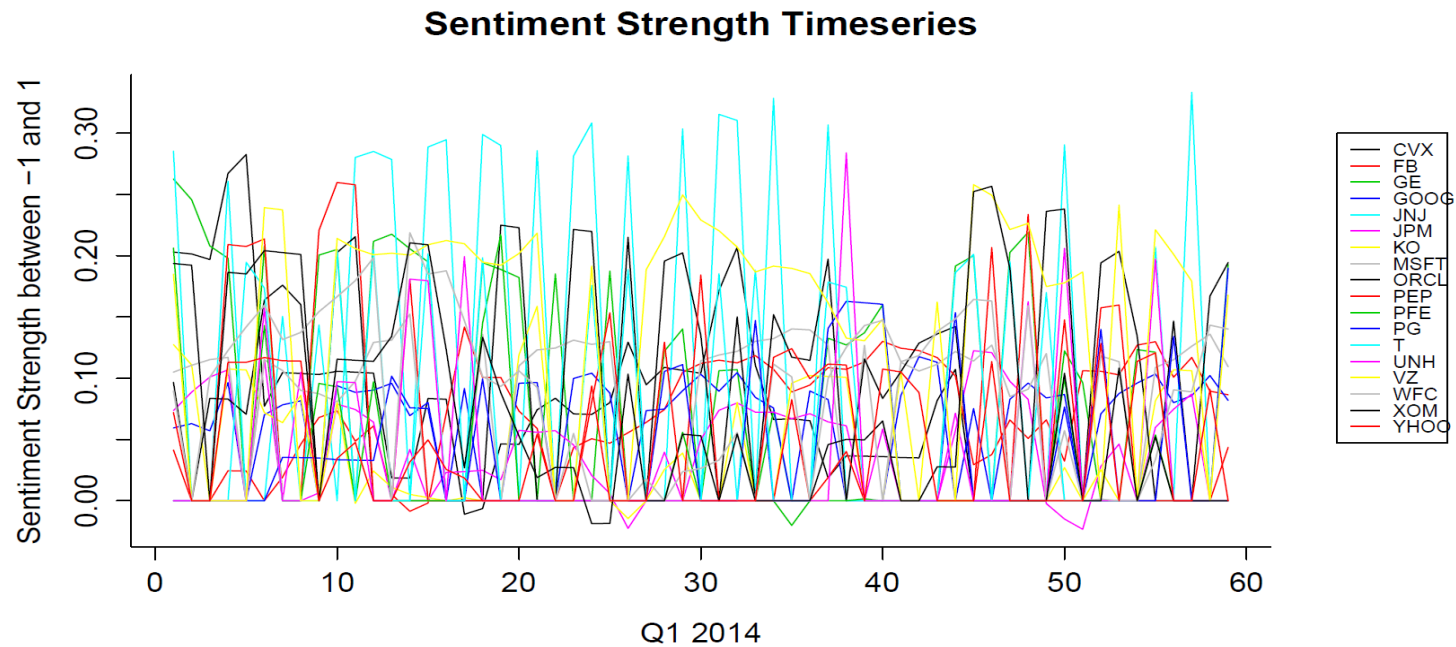- Research Rating
- Volatility

Fig.1 Time series for Sentiment Strength

As we can see in the time series shown Fig.1, sentiment strength for these securities is fluctuating between -1 and 1 where -1 indicates highly negative sentiment and +1 indicates a highly positive sentiment and 0 indicates neutral sentiment.

Justification of timeseries goes here..
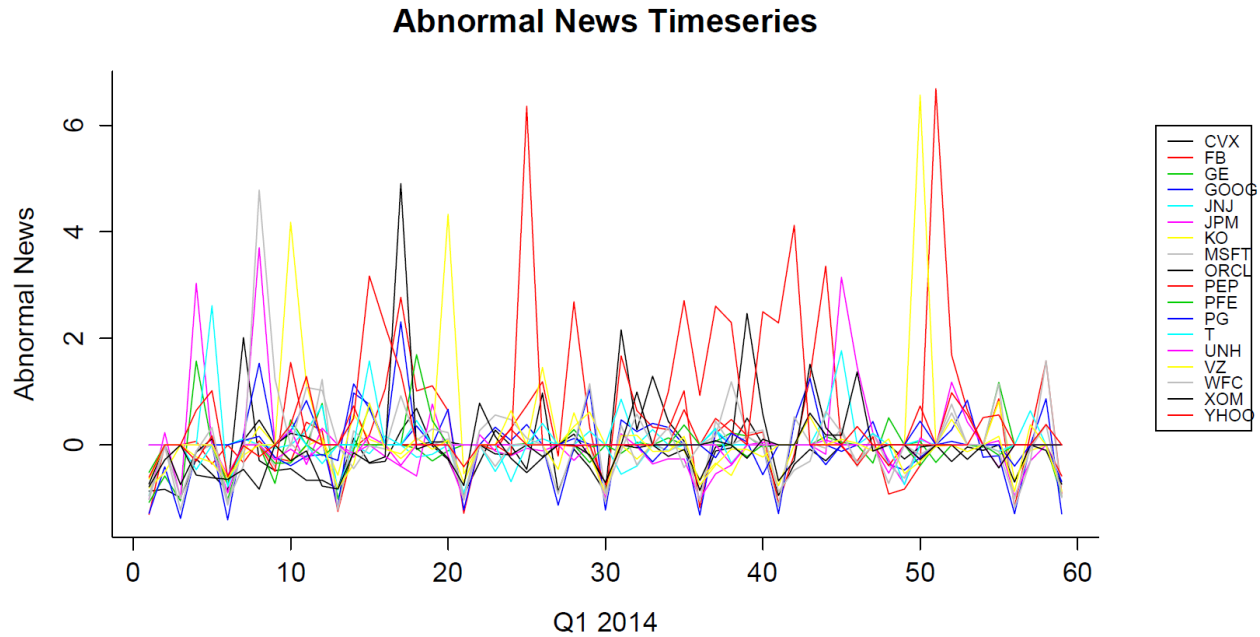
## Abnormal News Timeseries



Fig.2 Time series for Abnormal News

Time series graph for abnormal news volume is shown in Fig.2, A value of 0 means the volume is the same as normal. A positive number represents the number of standard deviations the volume is above the normal. A negative number represents the number of standard deviations the volume is below the normal.

Justification of timeseries goes here..

```
#Generating a Q-Q Plot for Abnormal News
qqnorm(sentimentdata$abnormal_news)
abline(0,1)

#Generating a Q-Q Plot for Sentiment Strength
qqnorm(sentimentdata$sentiment_strength)
abline(0,1))
```

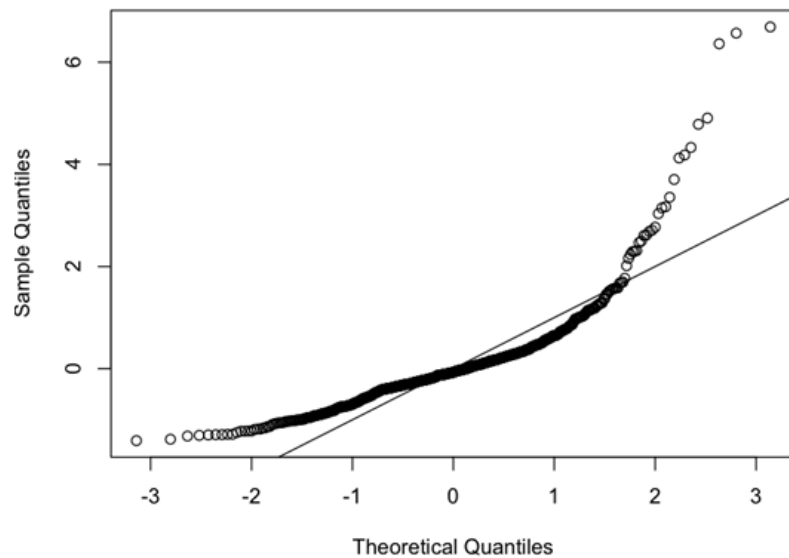**Q-Q Plot for Abnormal News**



Fig.3 Q-Q Plot for Abnormal News

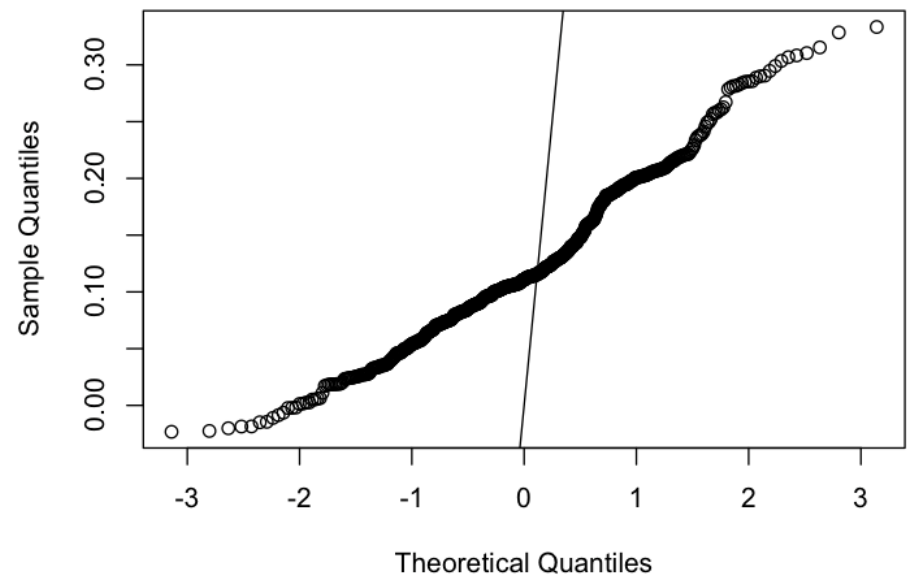**Q-Q Plot for Sentiment Strength**



Fig.4. Q-Q plot for Sentiment Strength

Q-Q plot for abnormal news in Fig.3 indicates that the data is not normally distributed, as there is lot of deviations noticed in this plot. Q-Q plot for sentiment strength shown in Fig.4 indicates the data is not normally distributed as well.

```
# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Return Pct Timeseries", xlab="Q1 2014",ylab='Return Per-
centage', col = 1:ncol(df.ts),bty='L')
```
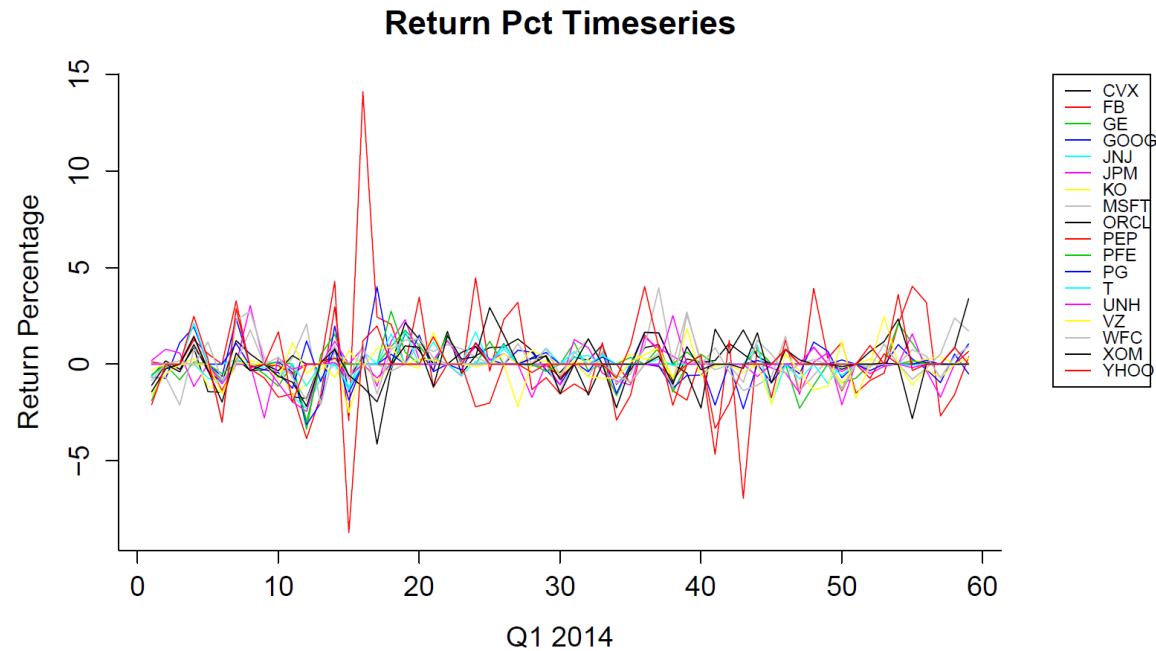


Fig.5 Time series for Return Percent

Time series graph for return percent is shown in Fig.5, A value of 0 means the return percent is flat whereas positive number represents profitable return. A negative number represents the stock returns were negative for that timeframe.

Justification of timeseries goes here..

9

```
# Q-Q plot for return percent
qqnorm(sentimentdata$return_pct)
abline(0,1)
```
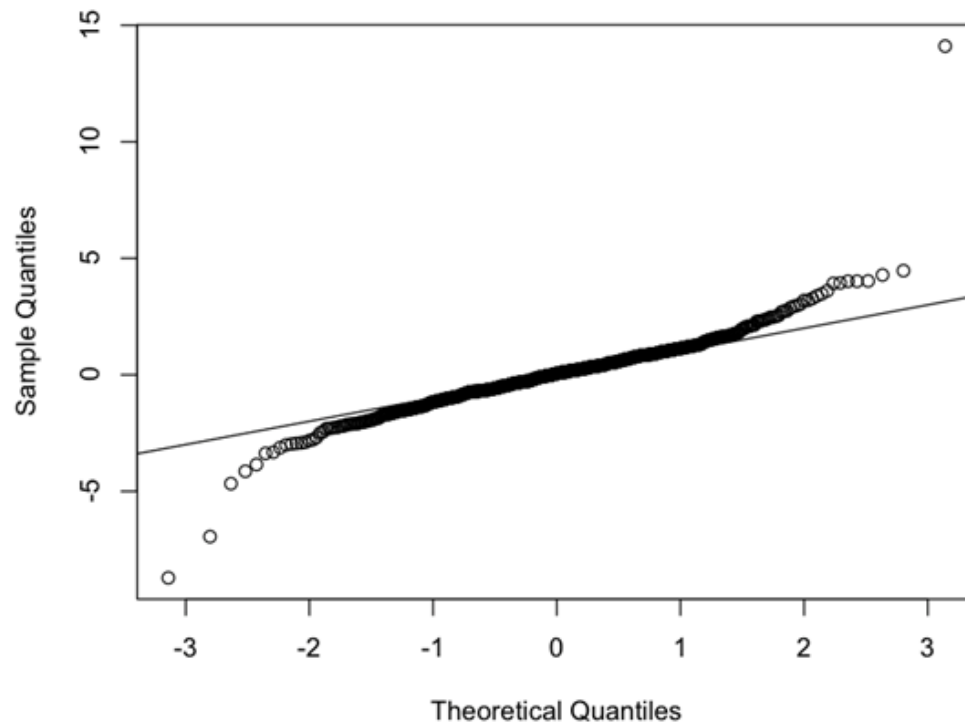
**Q-Q plot for Return percent.**



Fig.6 Q-Q plot for Return Percent

Q-Q plot for 1 day Return percent indicates the data may be distributed normally. It also indicates that there was no significant change in the returns for this time period.

# Plot the time series
```
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Trade Volume", xlab="Q1 2014",ylab='Net Trade Qty in
1000s', col = 1:ncol(df.ts),bty='L')
```
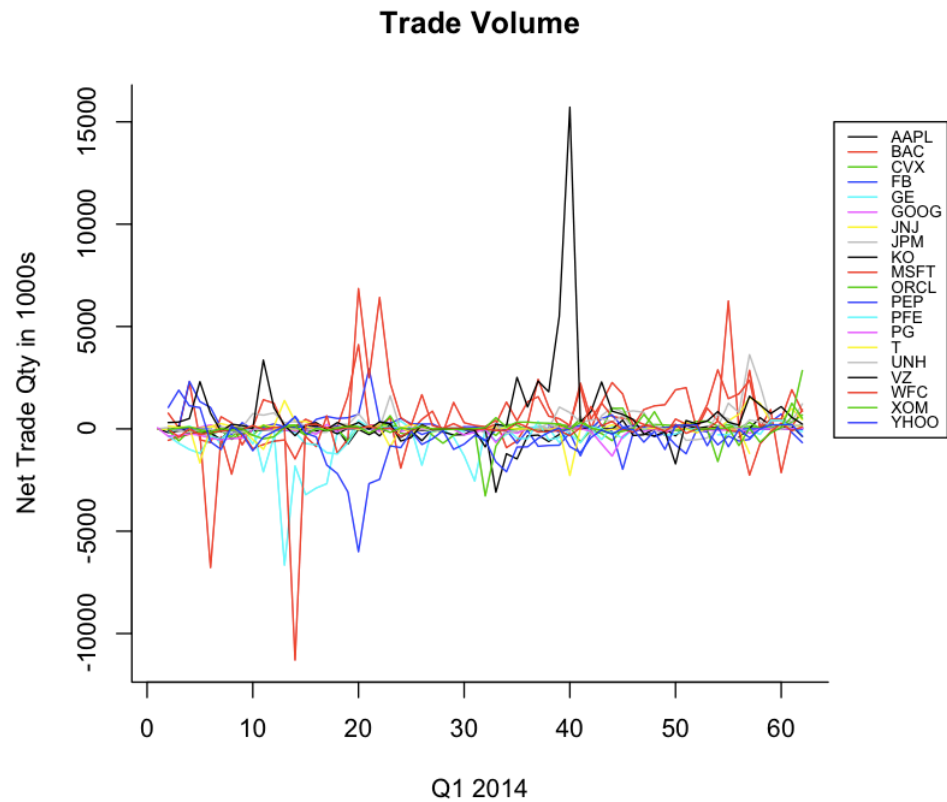


Fig.7  Time series for Trade Volume

Time series graph for net trade volume is shown in Fig.7. Net trade quantity for a given day is arrived by using the formula (Net Qty = Buy Qty – Sell Qty).  A –ve value indicates a SELL where as a +ve value indicates a BUY of the security.

Justification of timeseries goes here..

```
# Plot the time series for research data
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Research Timeseries", xlab="Q1 2014",ylab='Ratings', col
= 1:ncol(df.ts),bty='L')
```
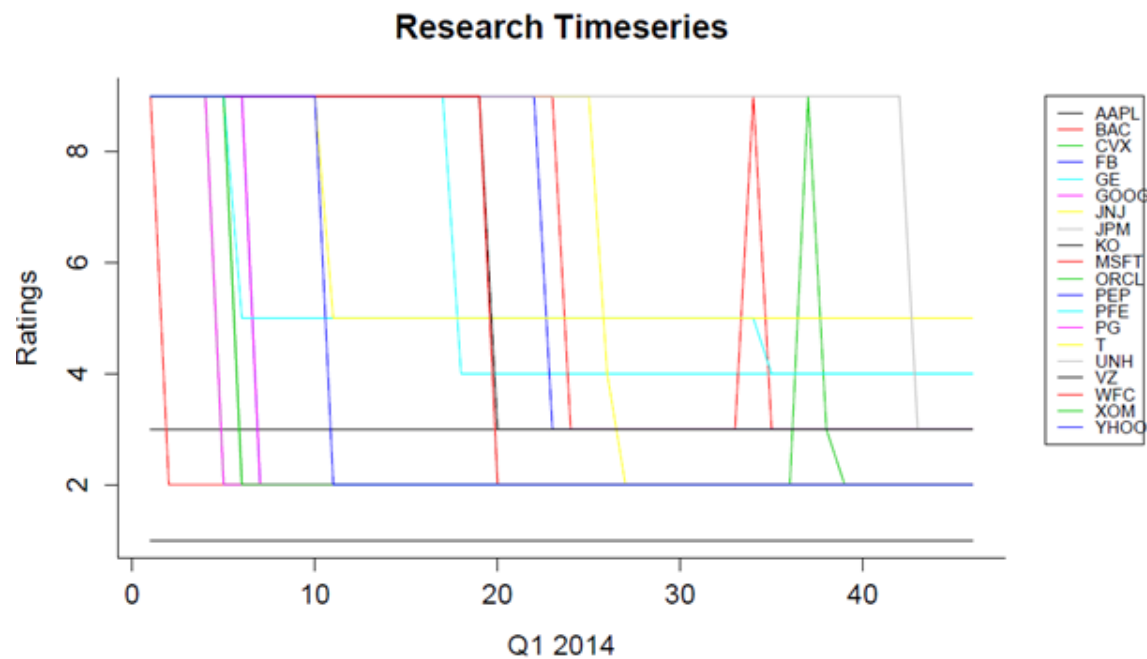


Fig.8  Time series for Research data

Research score is in house proprietary data. Score is usually represented on a 6-point scheme indicating from Strong Buy to Strong sell as well as Not Rated. 1 indicates a strong buy whereas 6 indicate a Strong sell. 9 indicate that the security is not rated. The time series depicted in Fig 8 shows the research scores for the list of securities chosen for this analysis.

Justification of timeseries goes here..

**Summary evaluating all time series goes here..**

## Assessment/interpretation/visualization relative to addressing the business question

We were able to take all trades for the list of 20 securities chosen and got corresponding sentiment score, abnormal news, return percent, research as well as volatility score to perform an extensive analysis of all these attributes using various combination of datasets.

We observe below correlation between volatility, sentiment, abnormal news and return percent.

| | volatility | sentiment_strength | abnormal_news | return_pct |
|---|---|---|---|---|
| volatility | 1.00000000 | -0.16868866 | 0.212394428 | -0.085977642 |
| sentiment_strength | -0.16868866 | 1.00000000 | -0.049765711 | -0.017287544 |
| abnormal_news | 0.21239443 | -0.04976571 | 1.000000000 | -0.008654543 |
| return_pct | -0.08597764 | -0.01728754 | -0.008654543 | 1.000000000 |
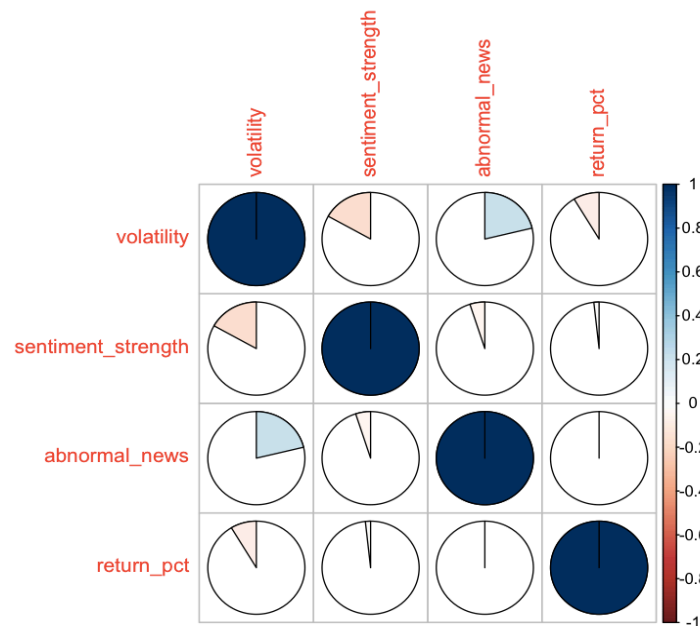
Fig.9 Correlation Plot[5] 1
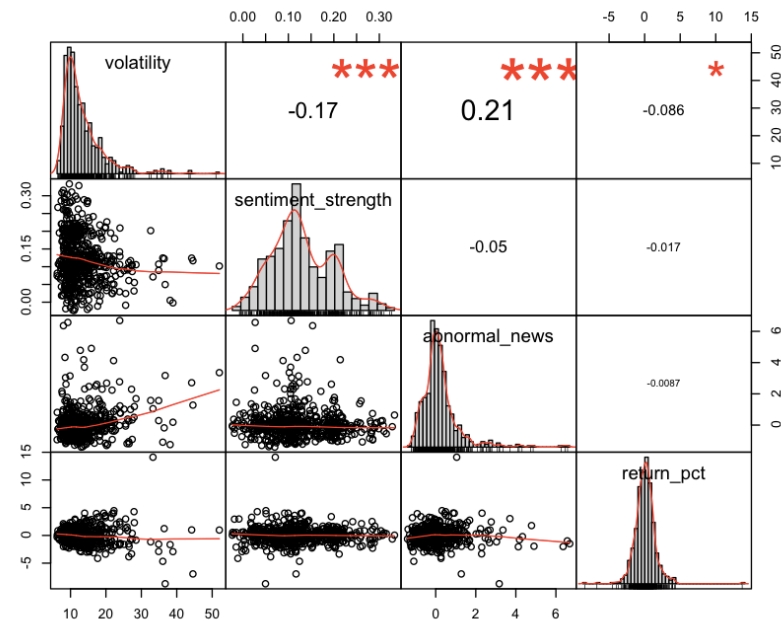


Fig.10 Scatter plot for Volatility, Sentiment Strength, Abnormal News and Return Percent

This observation indicates that there is no strong correlation between the variables. However the standard correlation may not work here based on the number of observation performed. We refer[6] an approach to further fit the correlation based on the number of observation.

In order to prove whether correlation exists between the variables we can use the below formula to calculate the coefficient of correlation.

$|r| >= 2 / \sqrt{n}$, where r=coefficient of correlation and n = number of observations.

In this case, the number of observations is 593

2/Sqrt (593) = 0.08213016

If the value of |r| > 0.08213016, then a relationship exists.

In this case we see a relationship between..

- Volatility and abnormal news
- Volatility and sentiment_strength
- Volatility and return_pct.

In our mash up dataset, sentiment with in house trading; we observe the below correlation

```
                     volatility      net_qty  sentiment_strength  abnormal_news   return_pct
volatility           1.00000000  -0.01134985         -0.17770889     0.22154463  -0.07674261
net_qty             -0.01134985   1.00000000          0.05676034    -0.05467748   0.20343541
sentiment_strength  -0.17770889   0.05676034          1.00000000    -0.05502920  -0.01530827
abnormal_news        0.22154463  -0.05467748         -0.05502920     1.00000000  -0.02344163
return_pct          -0.07674261   0.20343541         -0.01530827    -0.02344163   1.00000000
```
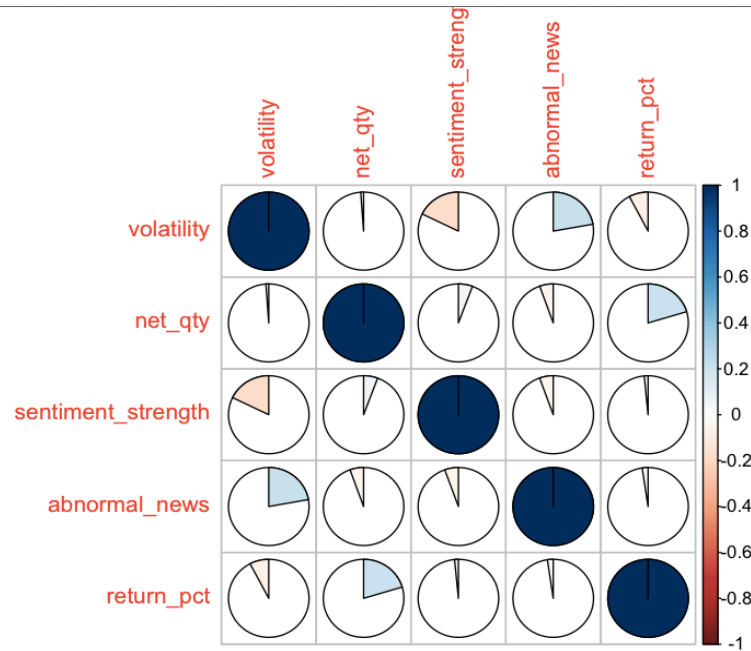
Fig.11 Correlation Plot [5] 2
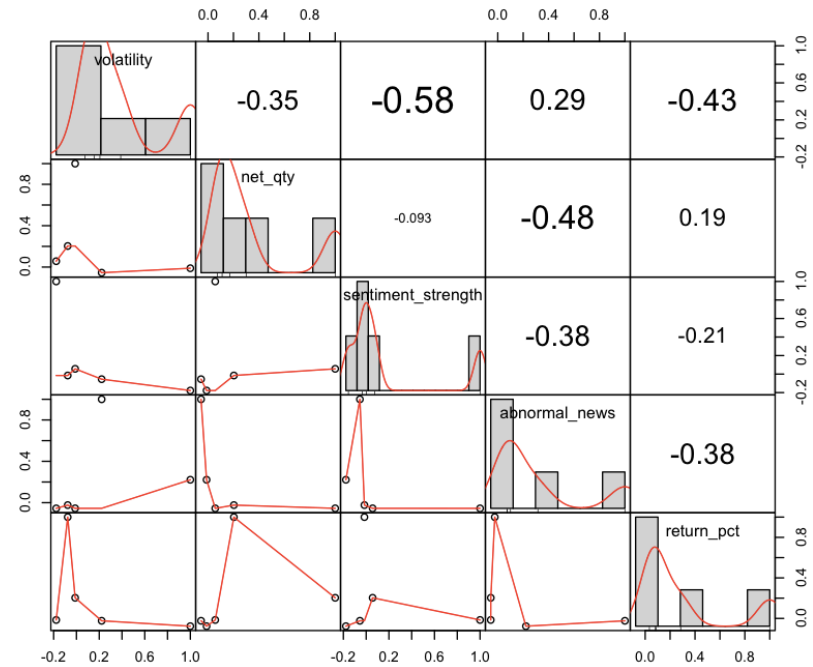


Fig.12 Scatter plot for Volatility, Trade volume, Sentiment Strength, Abnormal News and Return Percent

In order to prove whether correlation exists between the variables we can use the below formula to calculate the coefficient of correlation.

$|r| >= 2 / \sqrt{n}$, where r=coefficient of correlation and n = number of observations.

In this case, the number of observations is 548

$2/\sqrt{548} = 0.08543577$  If the value of $|r| > 0.08543577$, then a relationship exists.

In this case we see a relationship between..

- Volatility and abnormal news
- Volatility and sentiment_strength
- Volatility and return_pct.

We observe the same result that a correlation exists between abnormal news with the volatility. We do see a correlation exists between trading, sentiment and return percent. However we don't see any significant relation statistically between net quantity and volatility.
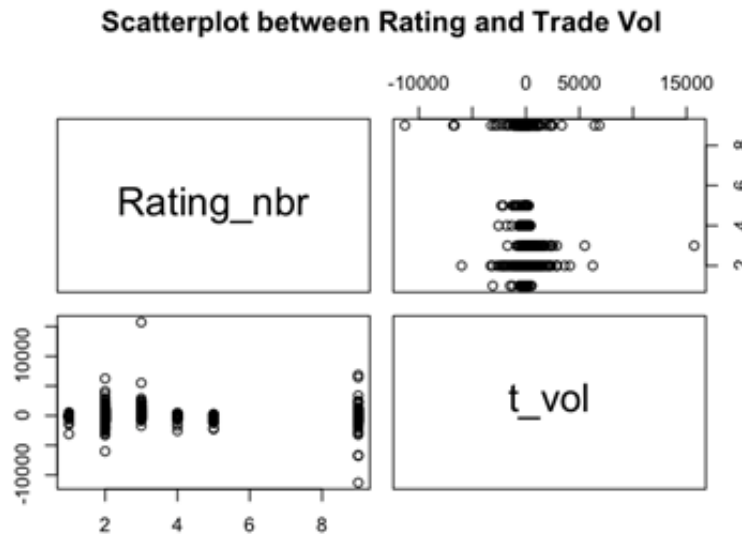


Fig.13 Scatter plot[4] between Rating and Trade Volume

Fig.13 represents scatter plot between Trade and research rating. Here we can observe some interesting fact. Even though some of the securities were rated as a Strong Buy (Rating 1), we see the net trade volume was in negative. That means the security was sold even though research rating was Buy. The same is true for Rating 2(Buy) and 3(Weak Buy) as well. As we analyzed the data we did find the specific security, which was rated strongly, has a negative sentiment at that period, which seems to be reflected portfolio managers trading decision.

## Linear Modeling [7]:

We used linear regression techniques to compare the following relationships

1. Volatility[8] Vs abnormal news
2. Volatility Vs Sentiment strength
3. Volatility Vs return percent

Results Summary:

```
Call:                                              Call:
lm(formula = volatility ~ sentiment_strength + abnormal_news +      lm(formula = volatility ~ sentiment_strength + abnormal_news)
    return_pct)

Residuals:                                         Residuals:
    Min      1Q  Median      3Q     Max                 Min      1Q  Median      3Q     Max
-13.377  -3.500  -1.428   2.003  34.883            -13.053  -3.528  -1.418   2.114  34.555

Coefficients:                                      Coefficients:
                   Estimate Std. Error t value Pr(>|t|)                         Estimate Std. Error t value Pr(>|t|)
(Intercept)         14.7751     0.4696  31.463  < 2e-16 ***    (Intercept)         14.7473     0.4709  31.314  < 2e-16 ***
sentiment_strength -13.1542     3.3018  -3.984 7.65e-05 ***    sentiment_strength -13.0273     3.3119  -3.933 9.40e-05 ***
abnormal_news        1.1992     0.2365   5.070 5.37e-07 ***    abnormal_news        1.2041     0.2373   5.075 5.25e-07 ***
return_pct          -0.3365     0.1552  -2.168   0.0306 *
---                                                ---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.601 on 573 degrees of freedom      Residual standard error: 5.619 on 574 degrees of freedom
  (16 observations deleted due to missingness)                  (16 observations deleted due to missingness)
Multiple R-squared:  0.07774,   Adjusted R-squared:  0.07291    Multiple R-squared:  0.07017,   Adjusted R-squared:  0.06694
F-statistic:  16.1 on 3 and 573 DF,  p-value: 4.638e-10         F-statistic: 21.66 on 2 and 574 DF,  p-value: 8.534e-10
```

Fig.14 Linear Model on Sentiment Data (All attributes)          Fig.15 Linear Model on Sentiment Data (2 attributes)

From the above results summary it can be concluded that volatility had a very high correlation with abnormal news, sentiment strength as well as return percent.

The linear model is in agreement with our analysis. Linear model represented above in Fig.14 is the best fit model as it has adjusted R-squared value of 0.07291.

```
Call:
lm(formula = volatility ~ sentiment_strength + abnormal_news +
    net_qty + return_pct)

Residuals:
   Min     1Q  Median     3Q    Max
-13.672  -3.530  -1.396   1.998  34.481

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.493e+01  4.876e-01  30.609  < 2e-16 ***
sentiment_strength -1.379e+01  3.369e+00  -4.092 4.93e-05 ***
abnormal_news       1.255e+00  2.441e-01   5.138 3.87e-07 ***
net_qty             1.438e-04  2.328e-04   0.618   0.5369
return_pct         -3.103e-01  1.636e-01  -1.897   0.0584 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.653 on 543 degrees of freedom
Multiple R-squared:  0.08273,   Adjusted R-squared:  0.07597
F-statistic: 12.24 on 4 and 543 DF,  p-value: 1.542e-09
```

```
Call:
lm(formula = volatility ~ abnormal_news + net_qty)

Residuals:
   Min     1Q  Median     3Q    Max
-13.527  -3.654  -1.610   2.202  34.425

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.318e+01  2.467e-01  53.437  < 2e-16 ***
abnormal_news 1.312e+00  2.478e-01   5.297 1.71e-07 ***
net_qty       4.235e-06  2.313e-04   0.018    0.985
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.745 on 545 degrees of freedom
Multiple R-squared:  0.04908,   Adjusted R-squared:  0.04559
F-statistic: 14.07 on 2 and 545 DF,  p-value: 1.106e-06
```

Fig.16 Linear Model on Mashup Data (All attribute)

Fig.17 Linear Model on Mashup Data (excluding sentiment and return pct.)

We then extend linear analysis as shown in Fig.16 and Fig.17 to our mashup data set (Sentiment with Trading), which shows some correlation but not a strong correlation with in house trading in line with our earlier observation in the correlation analysis. However, from Fig 13 (scatter plot between rating and trade volume), we observe that in house rating does not necessary impacting the trading volume. When we track that specific security we do see that at that timeframe the sentiment outlook was negative which seems to influence trading decision to go against the research rating.

## Recommendations to Business

The most important observation of this study is the correlation between the 4 major attributes chosen for this analysis. Our observation indicates a very high correlation between market volatility, sentiment strength, and abnormal news volume as well as return percent. The observation also suggests that a certain amount of conservatism could be beneficial in trading during a high volatile market, which might adversely impact a portfolio.

Our findings prove that market sentiment plays considerable role in in-house trading decision. This also leads to an anti-pattern where long term rating suggestion is ignored constantly for a quarter due to negative market sentiment. Even though we theoretically agree that there may be several factors like funds cash position, sudden shareholder redemption that could trigger buy/sell decision; it may also be possible that portfolio pumping [9] is performed. In this paper, we knowingly ignore those additional attributions as out of scope for this analysis.

Based on our findings our recommendation would be to consider market sentiment and volatility while making investment decision.

# References

[1] - http://www.investopedia.com/terms/a/arbitrageur.asp

[2] - New York Times, December 22, 2001.

[3] - New York Times, October 31, 2001.

[4]- http://www.statmethods.net/graphs/scatterplot.html

[5] - http://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

[6] – Correlation coefficient (https://www.youtube.com/watch?v=4EXNedimDMs)

[7] - http://blog.yhathq.com/posts/r-lm-summary.html

[8] - http://www.investopedia.com/terms/v/volatility.asp#ixzz3Z2ow9WQe

[9] – http://www.investopedia.com/terms/p/portfoliopumping.asp

Investor Sentiment, Regimes and Stock Returns; San-Lin Chung_ and Chung-Ying Yehy; December, 2008

# Appendix

**Data sources**

       Raven pack (Market Sentiment, Trade Volatility)
       Internal (Trade and Research)

       List of Securities chosen for analysis:

       PFIZER INC
       GENERAL ELECTRIC CO
       JOHNSON & JOHNSON
       FACEBOOK INC
       VERIZON COMMUNICATIONS INC
       UNITEDHEALTH GROUP INC
       BANK OF AMERICA CORPORATION
       PROCTER & GAMBLE CO
       COCA COLA CO
       CHEVRON CORP
       YAHOO INC
       EXXON MOBIL CORP
       JPMORGAN CHASE & CO
       WELLS FARGO & CO
       MICROSOFT CORP
       ORACLE CORP
       GOOGLE INC A
       PEPSICO INC
       APPLE INC
       AT&T INC

**Meta Data (Fields & Descriptions)**

**Security CUSIP** – CUSIP stands for Committee on Uniform Securities Identification Procedures. Formed in 1962, this committee developed a system (implemented in 1967) that identifies securities, specifically U.S. and Canadian registered stocks, and U.S. government and municipal bonds.

The CUSIP number consists of a combination of nine characters, both letters and numbers, which act as a sort of DNA for the security - uniquely identifying the company or issuer and the type of security. The first six characters identify the issuer and are assigned in an alphabetical fashion; the seventh and eighth characters (which can be alphabetical or numerical) identify the type of issue; and the last digit is used as a check digit.

The CUSIP Service Bureau is operated by Standard & Poor's on behalf of the American Bankers Association (ABA). When setting out to develop the CUSIP system of identification, the ABA basically had two main criteria it was trying to meet. First, it wanted the identification to contain the fewest number of characters possible and to be linked to an alphabetical sequence of issuer names. Secondly, it recognized that the system should be adequate given the current operating requirements while having the flexibility to adapt to any future needs or changes in the operating systems. For more information on the CUSIP system, visit Standard & Poor's CUSIP Service Bureau.

**Ticker** – Alternate Security Identifier. An arrangement of characters (usually letters) representing a particular security listed on an exchange or otherwise traded publicly. When a company issues securities to the public marketplace, it selects an available ticker symbol for its securities which investors use to place trade orders. Every listed security has a unique ticker symbol, facilitating the vast array of trade orders that flow through the financial markets every day.

**Trade Execution Date** – The month, day and year that an order is executed in the market. The trade date is when an order to purchase, sell or otherwise acquire a security is performed. The trade date can apply to the purchase, sale or transfer of bonds, equities, foreign exchange instruments, commodities, futures, etc. In some cases, the trade date will be recorded on the previous day, for trades that are executed very early, or on the next day, in the case of orders that are executed very late in the day.

**Trade Side** – Trade Side could be Buy or Sell

**Price** – A share price is the price of a single share of a number of saleable stocks of a company, derivative or other financial asset. In layman's terms, the stock price is the highest amount someone is willing to pay for the stock, or the lowest amount that it can be bought for.

**Market Cap (Market Capitalization)** – The total dollar market value of all of a company's outstanding shares. Market capitalization is calculated by multiplying a company's shares outstanding by the current market price of one share. The investment community uses this figure to determine a company's size, as opposed to sales or total asset figures. If a company has 35 million shares outstanding, each with a market value of $100, the company's market capitalization is $3.5 billion (35,000,000 x $100 per share).

**Market Cap Type** - Company size is a basic determinant of asset allocation and risk-return parameters for stocks and stock mutual funds. The term should not be confused with a company's "capitalization," which is a financial statement term that refers to the sum of a company's shareholders' equity plus long-term debt.

The stocks of large, medium and small companies are referred to as large-cap, mid-cap, and small-cap, respectively. Investment professionals differ on their exact definitions, but the current approximate categories of market capitalization are:

Large Cap: $10 billion plus and include the companies with the largest market capitalization.

Mid Cap: $2 billion to $10 billion

Small Cap: Less than $2 billion

In order to make an investment decision, you may need to factor in the market cap of some investments.

**Volatility** [8] – A statistical measure of the dispersion of returns for a given security or market index. Volatility can either be measured by using the standard deviation or variance between returns from that same security or market index. Commonly, the higher the volatility, the riskier the security.

A variable in option pricing formulas showing the extent to which the return of the underlying asset will fluctuate between now and the option's expiration. Volatility, as expressed as a percentage coefficient within option-pricing formulas, arises from daily trading activities. How volatility is measured will affect the value of the coefficient used.

**Trade Volume** - Trading volume, or volume, is the number of shares or contracts that indicates the overall activity of a security or market for a given period. Trading volume is an important technical indicator an investor uses to confirm a trend or trend reversal. Volume gives an investor an idea of the price action of a security and whether he should buy or sell the security.

Trading volume can help an investor identify momentum in a stock and confirm a trend. If trading volume increases, prices generally move in the same direction. That is, if a security is continuing higher in an uptrend, the volume of the security should also increase and vice versa.

**Research Score –** Research score is in house proprietary data. Score is usually represented on a 6-point scheme indicating from Strong Buy to Not Rated.

| | |
|---|---|
| 1 | Strong Buy |
| 2 | Buy |
| 3 | Weak Buy |
| 4 | Weak Sell |
| 5 | Sell |
| 6 | Strong Sell |
| 9 | Not Rated |

**Sentiment Indicator**

Measure the media sentiment for a company over the short and long-term horizons. Two types of sentiment indicators are provided:

•       A strength indicator which considers novel news events over the previous 91 day period and incorporates a decay function to give more weight to recent news.

•       A daily average sentiment indicator which gives a sentiment score for a day and allows customized strength indicators to be built for longer trading horizons.

Abnormality Indicators

These indicators tell when the sentiment or news coverage for a particular company has deviated from the norm and are useful for finding companies that are moving into or out of the media spotlight and are therefore subject to change.

**Sentiment Strength**

A nullable numeric value between -1 and +1 with up to 5 decimal points representing the aggregated sentiment strength over the previous 91 days. A value of -1 is highly negative, a value of +1 is highly positive and a value of 0 is neutral. An empty value means that there was no news events over the last 91 days and therefore no sentiment associated with the company.

Only novel news items that have non-neutral sentiment are included in this computation. Certain types of news stories categorized as "Order Imbalance", "Insider Trading" and "Technical Analysis" are excluded as they tend to add noise given their lack of sentiment, high volume and frequency. To allow for more recent events to have a greater impact on the company's sentiment, an exponential sentiment decay and a time weight function are incorporated. Both are based on the company's event volume, and hence sentiment will decay more rapidly with additional company events.

**Abnormal News**

A numerical value with up to 5 decimal places that captures how different the actual news volume over the past 24 hours is when compared to the normal news volume for the entity. A value of 0 means the volume is the same as normal. A positive number represents the number of standard deviations the volume is above the normal. A negative number represents the number of standard deviations the volume is below the normal. The value is calculated by taking the difference between the company's news volume over the previous 24 hour period (NEWS_VOLUME_1D) vs. the company's average 24 hour news volume over the past 365 days. The difference is divided by the volatility of the 24 hour news volume from the past 365 days (i.e. Z-Score calculation).

**Return Percent (Day)**

The gain or loss of a security in a particular period(1 DAY for this analysis). The return consists of the income and the capital gains relative on an investment. It is usually quoted as a percentage.

The general rule is that the more risk you take, the greater the potential for higher return - and loss.

**R packages utilized**

1. PerformanceAnalytics
   o   For correlation chart

2. corrplot
    o For correlation plot
3. Sp
    o Prerequisite for sqldf
4. Gsubfn
    o Prerequisite for sqldf
5. Tcltk
    o Prerequsite for sqldf
6. Sqldf
    o Manipulate data frames using sql statements and insert into SQL server
7. qqnorm
    o Create Q-Q plots
8. zoo
    o na.locf - Generic function for replacing each NA with the most recent non-NA prior to it.

# Project R Code

```
sentimentdata <- read.csv("sentiment_data.csv", stringsAsFactors=FALSE)
str(sentimentdata)
sentimentdata$trade_dt <- as.Date(sentimentdata$trade_dt,'%m/%d/%y')
sentimentdata$eff_date <- as.Date(sentimentdata$eff_date,'%m/%d/%y')

str(sentimentdata)

## Get the sentiment strength data
senti_strength <- sentimentdata[,c('symbol','eff_date','sentiment_strength')]

## Convert the Long Fomrat to Wide Format using the reshape
senti_strength.ts <- reshape(senti_strength, timevar = "eff_date", idvar = c("symbol"),direction = "wide")
```

```
## Name the columns
names(senti_strength.ts) <-
c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","24","25","26","27","28","29","30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45","46","47","48","49","50","51","52","53","54","55","56","57","58","59","60")

# Transpose the data set
df <- t(senti_strength.ts[1:18,2:60])

# Convert it to a data frame
df <- as.data.frame(df)

# Put the column names as the names of the securities
names(df) <-
c("CVX","FB","GE","GOOG","JNJ","JPM","KO","MSFT","ORCL","PEP","PFE","PG","T","UNH","VZ","WFC","XOM","YHOO")

# assigning 0 to NA
df[is.na(df)] <- 0

# Create the timeseries
df.ts <- ts(df, start = 1, frequency = 1)

# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Sentiment Strength Timeseries", xlab="Q1 2014",ylab='Sentiment Strength between -1 and 1', col = 1:ncol(df.ts),bty='L')

#Add the Legends
legend(65,0.30, inset=c(-0.3,-0.2), colnames(df), col=1:ncol(df), lty=1, cex=.65)
qqnorm(sentimentdata$sentiment_strength)
abline(0,1)

## Get the abnormal news data
abnormal_news <- sentimentdata[,c('symbol','eff_date','abnormal_news')]

## Convert the Long Fomrat to Wide Format using the reshape
```

```
abnormal_news.ts <- reshape(abnormal_news, timevar = "eff_date", idvar = c("symbol"),direction = "wide")

## Name the columns
names(abnormal_news.ts) <-
c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","
 24","25","26","27","28","29","30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45"
 ,"46","47","48","49","50","51","52","53","54","55","56","57","58","59","60")

# Transpose the data set
df <- t(abnormal_news.ts[1:18,2:60])

# Convert it to a data frame
df <- as.data.frame(df)

# Put the column names as the names of the securities
names(df) <-
c("CVX","FB","GE","GOOG","JNJ","JPM","KO","MSFT","ORCL","PEP","PFE","PG","T","UNH","VZ","WFC","XOM","YHOO")

# assigning 0 to NA
df[is.na(df)] <- 0

# Create the timeseries
df.ts <- ts(df, start = 1, frequency = 1)

# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Abnormal News Timeseries", xlab="Q1 2014",ylab='Abnormal News', col =
1:ncol(df.ts),bty='L')

#Add the Legends
legend(65,6, inset=c(-0.3,-0.2), colnames(df), col=1:ncol(df), lty=1, cex=.65)

#Draw Q-Q plot
qqnorm(sentimentdata$abnormal_news)
abline(0,1)
```

```
## Get the Return Percentage Data
return_pct <- sentimentdata[,c('symbol','eff_date','return_pct')]

## Convert the Long Fomrat to Wide Format using the reshape
return_pct.ts <- reshape(return_pct, timevar = "eff_date", idvar = c("symbol"),direction = "wide")

## Name the columns
names(return_pct.ts) <- c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","
 24","25","26","27","28","29","30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45"
 ,"46","47","48","49","50","51","52","53","54","55","56","57","58","59","60")

# Transpose the data set
df <- t(return_pct.ts[1:18,2:60])

# Convert it to a data frame
df <- as.data.frame(df)

# Put the column names as the names of the securities
names(df) <-
c("CVX","FB","GE","GOOG","JNJ","JPM","KO","MSFT","ORCL","PEP","PFE","PG","T","UNH","VZ","WFC","XOM","YHOO")
# assigning 0 to NA

df[is.na(df)] <- 0
# Create the timeseries

df.ts <- ts(df, start = 1, frequency = 1)

# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Return Pct Timeseries", xlab="Q1 2014",ylab='Return Percentage', col =
1:ncol(df.ts),bty='L')

#Add the Legends
legend(65,15, inset=c(-0.3,-0.2), colnames(df), col=1:ncol(df), lty=1, cex=.65)

#Draw Q-Q plot
```

```
qqnorm(sentimentdata$return_pct)
abline(0,1)

#Derive correlation
mydata <- sentimentdata[,c(2,8,9,10)]
mydata <- mydata[complete.cases(mydata),]
cor(mydata)

#Draw plot
#http://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
install.packages("corrplot")
library(corrplot)
m <- cor(mydata)
corrplot(m, method = "pie")

install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
chart.Correlation(mydata)

#Apply Linear model
attach(sentimentdata)
lmd <- lm(volatility ~ sentiment_strength + abnormal_news + return_pct)
lmd
summary(lmd)
lmd <- lm(volatility ~ sentiment_strength + abnormal_news)
lmd
summary(lmd)

#Read Trade data
tdata <- read.csv("trade_data.csv", stringsAsFactors=FALSE)
str(tdata)
tdata$trade_execution_dt <- as.Date(tdata$trade_execution_dt,'%m/%d/%y')
#Take net quantity as '000 to avoid exponential values
tdata$net_qty <- tdata$net_qty/1000
tdata.ts <- reshape(tdata, timevar = "trade_execution_dt", idvar = c("symbol"),direction = "wide")
```

```
names(tdata.ts) <-
c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","24","25","26","27","28","29","
30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45","46","47","48","49","50","51","52","53","54","55","56","5
7","58","59","60","61","62","63")
# Transpose the data set
df <- t(tdata.ts[1:20,2:63])
# Convert it to a data frame
df <- as.data.frame(df)
# Put the column names as the names of the securities
names(df) <-
c("AAPL","BAC","CVX","FB","GE","GOOG","JNJ","JPM","KO","MSFT","ORCL","PEP","PFE","PG","T","UNH","VZ","WFC","XOM","
YHOO")
# Create the timeseries
df.ts <- ts(df, start = 1, frequency = 1)
# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Trade Volume", xlab="Q1 2014",ylab='Net Trade Qty in 1000s', col = 1:ncol(df.ts),bty='L')
#Add the Legends
legend(65,15000, inset=c(-0.3,-0.2), colnames(df), col=1:ncol(df), lty=1, cex=.65)

#Now link both the data set
#Sentiment and Trade data using sqldf

m <- sentimentdata
z <- tdata
str(z)
str(m)
#Library required for sqldf
library(sp)
library(gsubfn)
library(tcltk)
library(sqldf)

#Create merged data set (Trade and Sentiment)
tempdata <- sqldf("select a.volatility,b.net_qty,a.sentiment_strength,a.abnormal_news,a.return_pct
from m a, z b
```

```
where a.symbol = b.symbol
and a.trade_dt = b.trade_execution_dt
")

cor(tempdata)
m1 <- cor(tempdata)
corrplot(m1, method = "pie")
chart.Correlation(m1)

#Linear Model
lmd2 <- lm(volatility ~ sentiment_strength + abnormal_news + net_qty+return_pct)
lmd2
summary(lmd2)
lmd3 <- lm(volatility ~ abnormal_news + net_qty)
lmd3

# Load reasearch Data
library(zoo)

## Read the research Data
researchdata <- read.csv("research_data2.csv", stringsAsFactors=FALSE)

## Get the research Data for columns needed
researchdata <- researchdata[,c('Symbol','Rating_nbr','publish_date')]

## Convert the Long Format to Wide Format using the reshape
research.ts <- reshape(researchdata, timevar = "publish_date", idvar = c("Symbol"),direction = "wide")

## Name the columns
names(research.ts) <-
c("Symbol","1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","
        24","25","26","27","28","29","30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45"
        ,"46")
# Replace eash NA with the most recent non-NA prior
research.ts$`1`[is.na(research.ts$`1`)] <- 9
```

```
# Transpose the data set
df <- t(research.ts[1:20,2:47])

# Convert it to a data frame
df <- as.data.frame(df)

# assigning previous values

for(i in 1:20) { df[,i] <- na.locf(df[,i])}

names(df) <- c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20")


# Create the timeseries

df.ts <- ts(df, start = 1, frequency = 1)

# Plot the time series
par(xpd=T, mar=par()$mar+c(0,0,0,5))
plot.ts(df.ts, plot.type="single", main="Research Timeseries", xlab="Q1 2014",ylab='Ratings', col = 1:ncol(df.ts),bty='L')

#Now read research rating data
rdata <- read.csv("research_data2.csv",stringsAsFactors=FALSE)
str(rdata)

rdata$publish_date <- as.Date(rdata$publish_date,'%m/%d/%y')

#Link Trade and Research using sqldf
z <- tdata
r <- rdata
#Create merged data set (Trade and research)
tempdata1 <- sqldf("select Rating_nbr,b.net_qty t_vol
from r a, z b
where a.Symbol = b.symbol
and publish_date = b.trade_execution_dt
")
```

```
m2 <- cor(tempdata1)
corrplot(m2, method = "pie")
attach(tempdata1)

pairs(~Rating_nbr+t_vol,data=tempdata1,
    main="Scatterplot between Rating and Trade Vol")

plot(Rating_nbr, t_vol, xlab="x-label", ylab="y-label", pch=21)
abline(lm(Rating_nbr ~ t_vol)
```

# Pre cursor deliverables

## Final Project Data Discovery: 1 Page Proposal

## Correlation between Market Sentiment, Research and Trading / Predicting Trading Volatility

**Selected dataset(s)**

- Large cap securities being researched by Fidelity
- Market sentiment score
- Proprietary Research
- Proprietary Trade

**Source(s)**

- Fidelity Investments
- Ravenpack -

**Business context**

Financial analysts have always indicated that market volatility impacts trading to a large extent. Access to volatility prediction will help Fidelity fund managers in taking appropriate trading decision in the best interest of the shareholders.

We are trying to analyze market sentiment (Ravenpack), proprietary research and trade data to provide meaningful correlations that could lead to prediction of trading volatility.

Fidelity Investments has in house research capabilities, which provide research advisory (Buy/Sell indicator) for various instruments that could be linked to market sentiment and historical trade information to predict trade volatility.

**Assessment of Data Quality**

Data is available in public domain and in house for this work. Data is currently in a relational format. For the purpose of this project data will be extracted in a csv format. Historical data will be analyzed for meaningful correlation and provide predictions.

**Relationship/affinity of data sets (if more than one data set):**

Data linkage is possible using Security CUSIP and ticker attributes.

Initial data attributes identified for this project are attached. Team is working on refining and cleansing the dataset.

Attributes in Research Dataset:

| Seq Num | Data Attribute | Data Type | Attribute Description |
|---|---|---|---|
| 1 | Security CUSIP | VARCHAR2 | Security Identifier |
| 2 | Ticker | VARCHAR2 | Security Identifier |
| 3 | Publication Date | Data | Research publication date |
| 4 | Rating | VARCHAR2 | Research Rating (Buy, Sell, Strong Buy, Strong Sell, Weak Buy, Weak Sell) |

Attributes in Trade Dataset:

| Seq | Data Attribute | Data Type | Attribute Description |
|---|---|---|---|

| Num | | | |
|---|---|---|---|
| 1 | Security CUSIP | VARCHAR2 | Security Identifier |
| 2 | Ticker | VARCHAR2 | Security Identifier |
| 3 | Trade Date | Date | Date of Trade |
| 4 | Trade Side Code | VARCHAR2 | Trade Side Code (BUY/SELL) |
| 5 | Trade Qty | NUMBER | Quantity of Trade |
| 6 | Gross USD Amt | NUMBER | Trade Amount in USD |

Attributes in Sentiment Dataset:

| Seq Num | Data Attribute | Data Type | Attribute Description |
|---|---|---|---|
| 1 | Security CUSIP | VARCHAR2 | Security Identifier |
| 2 | Ticker | VARCHAR2 | Security Identifier |
| 3 | Trade Date | DATE | Trade Execution date |
| 4 | Trade Side | VARCHAR2 | Sell/Buy |
| 5 | Price | NUMBER | Trade price |
| 6 | Market Cap | NUMBER | Market Cap Value |
| 7 | Market Cap Type | VARCHAR2 | Market Cap Types like Large Cap, Mid Cap, Small Cap |
| 8 | Volatility | VARCHAR2 | Volatility  score in low, high |
| 9 | Volume | NUMBER | Trading Volume |
| 10 | Sentiment | NUMBER | Sentiment score from Raven pack |

# Final Project Data Analysis Plan - 2 page Proposal

## Correlation between Market Sentiment, Research and Trading / Predicting Trading Volatility

### Business Question
How can active fund managers make informed trading decisions based on market sentiment, research and volatility?

### Understanding the Attributes

**Sentiment Indicators**

Measure the media sentiment for a company over the short and long-term horizons. Two types of sentiment indicators are provided:

- A *strength* indicator which considers novel news events over the previous 91 day period
and incorporates a decay function to give more weight to recent news.
- A *daily average* sentiment indicator which gives a sentiment score for a day and allows
customized strength indicators to be built for longer trading horizons.

**Abnormality Indicators**

These indicators tell when the sentiment or news coverage for a particular company has deviated from the norm and are useful for finding companies that are moving into or out of the media spotlight and are therefore subject to change.

**SENTIMENT_STRENGTH_91D**

A nullable numeric value between -1 and +1 with up to 5 decimal points representing the
aggregated sentiment strength over the previous 91 days. A value of -1 is highly negative, a value of +1 is highly positive and a value of 0 is neutral. An empty value means that there was no news events over the last 91 days and therefore no sentiment associated with the company.

Only novel news items that have non-neutral sentiment are included in this computation. Certain
types of news stories categorized as "Order Imbalance", "Insider Trading" and "Technical
Analysis" are excluded as they tend to add noise given their lack of sentiment, high volume and
frequency. To allow for more recent events to have a greater impact on the company's sentiment, an exponential sentiment decay and a time weight function are incorporated. Both are based on the company's event volume, and hence sentiment will decay more rapidly with additional company events.

**ABNORMAL_NEWS_VOLUME_1D**

A numerical value with up to 5 decimal places that captures how different the actual news volume over the past 24 hours is when compared to the normal news volume for the entity. A value of 0 means the volume is the same as normal. A positive number represents the number of standard deviations the volume is above the normal. A negative number represents the number of standard deviations the volume is below the normal. The value is calculated by taking the difference between the company's news volume over the previous 24 hour period (NEWS_VOLUME_1D) vs. the company's average 24 hour news volume over the past 365 days. The difference is divided by the volatility of the 24 hour news volume from the past 365 days (i.e. Z-Score calculation).

**SECURITY_RETURNS_1D_PCT**
The gain or loss of a security in a particular period(1 DAY for this analysis). The return consists of the income and the capital gains relative on an investment. It is usually quoted as a percentage.

The general rule is that the more risk you take, the greater the potential for higher return - and loss.

# Data Acquisition Sequence

Discussions with subject matter experts provided some leads in to the data acquisition.

Raven pack sentiment data(subscription) was used to get sentiment score and abnormal news

Fidelity proprietary research is in house

Trade data is in house

This data is pulled in a csv format to be able to read in to R

# Inferences from Initial Data Analysis

What are the data sets for linkage/analysis

- Sentiment Score (Date, Sentiment Strength)
- Abnormal news (Date, Abnormal News Volume)
- Returns data (Date, Return percentage)
- Research Score (Date, Research Score)
- Trade Volume (Date, Trade Volume)

What's common among the data sets to provide the linkage

- Visual data analysis was performed on this data and the following inferences are made.

Initial plan was to use security identifier to link the datasets; however after initial analysis of each of these datasets we see a better a better linkage of data is possible using date attribute available in this datasets.

## Review Submission History: Final Project 3 Page Project Report

### Correlation between Market Sentiment, Research and Trading / Predicting Trading Volatility

### Pre-processing of Data

Manual pre-processing of data was performed to get the data in the desired format. SQL queries were run against database to get the data and saved the result as a csv file to be able to upload in to R.

### Data Acquisition/Clean up and Visualization

Using data for a single security we tried to observe

- Sentiment Data
- Abnormal News Volume
- Return Pct
- Trade volume
- Research Rating

As part of data cleansing the research indicator was removed from the Sentimental dataset.

We then observe the time series graph for Sentiment Data and Abnormal News Volume.
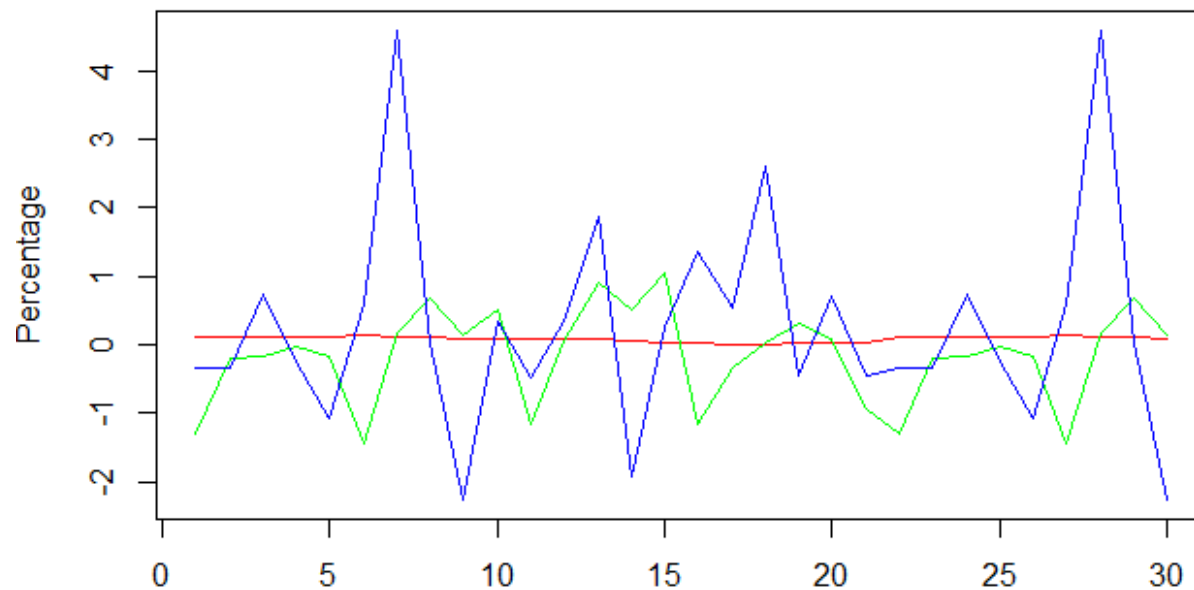
**Sentiment Strength**

**Abnormal NewsVolume**
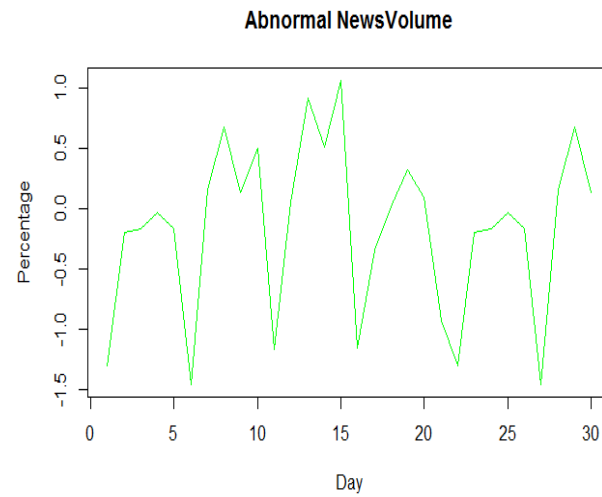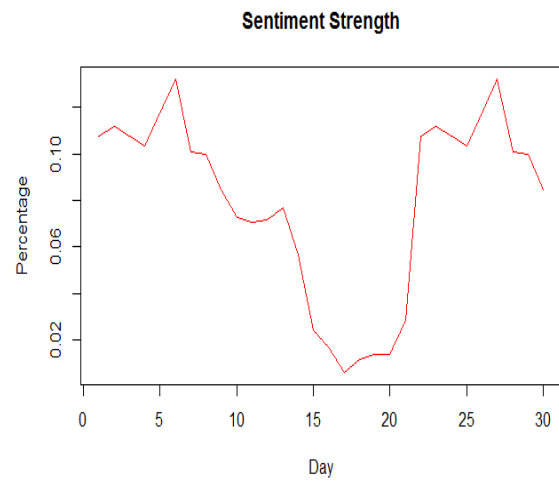
Furthermore we observe the return percentage
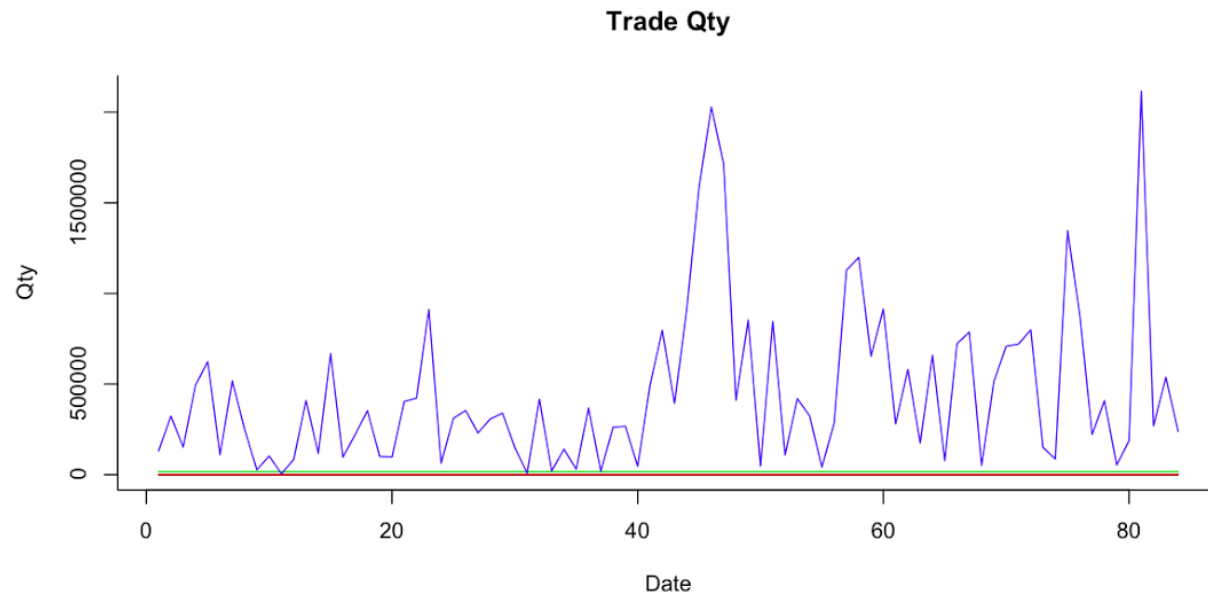
**Return Percentage**
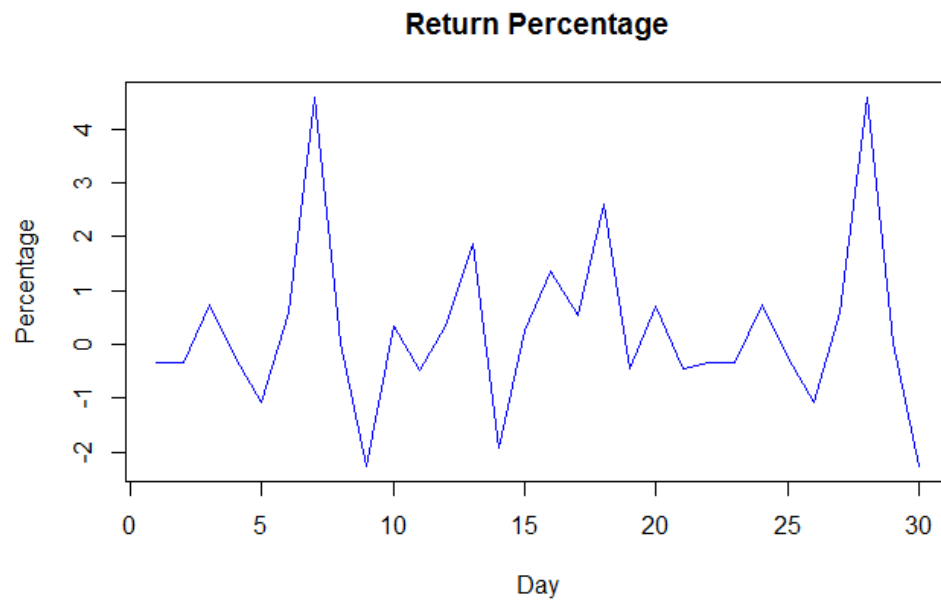
We can clearly see the correlation with Sentiment, Abnormal News and Returns.

## Sentiment, Abnormal News and Returns



For the same time period in-house Trading process shows the alignment with the sentiments.

**Trade Qty**

However in this graph we have not considered the net quantity (Buy-Sell), which we will refine in future stages of this project.

The in house research data usually valid for 40days after it become publish. To prepare a time series we have generated data for all days. Below graph was plotted based on data from 3/2013 till 09/2014 for a single security. We observe that rating was not changed after 7/2013 onwards.

The Research Ratings for a Security
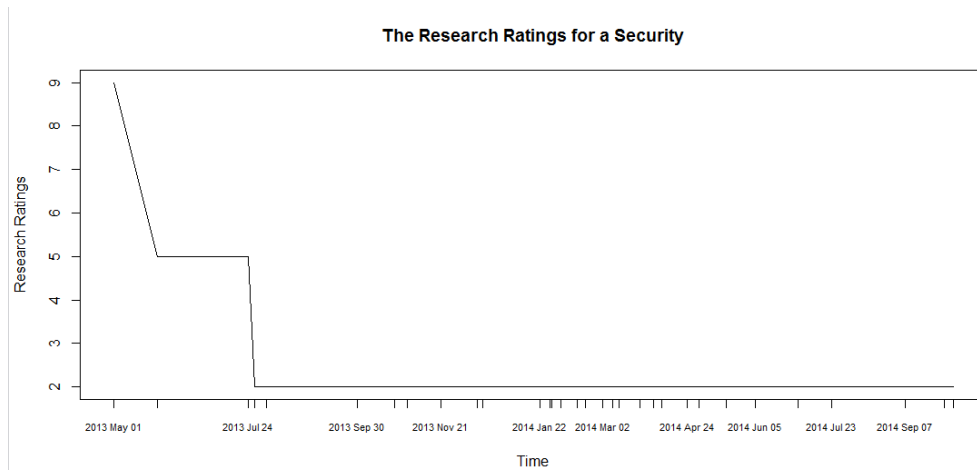
## R Script

#Reading in sentiment data for a single security.

sen_abn_res<-read.csv("c:/sent_res_fb_data.csv")

#Data clean up and preparation steps

sen_res_data<-as.data.frame(t(sen_abn_res), header=TRUE)

names(sen_res_data)<-sen_res_data[1, ]

colnames(sen_res_data) = unlist(sen_res_data[1, ])

sen_res_data<-sen_res_data[-1, ]

# Removed the research indicator as it did not add value to the dataset.

sen_res_data<-sen_res_data[-4, ]

**Preparing time series data frames for each of the datasets**

# Sentiment Strength time series plot

sent_data_ts<-ts(as.vector(t(as.matrix(sen_res_data[1, ]))),start=c(1,1), end=c(30,1), frequency=1)

plot(sent_data_ts, plot.type="single", main="Sentiment Strength",  xlab="Day",ylab='Percentage', col=c("red") , lty=1)

#Abnormal News Volume Timeseries Plot

abn_data_ts<-ts(as.vector(t(as.matrix(sen_res_data[2, ]))),start=c(1,1), end=c(30,1), frequency=1)

plot(abn_data_ts, plot.type="single", main="Abnormal NewsVolume",  xlab="Day",ylab='Percentage', col=c("green") , lty=1)


#Returns Data Timeseries plot

ret_data_ts<-ts(as.vector(t(as.matrix(sen_res_data[3, ]))),start=c(1,1), end=c(30,1), frequency=1)

plot(res_data_ts, plot.type="single", main="Return Percentage",  xlab="Day",ylab='Percentage', col=c("blue") , lty=1)

#Combined Time series plot

sen_abn_ret_ts <- cbind(sent_data_ts, abn_data_ts, ret_data_ts)

plot.ts(sen_abn_ret_ts, plot.type="single", main="Sentiment, Abnormal News and Returns",  xlab="Day", ylab='Percentage', col=c("red", "green", "blue"), lty=1)

# Trade data acquisition

tstrd <- read.csv("trade_f_ts_2014.csv")

# Data cleansing and formatting of trade data

tstrd$TRADE_EXECUTION_DT <- as.Date(tstrd$TRADE_EXECUTION_DT,format="%d-%b-%y")

# Developing a time series plot for trade data

```
tstrd.ts <- ts(tstrd)

plot.ts(tstrd.ts, plot.type="single", main="Trade Qty", xlab="Date",ylab='Qty', col = 1:ncol(tstrd.ts),bty='L')

#Acquisition of Research Data

r <- read.csv("r_ts.csv")

#Data cleansing and formatting of research data

r1 <- r[,c('MONTH_DATA', 'RATING_NBR')]

r1$MONTH_DATA <- as.Date(r1$MONTH_DATA,"%m/%d/%Y")

#Developing a plot of Research data

plot(RATING_NBR ~ MONTH_DATA, r1, xaxt = "n", type = "l",main = "The Research Ratings for a Security", xlab =
"Time", ylab = "Research Ratings")

axis(1, r1$MONTH_DATA, format(r1$MONTH_DATA, "%Y %b %d"), cex.axis = .7)
```

## Sample Dataset

### Sentiment Data

| CUSIP | TMSTMP_UTC | SENTIMENT_STRENGTH_91D | ABNORMAL_NEWS_1D | RETURN_1D_PCT |
|---|---|---|---|---|
| 12345 | 5/19/2014 | 0.87154 | -1.28073 | 1.18492 |
| 12345 | 5/20/2014 | 0.09358 | -0.42608 | 0.0198484 |
| 12345 | 5/7/2013 | 0.10009 | -0.32795 | -0.445402 |

**Rating Data**

| CUSIP | INSTRUMENT_LONG_NAME | RATING_SCHEME_NAME | RATING_NAME | PUBLICATION_TMSTMP |
|---|---|---|---|---|
| 12345 | ABCD INC | 6 POINT | Weak Buy | 10/20/2014 20:30 |
| 12345 | ABCD INC | 6 POINT | Weak Buy | 10/17/2014 2:35 |
| 12345 | ABCD INC | 6 POINT | Weak Buy | 9/29/2014 19:59 |

**Trade Data**

| CUSIP | MAR-KET_EXCHANGE_CD | TRADE_EXECUTION_DT | TRADE_EXECUTION_TMSTMP | TRADE_SIDE_CD | TRADE_CURRENCY_CD | TRADE_QTY | GROSS_USD_AMT |
|---|---|---|---|---|---|---|---|
| 12345 | XNGS | 8-Oct-14 | 08-OCT-14 10.38.46 AM | SEL | USD | 100 | 7610.75 |
| 12345 | XNGS | 8-Oct-14 | 08-OCT-14 03.49.00 PM | BUY | USD | 700 | 54279.26 |
| 12345 | XNGS | 8-Oct-14 | 08-OCT-14 03.48.59 PM | BUY | USD | 100 | 7699.28 |
| 12345 | XNGS | 8-Oct-14 | 08-OCT-14 03.48.59 PM | BUY | USD | 100 | 7698.4 |