Capstone Project

Financial News Analytics

Trend Analysis of Market Securities based on Sentiment Analysis of News and Social Media

Submitted by:

Name	Student Id
Abhishek Khanna	71510003
Sonu Krishnan	71510076
Sumit Janmejai	71510081
Venkata Pilla	71510088

Acknowledgements

We would like to take this opportunity to express our profound gratitude and deep regards to our project sponsor, Deloitte Consulting, for providing the opportunity to learn and work on challenging Financial News Analytics – Sentiment Analysis and also to our project mentor Reema, for her efforts in facilitating this Capstone Opportunity. We are deeply indebted to all of them and welcome this opportunity to benefit further from their contribution. We wish to express our special thanks to Hemanth from ISB, Suresh Kumar, Chandra Narra and Sandeep Kumar Sharma from Deloitte and our other batch mates who have provided their valuable suggestions.

- Abhishek, Sonu, Sumit, Venkat

Contents

Introduction

News from various parts of the world has a strong relationship with the market (investor) sentiments as markets react sharply and promptly to the news and the sentiment they carry. Effective models that incorporate news data are required by all investors (be it Fund Houses, Brokers or Investment Bank Managers). The idea behind News Analytics is to automate human thinking and reasoning to come up with a quantifiable model and enable users with the decisive capability to predict gains or control risk.

The Project Objectives are:

(i) To Perform Trend Analysis of securities based on Market Sentiments through Sentiment Analysis on Financial News and Social Media updates

(ii) To come up with a Sentiment score or Sentiment polarity for each security, sector or benchmark by analyzing Online Financial and Generic news, Market data and Social Media updates.

This project aims to use Statistical and Analytical tools in Text Analytics on the sourced data to perform the following:

1. Sentiment Score and Mapping for each security/sector/benchmark
2. Overall market trends related to news
3. Visualizations and dashboards complementing the above

Areas of Impact

- The portfolio specific investment advice given by Investment Management firms based on an overall sentiment from a model that captures the sentiment from financial news and social media content provides much more qualitative insights to potential investors in the markets.

- A study of the variation of security prices corresponding to the variation of sentiment polarity for securities in each sector and markets would provide good insights to Investment Management firms and Fund Broker houses.

Project Approach

The approach taken is defined below:

- Identification of a Specific Market/Region and Securities to focus on.
- Identification of various Financial and General News websites, Social Media sites to source the information from.

- ➢ Assigning weightage to each source of data based on the Reliability factor.
- ➢ Design and develop scripts in "R" to collect the data from the identified sources (data scraping), pre-process (data cleaning) and making it ready for analysis.
- ➢ Perform various Text Analytics on the data using NLP techniques. Define a data dictionary and generation of word cloud, bi-grams and tri-grams for analysis.
- ➢ Arriving at a final score ranging from -1 to +1 for each Security that depicts the negative or the positive online sentiment based on the information.
- ➢ Perform the above analysis for Securities as well as various sectors identified.
- ➢ Support various visualizations using R and Tableau.

Project Workflow

Phase 1: Exploratory Analysis and Data Collection

- ✓ Identification of a specific market/region and securities to focus on.
- ✓ Identification of various Financial and General News websites, Social Media sites to source the information from.

Phase 2: Data preparation

- ✓ Designing of various scripts in R to collect the data from the identified sources (data scraping), pre-process (data cleaning) and making it ready for analysis

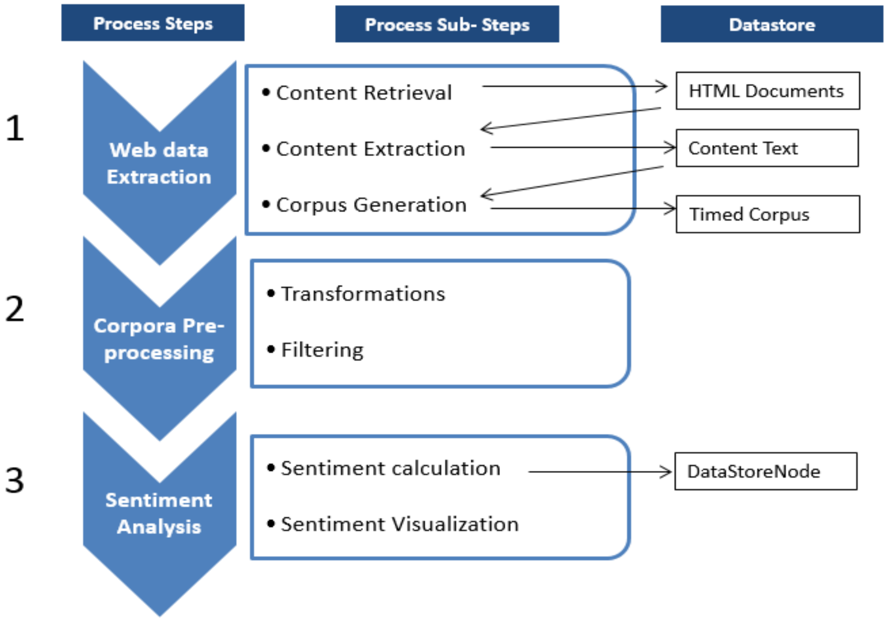
Phase 3: Model building and validation

- ✓ Perform various text analytics on that using NLP techniques. Defining a data dictionary and generation of word-cloud, bi-grams and tri-grams for analysis.
- ✓ Arriving at a final score ranging from -1 to +1 for each security that depicts the negative or the positive online sentiment about a particular security

Phase 4: Writing the report and summarizing the findings

- ✓ Development of dashboards and visualizations in Tableau to see the variation in market securities against the sentiment polarities
- ✓ Observations, visual analysis and project report documentation

Workflow



Project Phases

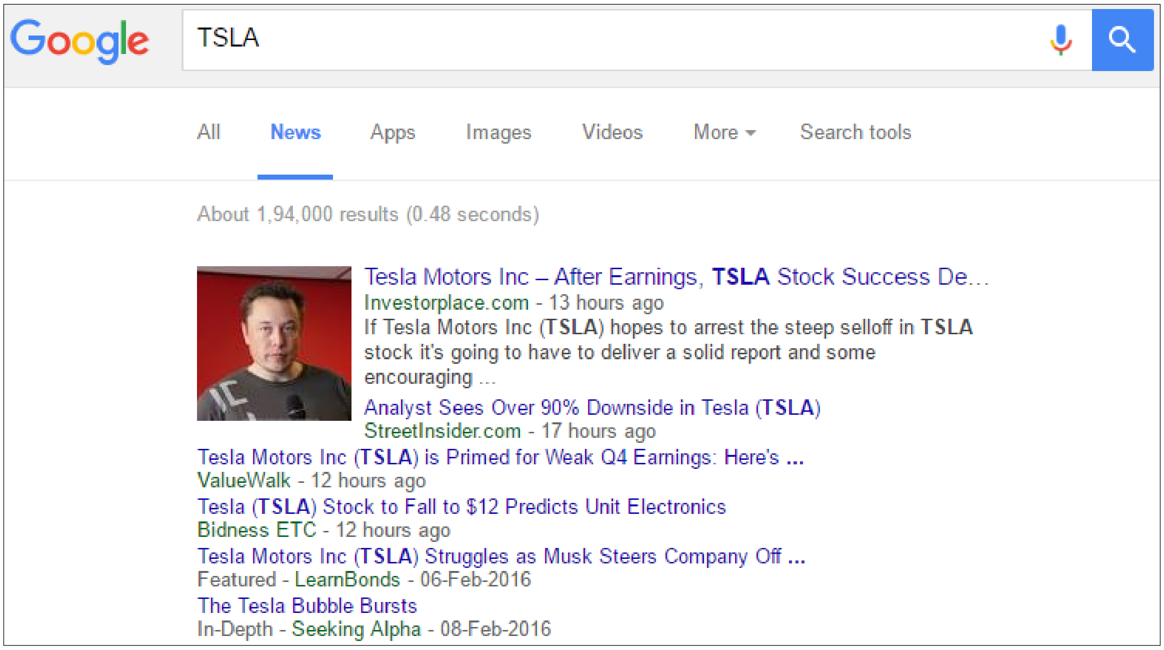
Phase I: Exploratory Analysis and Data Collection

Description of Data Collected:

- The Market and Sector identified for the Project Scope is the US Markets and Automobile sector respectively.
- The Securities identified are :
    - Ford Motor Company (Stock code: F)
    - Tesla Motors Inc. (Stock code: TSLA)
    - General Motors Company (Stock code: GM)
- Data sources
    - **Online news:**

      Google News is identified as News source as it is a News aggregator and has a collection of News content from various websites. Also this is being widely used by Investor community for Market and Industry updates.

Sample Google News for Tesla (TSLA)



About 1,94,000 results (0.48 seconds)

Tesla Motors Inc – After Earnings, TSLA Stock Success De...
Investorplace.com - 13 hours ago
If Tesla Motors Inc (TSLA) hopes to arrest the steep selloff in TSLA stock it's going to have to deliver a solid report and some encouraging ...

Analyst Sees Over 90% Downside in Tesla (TSLA)
StreetInsider.com - 17 hours ago

Tesla Motors Inc (TSLA) is Primed for Weak Q4 Earnings: Here's ...
ValueWalk - 12 hours ago

Tesla (TSLA) Stock to Fall to $12 Predicts Unit Electronics
Bidness ETC - 12 hours ago

Tesla Motors Inc (TSLA) Struggles as Musk Steers Company Off ...
Featured - LearnBonds - 06-Feb-2016

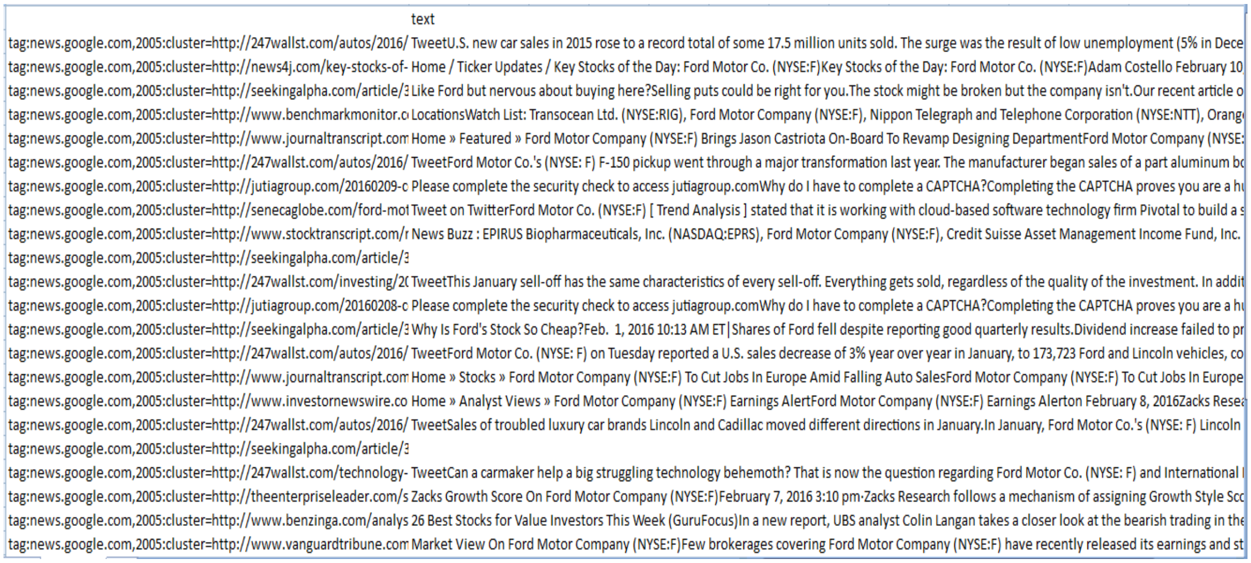
The Tesla Bubble Bursts
In-Depth - Seeking Alpha - 08-Feb-2016

From the above news, we can clearly see that there are both positive and negative news for a particular day where Business forums are speculating the Q4 performance and thereby the Stock Price.

URL ->

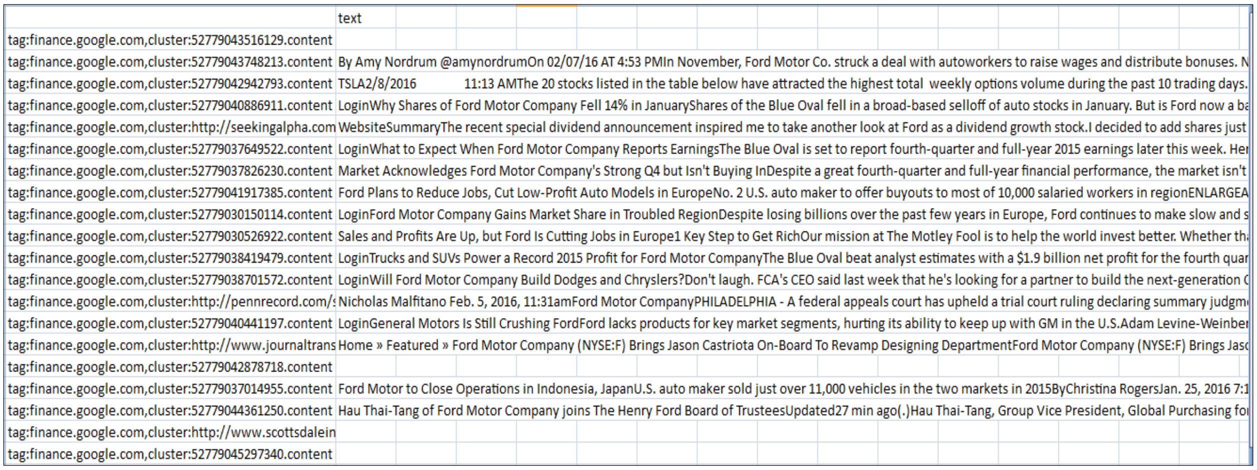
https://www.google.co.in/search?hl=en&gl=in&tbm=nws&authuser=0&q=TSLA&oq=TSLA&gs_l=news-cc.3..43j0l2j43i53.2556.4495.0.5169.4.4.0.0.0.0.420.824.2j0j1j0j1.4.0...0.0...1ac.1.gSblk98vZUA

Sample Google News Extract from Internet for Ford Motors



	text
tag:news.google.com,2005:cluster=http://247wallst.com/autos/2016/	TweetU.S. new car sales in 2015 rose to a record total of some 17.5 million units sold. The surge was the result of low unemployment (5% in Dece
tag:news.google.com,2005:cluster=http://news4j.com/key-stocks-of-	Home / Ticker Updates / Key Stocks of the Day: Ford Motor Co. (NYSE:F)Key Stocks of the Day: Ford Motor Co. (NYSE:F)Adam Costello February 10
tag:news.google.com,2005:cluster=http://seekingalpha.com/article/3	Like Ford but nervous about buying here?Selling puts could be right for you.The stock might be broken but the company isn't.Our recent article o
tag:news.google.com,2005:cluster=http://www.benchmarkmonitor.c	LocationsWatch List: Transocean Ltd. (NYSE:RIG), Ford Motor Company (NYSE:F), Nippon Telegraph and Telephone Corporation (NYSE:NTT), Orang
tag:news.google.com,2005:cluster=http://www.journaltranscript.com	Home » Featured » Ford Motor Company (NYSE:F) Brings Jason Castriota On-Board To Revamp Designing DepartmentFord Motor Company (NYSE:
tag:news.google.com,2005:cluster=http://247wallst.com/autos/2016/	TweetFord Motor Co.'s (NYSE: F) F-150 pickup went through a major transformation last year. The manufacturer began sales of a part aluminum bc
tag:news.google.com,2005:cluster=http://jutiagroup.com/20160209-c	Please complete the security check to access jutiagroup.comWhy do I have to complete a CAPTCHA?Completing the CAPTCHA proves you are a hu
tag:news.google.com,2005:cluster=http://senecaglobe.com/ford-mot	Tweet on TwitterFord Motor Co. (NYSE:F) [Trend Analysis] stated that it is working with cloud-based software technology firm Pivotal to build a s
tag:news.google.com,2005:cluster=http://www.stocktranscript.com/r	News Buzz : EPIRUS Biopharmaceuticals, Inc. (NASDAQ:EPRS), Ford Motor Company (NYSE:F), Credit Suisse Asset Management Income Fund, Inc.
tag:news.google.com,2005:cluster=http://seekingalpha.com/article/3	
tag:news.google.com,2005:cluster=http://247wallst.com/investing/2(	TweetThis January sell-off has the same characteristics of every sell-off. Everything gets sold, regardless of the quality of the investment. In addit
tag:news.google.com,2005:cluster=http://jutiagroup.com/20160208-c	Please complete the security check to access jutiagroup.comWhy do I have to complete a CAPTCHA?Completing the CAPTCHA proves you are a hu
tag:news.google.com,2005:cluster=http://seekingalpha.com/article/3	Why Is Ford's Stock So Cheap?Feb. 1, 2016 10:13 AM ET|Shares of Ford fell despite reporting good quarterly results.Dividend increase failed to pr
tag:news.google.com,2005:cluster=http://247wallst.com/autos/2016/	TweetFord Motor Co. (NYSE: F) on Tuesday reported a U.S. sales decrease of 3% year over year in January, to 173,723 Ford and Lincoln vehicles, co
tag:news.google.com,2005:cluster=http://www.journaltranscript.com	Home » Stocks » Ford Motor Company (NYSE:F) To Cut Jobs In Europe Amid Falling Auto SalesFord Motor Company (NYSE:F) To Cut Jobs In Europe
tag:news.google.com,2005:cluster=http://www.investornewswire.co	Home » Analyst Views » Ford Motor Company (NYSE:F) Earnings AlertFord Motor Company (NYSE:F) Earnings Alerton February 8, 2016Zacks Resea
tag:news.google.com,2005:cluster=http://247wallst.com/autos/2016/	TweetSales of troubled luxury car brands Lincoln and Cadillac moved different directions in January.In January, Ford Motor Co.'s (NYSE: F) Lincoln
tag:news.google.com,2005:cluster=http://seekingalpha.com/article/3	
tag:news.google.com,2005:cluster=http://247wallst.com/technology-	TweetCan a carmaker help a big struggling technology behemoth? That is now the question regarding Ford Motor Co. (NYSE: F) and International I
tag:news.google.com,2005:cluster=http://theenterpriseleader.com/s	Zacks Growth Score On Ford Motor Company (NYSE:F)February 7, 2016 3:10 pm·Zacks Research follows a mechanism of assigning Growth Style Scc
tag:news.google.com,2005:cluster=http://www.benzinga.com/analys	26 Best Stocks for Value Investors This Week (GuruFocus)In a new report, UBS analyst Colin Langan takes a closer look at the bearish trading in the
tag:news.google.com,2005:cluster=http://www.vanguardtribune.com	Market View On Ford Motor Company (NYSE:F)Few brokerages covering Ford Motor Company (NYSE:F) have recently released its earnings and st

<u>Sample Google Finance Extract from Internet for Ford Motors</u>



	text
tag:finance.google.com,cluster:52779043516129.content	
tag:finance.google.com,cluster:52779043748213.content	By Amy Nordrum @amynordrumOn 02/07/16 AT 4:53 PMIn November, Ford Motor Co. struck a deal with autoworkers to raise wages and distribute bonuses. N
tag:finance.google.com,cluster:52779042942793.content	TSLA2/8/2016 11:13 AMThe 20 stocks listed in the table below have attracted the highest total weekly options volume during the past 10 trading days.
tag:finance.google.com,cluster:52779040886911.content	LoginWhy Shares of Ford Motor Company Fell 14% in JanuaryShares of the Blue Oval fell in a broad-based selloff of auto stocks in January. But is Ford now a ba
tag:finance.google.com,cluster:http://seekingalpha.com	WebsiteSummaryThe recent special dividend announcement inspired me to take another look at Ford as a dividend growth stock.I decided to add shares just
tag:finance.google.com,cluster:52779037649522.content	LoginWhat to Expect When Ford Motor Company Reports EarningsThe Blue Oval is set to report fourth-quarter and full-year 2015 earnings later this week. Her
tag:finance.google.com,cluster:52779037826230.content	Market Acknowledges Ford Motor Company's Strong Q4 but Isn't Buying InDespite a great fourth-quarter and full-year financial performance, the market isn't
tag:finance.google.com,cluster:52779041917385.content	Ford Plans to Reduce Jobs, Cut Low-Profit Auto Models in EuropeNo. 2 U.S. auto maker to offer buyouts to most of 10,000 salaried workers in regionENLARGEA
tag:finance.google.com,cluster:52779030150114.content	LoginFord Motor Company Gains Market Share in Troubled RegionDespite losing billions over the past few years in Europe, Ford continues to make slow and s
tag:finance.google.com,cluster:52779030526922.content	Sales and Profits Are Up, but Ford Is Cutting Jobs in Europe1 Key Step to Get RichOur mission at The Motley Fool is to help the world invest better. Whether th
tag:finance.google.com,cluster:52779038419479.content	LoginTrucks and SUVs Power a Record 2015 Profit for Ford Motor CompanyThe Blue Oval beat analyst estimates with a $1.9 billion net profit for the fourth quar
tag:finance.google.com,cluster:52779038701572.content	LoginWill Ford Motor Company Build Dodges and Chryslers?Don't laugh. FCA's CEO said last week that he's looking for a partner to build the next-generation C
tag:finance.google.com,cluster:http://pennrecord.com/s	Nicholas Malfitano Feb. 5, 2016, 11:31amFord Motor CompanyPHILADELPHIA - A federal appeals court has upheld a trial court ruling declaring summary judgm
tag:finance.google.com,cluster:52779040441197.content	LoginGeneral Motors Is Still Crushing FordFord lacks products for key market segments, hurting its ability to keep up with GM in the U.S.Adam Levine-Weinber
tag:finance.google.com,cluster:http://www.journaltrans	Home » Featured » Ford Motor Company (NYSE:F) Brings Jason Castriota On-Board To Revamp Designing DepartmentFord Motor Company (NYSE:F) Brings Jasc
tag:finance.google.com,cluster:52779042878718.content	
tag:finance.google.com,cluster:52779037014955.content	Ford Motor to Close Operations in Indonesia, JapanU.S. auto maker sold just over 11,000 vehicles in the two markets in 2015ByChristina RogersJan. 25, 2016 7:1
tag:finance.google.com,cluster:52779044361250.content	Hau Thai-Tang of Ford Motor Company joins The Henry Ford Board of TrusteesUpdated27 min ago(.)Hau Thai-Tang, Group Vice President, Global Purchasing fo
tag:finance.google.com,cluster:http://www.scottsdalein	
tag:finance.google.com,cluster:52779045297340.content	

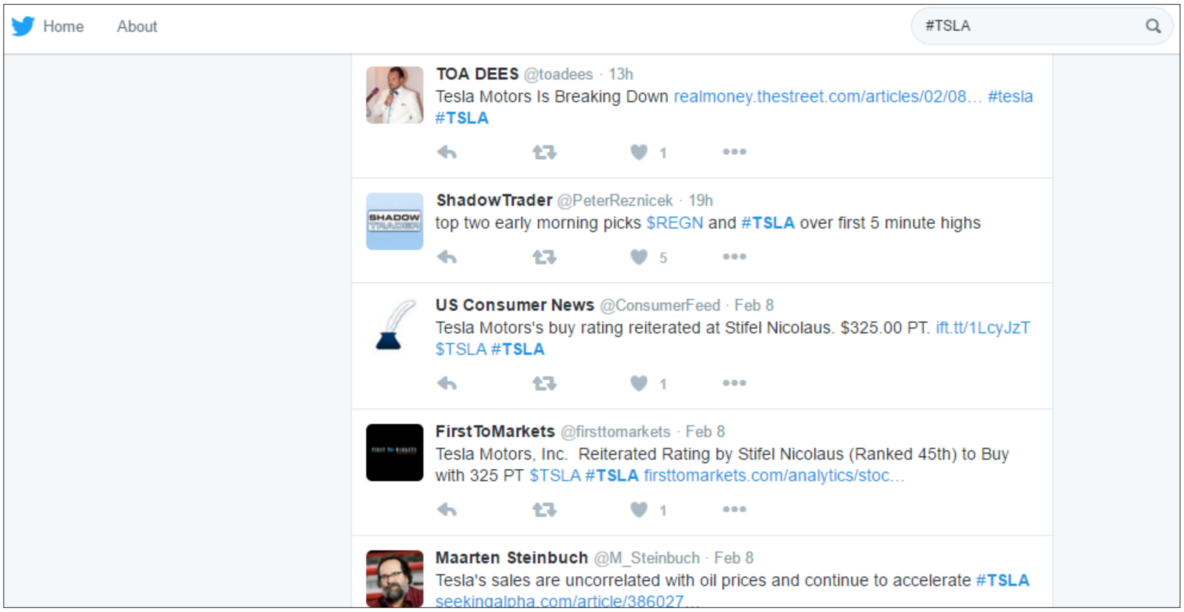
- **Social Media:**

  We sourced data from **Twitter** based on the securities' company names and stock codes both. **Challenges:** The Twitter API used allowed us to fetch tweets only up to one week in the past. Hence to overcome this limitation we set up a cluster and engine to run the API once every week. We were able to have tweets collected for the past several weeks.

  <u>Sample Twitter Output</u>

  Similar to the Google News, the below snapshot of the tweets from Twitter website for #TSLA also clearly shows Positive as well as Negative sentiment tweets. There are tweets reinstating the "Buy" status on the stock while there is a tweet stating "Tesla Motors is breaking down".

  URL -> https://twitter.com/search?q=%23TSLA&src=typd



TOA DEES @toadees · 13h
Tesla Motors Is Breaking Down realmoney.thestreet.com/articles/02/08… #tesla #TSLA
1

ShadowTrader @PeterReznicek · 19h
top two early morning picks $REGN and #TSLA over first 5 minute highs
5

US Consumer News @ConsumerFeed · Feb 8
Tesla Motors's buy rating reiterated at Stifel Nicolaus. $325.00 PT. ift.tt/1LcyJzT $TSLA #TSLA
1

FirstToMarkets @firsttomarkets · Feb 8
Tesla Motors, Inc. Reiterated Rating by Stifel Nicolaus (Ranked 45th) to Buy with 325 PT $TSLA #TSLA firsttomarkets.com/analytics/stoc…
1

Maarten Steinbuch @M_Steinbuch · Feb 8
Tesla's sales are uncorrelated with oil prices and continue to accelerate #TSLA seekingalpha.com/article/386027…

- **Web Data extraction:**

  This is done through RSS Feeds and APIs.

  - **RSS feeds :**

    News Sources such as Google News, Google Finance, Yahoo News and Yahoo Finance provide RSS (**R**ich **S**ite **S**ummary; originally RDF Site Summary; often called **R**eally **S**imple **S**yndication) feed options. The feeds provide Metadata of News content about any topic i.e. the URLs of external sites where content resides.

    This is a 2-step procedure involving:

    (i) Downloading metadata feeds

    (ii) Downloading content sites

  - **APIs:**

    News Sources: Reuters Spotlight, NY Times, Yahoo BOSS, Bing

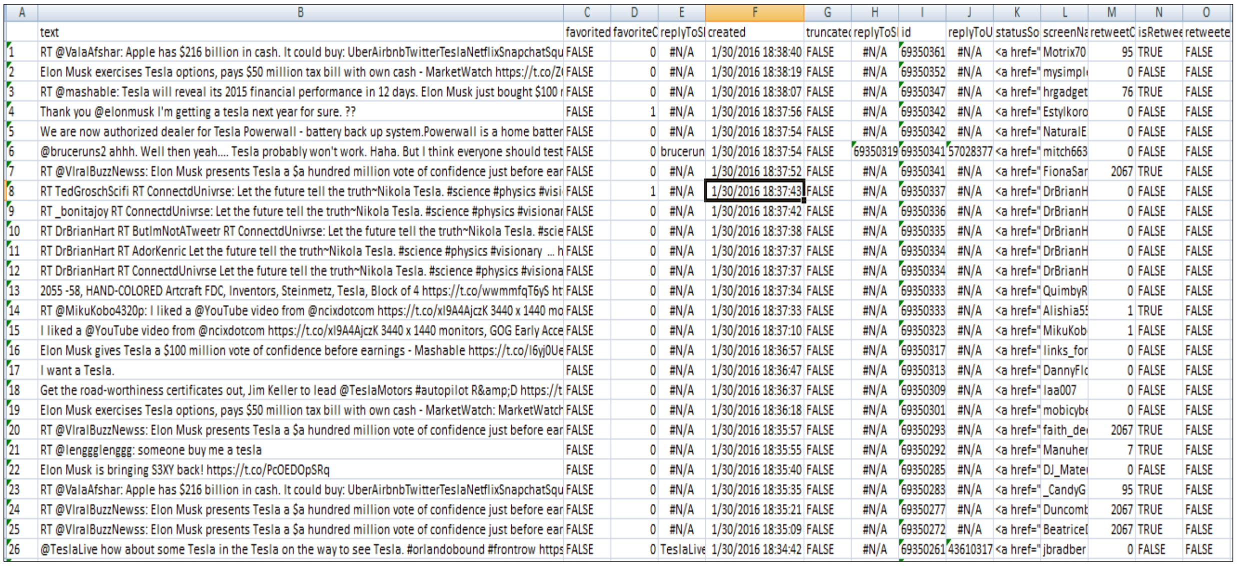
    Twitter: Publically available REST API. Twitter API poses the following limitations:

    (i) Number of tweets that can be fetched – rate limit

    (ii) How far back can we go while collecting data ~ 7 days

    Hence we need to set up a periodic run for collecting Twitter data on a weekly basis. More information on REST API can be found https://dev.twitter.com/rest/public

    Sample of the Twitter data scraped from the internet using the API is shown below,



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	text	favorited	favoriteC	replyToS	created	truncate	replyToS	id	replyToU	statusSo	screenNa	retweetC	isRetwee	retweete
1	RT @ValaAfshar: Apple has $216 billion in cash. It could buy: UberAirbnbTwitterTeslaNetflixSnapchatSqu	FALSE	0	#N/A	1/30/2016 18:38:40	FALSE	#N/A	69350361	#N/A	<a href="	Motrix70	95	TRUE	FALSE
2	Elon Musk exercises Tesla options, pays $50 million tax bill with own cash - MarketWatch https://t.co/Z	FALSE	0	#N/A	1/30/2016 18:38:19	FALSE	#N/A	69350352	#N/A	<a href="	mysimpl	0	FALSE	FALSE
3	RT @mashable: Tesla will reveal its 2015 financial performance in 12 days. Elon Musk just bought $100 r	FALSE	0	#N/A	1/30/2016 18:38:07	FALSE	#N/A	69350347	#N/A	<a href="	hrgadget	76	TRUE	FALSE
4	Thank you @elonmusk I'm getting a tesla next year for sure. ??	FALSE	1	#N/A	1/30/2016 18:37:56	FALSE	#N/A	69350342	#N/A	<a href="	Estylkoro	0	FALSE	FALSE
5	We are now authorized dealer for Tesla Powerwall - battery back up system.Powerwall is a home batter	FALSE	0	#N/A	1/30/2016 18:37:54	FALSE	#N/A	69350342	#N/A	<a href="	NaturalE	0	FALSE	FALSE
6	@bruceruns2 ahhh. Well then yeah.... Tesla probably won't work. Haha. But I think everyone should test	FALSE	0	brucerun	1/30/2016 18:37:54	FALSE	69350319	69350341	57028377	<a href="	mitch663	0	FALSE	FALSE
7	RT @VIralBuzzNewss: Elon Musk presents Tesla a $a hundred million vote of confidence just before ear	FALSE	0	#N/A	1/30/2016 18:37:52	FALSE	#N/A	69350341	#N/A	<a href="	FionaSar	2067	TRUE	FALSE
8	RT TedGroschScifi RT ConnectdUnivrse: Let the future tell the truth~Nikola Tesla. #science #physics #visi	FALSE	1	#N/A	1/30/2016 18:37:43	FALSE	#N/A	69350337	#N/A	<a href="	DrBrianH	0	FALSE	FALSE
9	RT _bonitajoy RT ConnectdUnivrse: Let the future tell the truth~Nikola Tesla. #science #physics #visionar	FALSE	0	#N/A	1/30/2016 18:37:42	FALSE	#N/A	69350336	#N/A	<a href="	DrBrianH	0	FALSE	FALSE
10	RT DrBrianHart RT ButImNotATweetr RT ConnectdUnivrse: Let the future tell the truth~Nikola Tesla. #scie	FALSE	0	#N/A	1/30/2016 18:37:38	FALSE	#N/A	69350335	#N/A	<a href="	DrBrianH	0	FALSE	FALSE
11	RT DrBrianHart RT AdorKenric Let the future tell the truth~Nikola Tesla. #science #physics #visionary … h	FALSE	0	#N/A	1/30/2016 18:37:37	FALSE	#N/A	69350334	#N/A	<a href="	DrBrianH	0	FALSE	FALSE
12	RT DrBrianHart RT ConnectdUnivrse Let the future tell the truth~Nikola Tesla. #science #physics #visiona	FALSE	0	#N/A	1/30/2016 18:37:37	FALSE	#N/A	69350334	#N/A	<a href="	DrBrianH	0	FALSE	FALSE
13	2055 -58, HAND-COLORED Artcraft FDC, Inventors, Steinmetz, Tesla, Block of 4 https://t.co/wwmmfqT6yS ht	FALSE	0	#N/A	1/30/2016 18:37:34	FALSE	#N/A	69350333	#N/A	<a href="	QuimbyR	0	FALSE	FALSE
14	RT @MikuKobo4320p: I liked a @YouTube video from @ncixdotcom https://t.co/xI9A4AjczK 3440 x 1440 mo	FALSE	0	#N/A	1/30/2016 18:37:33	FALSE	#N/A	69350333	#N/A	<a href="	Alishia5	1	TRUE	FALSE
15	I liked a @YouTube video from @ncixdotcom https://t.co/xI9A4AjczK 3440 x 1440 monitors, GOG Early Acce	FALSE	0	#N/A	1/30/2016 18:37:10	FALSE	#N/A	69350323	#N/A	<a href="	MikuKob	1	FALSE	FALSE
16	Elon Musk gives Tesla a $100 million vote of confidence before earnings - Mashable https://t.co/I6yj0Ue	FALSE	0	#N/A	1/30/2016 18:36:57	FALSE	#N/A	69350317	#N/A	<a href="	links_for	0	FALSE	FALSE
17	I want a Tesla.	FALSE	0	#N/A	1/30/2016 18:36:47	FALSE	#N/A	69350313	#N/A	<a href="	DannyFl	0	FALSE	FALSE
18	Get the road-worthiness certificates out, Jim Keller to lead @TeslaMotors #autopilot R&amp;D https://t	FALSE	0	#N/A	1/30/2016 18:36:37	FALSE	#N/A	69350309	#N/A	<a href="	laa007	0	FALSE	FALSE
19	Elon Musk exercises Tesla options, pays $50 million tax bill with own cash - MarketWatch: MarketWatch	FALSE	0	#N/A	1/30/2016 18:36:18	FALSE	#N/A	69350301	#N/A	<a href="	mobicyb	0	FALSE	FALSE
20	RT @VIralBuzzNewss: Elon Musk presents Tesla a $a hundred million vote of confidence just before ear	FALSE	0	#N/A	1/30/2016 18:35:57	FALSE	#N/A	69350293	#N/A	<a href="	faith_de	2067	TRUE	FALSE
21	RT @lenggglenggg: someone buy me a tesla	FALSE	0	#N/A	1/30/2016 18:35:55	FALSE	#N/A	69350292	#N/A	<a href="	Manuher	7	TRUE	FALSE
22	Elon Musk is bringing S3XY back! https://t.co/PcOEDOpSRq	FALSE	0	#N/A	1/30/2016 18:35:40	FALSE	#N/A	69350285	#N/A	<a href="	DJ_Mate	0	FALSE	FALSE
23	RT @ValaAfshar: Apple has $216 billion in cash. It could buy: UberAirbnbTwitterTeslaNetflixSnapchatSqu	FALSE	0	#N/A	1/30/2016 18:35:35	FALSE	#N/A	69350283	#N/A	<a href="	_CandyG	95	TRUE	FALSE
24	RT @VIralBuzzNewss: Elon Musk presents Tesla a $a hundred million vote of confidence just before ear	FALSE	0	#N/A	1/30/2016 18:35:21	FALSE	#N/A	69350277	#N/A	<a href="	Duncomb	2067	TRUE	FALSE
25	RT @VIralBuzzNewss: Elon Musk presents Tesla a $a hundred million vote of confidence just before ear	FALSE	0	#N/A	1/30/2016 18:35:09	FALSE	#N/A	69350272	#N/A	<a href="	Beatrice	2067	TRUE	FALSE
26	@TeslaLive how about some Tesla in the Tesla on the way to see Tesla. #orlandobound #frontrow https	FALSE	0	TeslaLive	1/30/2016 18:34:42	FALSE	#N/A	69350261	43610317	<a href="	jbradber	0	FALSE	FALSE

Code attached in Annexure 1.

Phase II: Data Preparation

Methods employed:

1) Content extraction from RSS Feeds

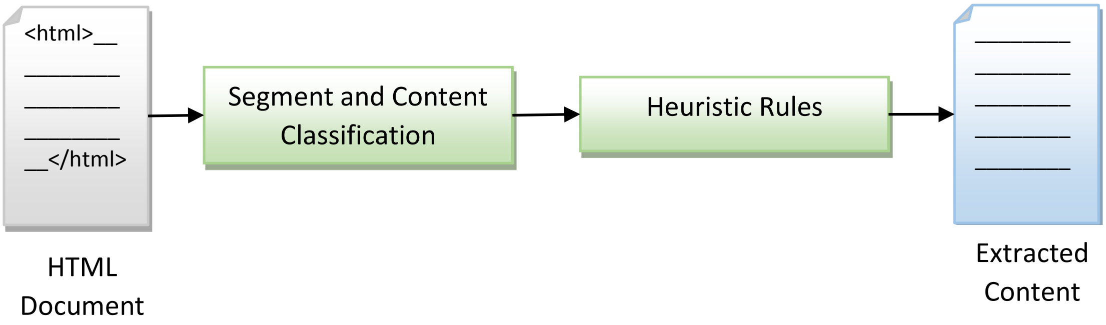
- Blogs and News Sites often contain side bars, suggested reads, Ads and other unwanted contents in addition to the required News content
- To extract the key content from the Web Page and ignore the noise, the following technique has been used.

  Reference:

  A detailed survey of HTML content extraction is given by [Gottron, 2008]:

  "Content Extraction: Identifying the Main Content in HTML Documents"

- The Technique:



As shown above, the combination of Largest Block of String (LBS) and classification tasks namely CS+LBS yield the best performance for Blogs and News datasets. The following steps are required:

- Examine each HTML sub node from top to bottom
- If text length/total length < threshold then drill down
- Select text from sub node with longest text length

2) Twitter REST API

- Twitter is one of the popular Social Media platforms where user can communicate with wider audience instantly. Additionally this platform is popular in terms of the Media attention it receives and hence makes it easier to find and follow conversations.
- Twitter has hashtag (#) norms which make it easier gathering, sorting and expanding searches. When collecting information related to major incidents, news stories, all such events tend to be centered on hashtag.
- However there is a limitation on volume and timeline of data that can be captured using the Twitter API.

Limitations:

- Twitter allows capture of only 300-400 tweets every 15 minutes per handle.
- Allows to capture data up to 7 days in past.

Our Approach:

1. Shortlisted the appropriate hashtags:

   I. FORD: #fordmotors #ford

   II. TESLA: #TeslaMotors #Tesla #TSLA

   III. General Motors_: #GeneralMotors #GM

2. Executed the code every weekend to capture the conversations on the above hashtags and dumped them in an excel file.

Phase III: Model Building and Validation

Pre-Processing:

Some cleaning and preparation of corpus required before sentiment classification operations.

Standard transformations like:

- Conversion to lower case
- Removal of stop-words
- Removal of punctuations
- Stemming
- Stripping of white space
- And more depending on quality of relevant text extraction

“tm” package in R has easy functions to achieve all the above.

Sentiment Analysis:

Sentiment Calculation

Bag of Words model is used for sentiment calculation. Advantages of such an approach are:

1. No need of labelled training data
2. Quick turn-around

Pre requisites for the sentiment calculation are:

1. Document Term Matrix (tm package in R)
2. Lexicons (dictionary of positive/negative terms)

SentiWordnet, General Inquirer from Harvard University, NTU Sentiment Dictionary, OpinionFinder’s Subjectivity Lexicon are some of the freely available and widely regarded lexicons.

Following indicators of sentiment are taken directly from the Lydia Sentiment Analysis framework as explained by Wenbin Zhang and Steven Skiena in *“Trading Strategies To Exploit News Sentiment” :*

“The Lydia sentiment data consists of time series of favorable (positive) and unfavorable (negative)

*words co-referenced with occurrences of each named entity (here denoting companies). Let **p** and **n** denote the number of raw positive and negative references to a given entity, which occurs a total of N times in the corpus (including neutral references).*
Then we derive the following natural sentiment/subjectivity measures from these raw counts:

- polarity = (p − n)/(p + n)
- subjectivity = (n + p)/N
- pos refs per ref = p/N
- neg refs per ref = n/N
- senti diffs per ref = (p − n)/N

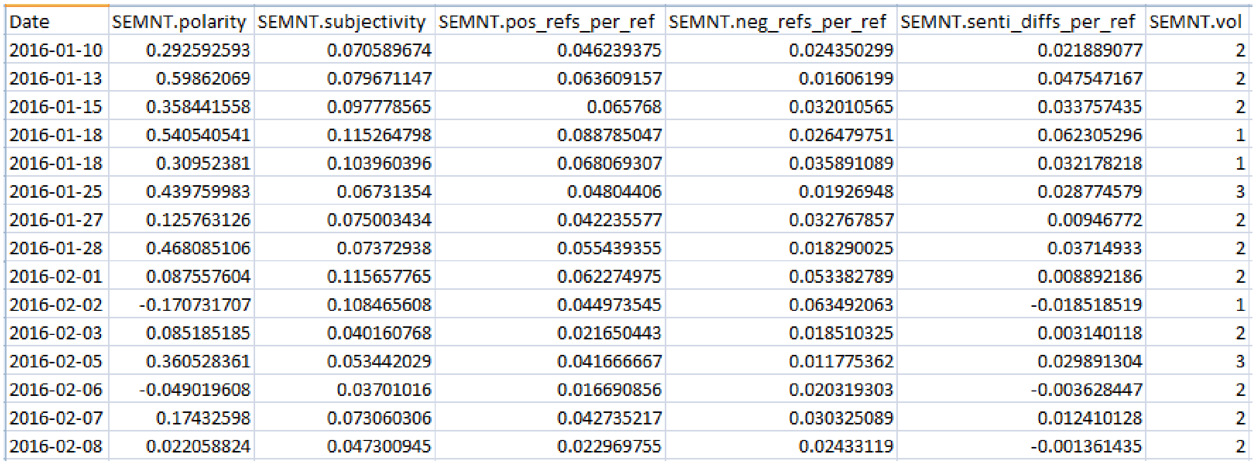
"These measures are not highly correlated with raw sentiment counts and they can provide additional information that raw data cannot. Therefore, with them we will be able to avoid multicollinearity during linear analysis"

Reference: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.8687&rep=rep1&type=pdf

Model Output

The output of the model would be a file that lists the daily values of net polarity and subjectivity values for each of the securities identified.

A snapshot of the output file is shown below:



Date	SEMNT.polarity	SEMNT.subjectivity	SEMNT.pos_refs_per_ref	SEMNT.neg_refs_per_ref	SEMNT.senti_diffs_per_ref	SEMNT.vol
2016-01-10	0.292592593	0.070589674	0.046239375	0.024350299	0.021889077	2
2016-01-13	0.59862069	0.079671147	0.063609157	0.01606199	0.047547167	2
2016-01-15	0.358441558	0.097778565	0.065768	0.032010565	0.033757435	2
2016-01-18	0.540540541	0.115264798	0.088785047	0.026479751	0.062305296	1
2016-01-18	0.30952381	0.103960396	0.068069307	0.035891089	0.032178218	1
2016-01-25	0.439759983	0.06731354	0.04804406	0.01926948	0.028774579	3
2016-01-27	0.125763126	0.075003434	0.042235577	0.032767857	0.00946772	2
2016-01-28	0.468085106	0.07372938	0.055439355	0.018290025	0.03714933	2
2016-02-01	0.087557604	0.115657765	0.062274975	0.053382789	0.008892186	2
2016-02-02	-0.170731707	0.108465608	0.044973545	0.063492063	-0.018518519	1
2016-02-03	0.085185185	0.040160768	0.021650443	0.018510325	0.003140118	2
2016-02-05	0.360528361	0.053442029	0.041666667	0.011775362	0.029891304	3
2016-02-06	-0.049019608	0.03701016	0.016690856	0.020319303	-0.003628447	2
2016-02-07	0.17432598	0.073060306	0.042735217	0.030325089	0.012410128	2
2016-02-08	0.022058824	0.047300945	0.022969755	0.02433119	-0.001361435	2

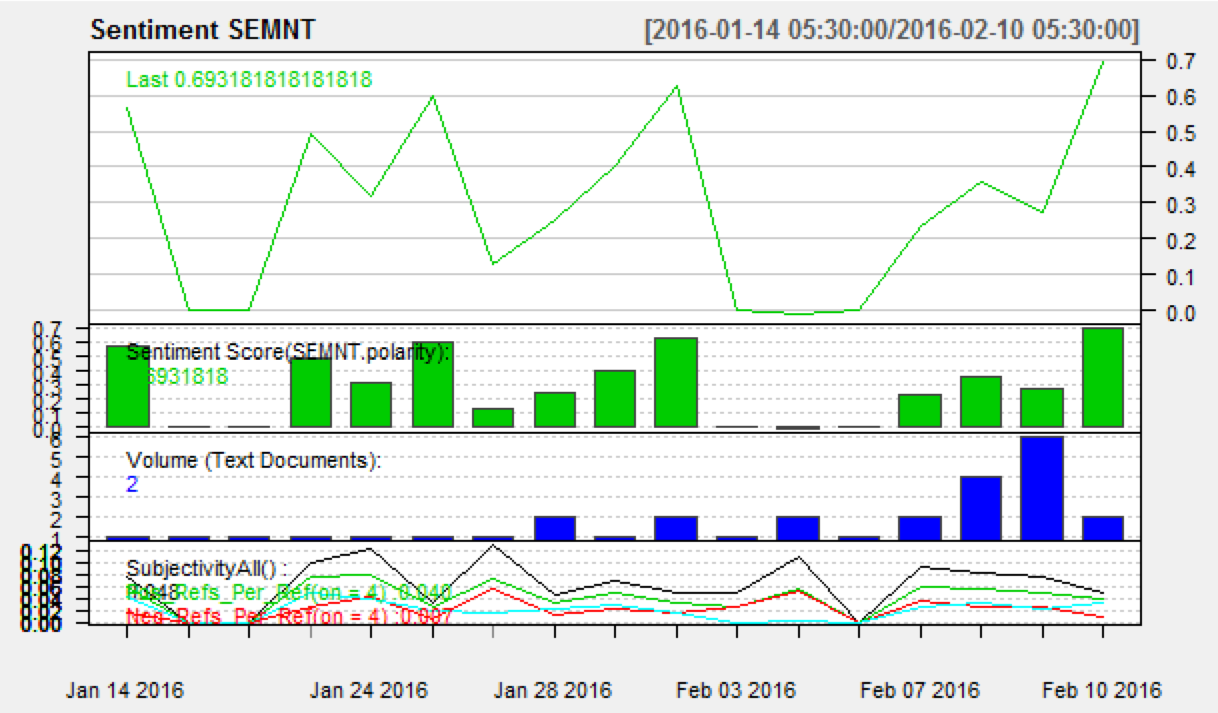
Phase IV: Visualization and Dashboards

The following visualizations have been derived from the model output and the values of the EOD security prices on respective days:

1. Daily volume trend of number of Tweets and number of web extraction for each of the securities
2. Time series trends visual on aggregated Sentiment Polarity for each of the securities
3. Time series trends visual on Sentiment Subjectivity for each security
4. A time series analysis of the variation of actual security prices on the markets with respect to the positive and negative sentiments for each of the securities
5. A time series analysis of the variation of actual security prices on the markets with respect to the aggregated sentiment polarity for each of the securities
6. A time series analysis of the variation of actual security prices against the aggregated sentiment polarity of the (T-1) day for each of the securities

The Model output graphs show the Sentiment, Polarity and Volume in the transaction. Below are the sample for the Google News and Google Finance data.

Google News Data Analysis



Google Finance Data Analysis

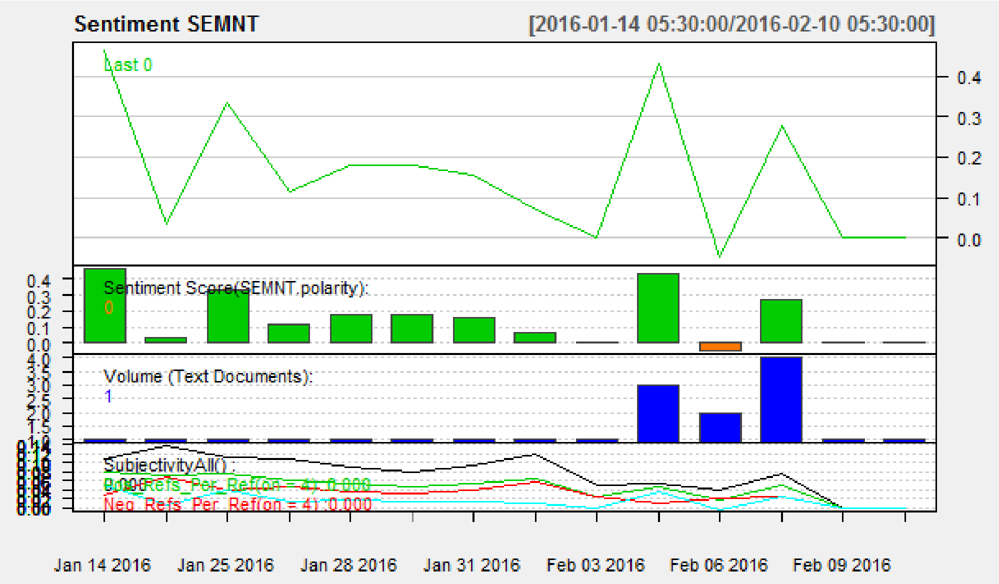
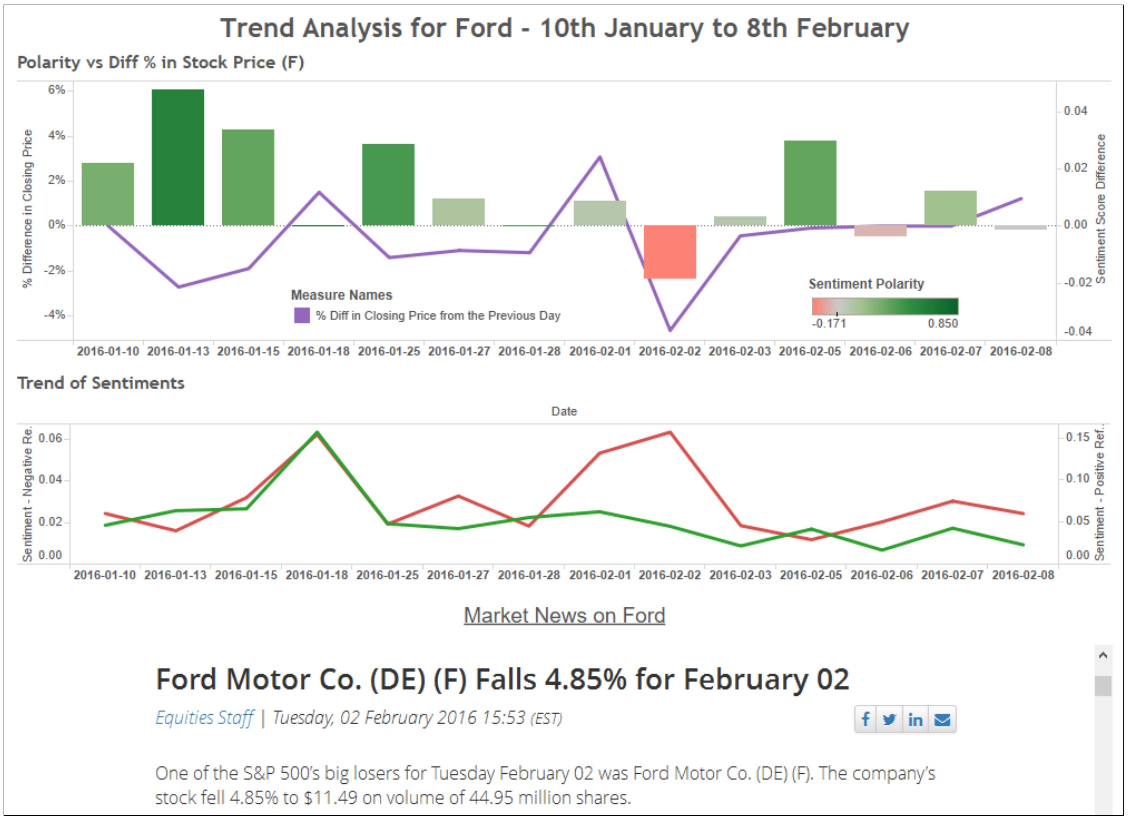


Tableau Dashboard showing Trend Analysis – Polarity vs % change in Stock Price for Ford

Below is dashboard created using Tableau showing the Polarity of News gathered and Percentage Difference in Stock Price for Ford (F) along with Trend in Sentiments. We can clearly see that the sentiments for Ford 2nd Feb was negative leading to drop in Stock Price by 4.85% (news from website also attached)



Ford Motor Co. (DE) (F) Falls 4.85% for February 02

Equities Staff | *Tuesday, 02 February 2016 15:53* (EST)

One of the S&P 500's big losers for Tuesday February 02 was Ford Motor Co. (DE) (F). The company's stock fell 4.85% to $11.49 on volume of 44.95 million shares.

Conclusion and Observations

One of the limitations observed while preparing the model is the reliability of Twitter Data. As Organizations are aware of the twitter trend and how the sentiments impact the trend of their stock market, they run various positive campaigns such as “Fiesta Movement” causing buzz of the product on the social media trying to divert focus on low Sales of cars during the season or Quarterly Earnings Results etc. (http://www.bu.edu/today/2013/how-ford-became-a-leader-in-social-media/). Also the amount of cleaning efforts is high for Twitter owing to various styles and lingo of tweets. This also brings focus to the relevancy of the data. Any news such as Product Recall is a negative sentiment.

Through Text Analytics, we are able to understand the correlation of the news sentiment to the movement of stock prices in the Market. The current approach given the limitation of time and resources is limited to retrospective processing of data i.e. scraps through the internet for relevant data from various sources, processes the same and derives the trend in the sentiments of the data. This is then correlated to the movement of the stock prices. As we can see with the dashboards, the correlation is strong.

The way-forward for this project is to be able to scrap the data from various data sources in real-time there by processing their sentiments and providing the trend as a tool to investors (typically the day traders) in taking their stand (either long or short) accordingly. One suggestion is to move the model to cloud services such as AWS and have it run through the day scrapping real time news and tweets and process the sentiment score and through the trend enable the user to take the buy/sell call accordingly.

References

1. Deloitte -> http://www2.deloitte.com/us/en.html
2. Content extraction - identifying the main content in HTML document by Thomas Gottron - https://www.researchgate.net/publication/220690546_Content_extraction_-_identifying_the_main_content_in_HTML_documents
3. Trading Strategies to Exploit Blog and News Sentiment by Wenbin Zhang, Steven Skiena - http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1529
4. Google News Link => https://news.google.co.in/nwshp?hl=en&tab=wn
5. Twitter Website => https://twitter.com/
6. REST API => https://dev.twitter.com/rest/public
7. Plugins => https://github.com/mannau/tm.plugin.sentiment

Annexure

Twitter Sniffer Program

```
############# Twitter Data Capture ###########
#TO capture tweets related to Tesla , Ford and General Motors
####### Handles being sniffed #########
### Tesla: #TeslaMotors #Tesla #TSLA
### FORD: #ford #fordmotors
### General Motors: #generalmotors #gm
# Fetching required packages
install.packages("twitteR")
install.packages("RCurl")
install.packages("RJSONIO")
install.packages("stringr")
install.packages('base64enc')
install.packages('jsonlite')
install.packages("xlsx")

library(jsonlite)
library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
library(base64enc)
library(xlsx)

# Declare Twitter API Credentials
api_key = "9besLwERF5I6UARDLVd1OjRBn"
api_secret = "6O9NlAp8owLS12IjZDZudOfDnmFAMgAfqpIaf9AqGUJM26Mkbm"
access_token = "141123071-s20Z93Nizv84W0h8WVMNeYJrMl8sEyzWfJpFqZfF"
token_secret = "7AB0Nw4fhEwmHhpHgeABj4ZBigfdnvXIyHpRxhBNEljCN"

twitter_log = file("twitter_sniffer.log",open = "a")

putMeTOSleepCounter_Ford = 0
putMeTOSleepCounter_GM = 0
putMeTOSleepCounter_Tesla = 0

sleepCounter = 0
keepCapturing = TRUE


# Create Twitter Connection
setup_twitter_oauth(api_key, api_secret, access_token, token_secret)

cat(paste("** Creating Securties Data Frame ",Sys.Date()," **\n"),file=twitter_log)
```

```
# Defining the securities to be captured
securities = c("Ford","Tesla","General Motors")
hastags = c("'fordmotors' OR #ford OR #fordmotors",
             "'TeslaMotors' OR 'Tesla' OR 'TSLA' OR #TeslaMotors OR #Tesla OR #TSLA",
             "'GeneralMotors' OR #generalmotors OR #gm")

ford_dataFrame = data.frame()
tesla_dataFrame = data.frame()
generalmotors_dataFrame = data.frame()

securities_toCapture = data.frame(securities,hastags)

names(securities_toCapture) <- c("Security","HashTags") # variable names

##Tweet sniffing

while(keepCapturing){
    if(putMeTOSleepCounter_Ford == 3 & putMeTOSleepCounter_GM == 3 & putMeTOSleepCounter_Tesla == 3 ){
      print("## Didnt get any data for last 3 tries.. Sleeping for 15 mins before retry.##")
      cat("## Didnt get any data for last 3 tries.. Sleeping for 15 mins before retry. ##\n",file =
twitter_log)
      Sys.sleep(900)
      putMeTOSleepCounter_Ford = 0
      putMeTOSleepCounter_GM = 0
      putMeTOSleepCounter_Tesla = 0
      sleepCounter = sleepCounter + 1
    }
    if(sleepCounter == 3){
      print("ERROR: Didnt fetch data after 15 mins of sleep 3 times , time to quit !!")
      cat("ERROR: Didnt fetch data after 15 mins of sleep 3 times , time to quit !!\n",file =
twitter_log)
      keepCapturing = FALSE
      break
    }

      #Capture previous data
    for(i in 1:nrow(securities_toCapture)) {
      data <- securities_toCapture[i,]

      temp_dataFrame = data.frame()

      switch(as.character(data$Security),
             'Ford'={
               if(putMeTOSleepCounter_Ford == 3){
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it!!\n"))
```

```
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it !!\n"),file = twitter_log)
                 next
               }

               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n")
               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n",file = twitter_log)

               tweets = searchTwitter(toString(data$HashTags) ,
                                      n=100, lang="en", maxID = data$TailId)
               temp_dataFrame = twListToDF(tweets)



               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n")
               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n",file = twitter_log)

               if(nrow(temp_dataFrame) == 1 ){
                 print(" ** Insufficient tweets ** Skipping this iteration")
                 cat(" ** Insufficient tweets ** Skipping this iteration\n",file = twitter_log)
                 putMeTOSleepCounter_Ford = putMeTOSleepCounter_Ford + 1
                 next
               }

               print("Updating Ford Data Frame")
               cat("Updating Ford Data Frame\n",file = twitter_log)

               ford_dataFrame = rbind(ford_dataFrame,temp_dataFrame)
               securities_toCapture[i,"HeadId"] = head(ford_dataFrame,n=1)["id"]
               securities_toCapture[i,"TailId"] = tail(ford_dataFrame,n=1)["id"]

               if(nrow(ford_dataFrame) %% 500 == 0){
                 print("##### Dumping Ford Data Frame Checkpoint#####")
                 cat(paste("##### Dumping Ford Data Frame Checkpoint#",nrow(ford_dataFrame) %/% 500,
"#####\n"),file = twitter_log)
                 write.xlsx(ford_dataFrame,paste("Ford",nrow(ford_dataFrame) %/% 500,".xlsx",sep =
""))
               }
             },
             'Tesla'={
               if(putMeTOSleepCounter_Tesla == 3){
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it!!\n"))
```

```
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it !!\n"),file = twitter_log)
                 next
               }

               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n")
               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n",file = twitter_log)

               tweets = searchTwitter(toString(data$HashTags) ,
                                      n=100, lang="en", maxID = data$TailId)
               temp_dataFrame = twListToDF(tweets)

               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n")
               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n",file = twitter_log)

               if(nrow(temp_dataFrame) == 1 ){
                 print(" ** Insufficient tweets ** Skipping this iteration")
                 cat(" ** Insufficient tweets ** Skipping this iteration\n",file = twitter_log)
                 putMeTOSleepCounter_Tesla = putMeTOSleepCounter_Tesla + 1
                 next
               }

               print("Updating Tesla Data Frame")
               cat("Updating Tesla Data Frame\n",file = twitter_log)

               tesla_dataFrame = rbind(tesla_dataFrame,temp_dataFrame)
               securities_toCapture[i,"HeadId"] = head(tesla_dataFrame,n=1)["id"]
               securities_toCapture[i,"TailId"] = tail(tesla_dataFrame,n=1)["id"]



               if(nrow(tesla_dataFrame) %% 500 == 0){
                 print("##### Dumping Tesla Data Frame Checkpoint#####")
                 cat(paste("##### Dumping Tesla Data Frame Checkpoint#",nrow(tesla_dataFrame) %/% 500,
"#####\n"),file = twitter_log)
                 write.xlsx(tesla_dataFrame,paste("Tesla",nrow(tesla_dataFrame) %/% 500,".xlsx",sep =
""))
               }
             },
             'General Motors'={
               if(putMeTOSleepCounter_GM == 3){
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it!!\n"))
```

```
                 cat(paste("WARNING: No data fetched for last 3 tries for ",data$Security,", skipping
it !!\n"),file = twitter_log)
                 next
               }

               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n")
               cat(paste("********* Capturing Past Data for: ",data$Security," for tags:
",data$HashTags)," *********\n",file = twitter_log)

               tweets = searchTwitter(toString(data$HashTags) ,
                                      n=100, lang="en", maxID = data$TailId)
               temp_dataFrame = twListToDF(tweets)

               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n")
               cat(paste("*** Fetched : ",nrow(temp_dataFrame)," tweets for Security:
",data$Security)," *** \n",file = twitter_log)

               if(nrow(temp_dataFrame) == 1 ){
                 print(" ** Insufficient tweets ** Skipping this iteration")
                 cat(" ** Insufficient tweets ** Skipping this iteration\n",file = twitter_log)
                 putMeTOSleepCounter_GM = putMeTOSleepCounter_GM + 1
                 next
               }

               print("Updating General Motors Data Frame")
               cat("Updating General Motors Data Frame\n",file = twitter_log)

               generalmotors_dataFrame = rbind(generalmotors_dataFrame,temp_dataFrame)
               securities_toCapture[i,"HeadId"] = head(generalmotors_dataFrame,n=1)["id"]
               securities_toCapture[i,"TailId"] = tail(generalmotors_dataFrame,n=1)["id"]

               if(nrow(generalmotors_dataFrame) %% 500 == 0){
                 print("##### Dumping GM Data Frame Checkpoint#####")
                 cat(paste("##### Dumping GM Data Frame Checkpoint#",nrow(generalmotors_dataFrame) %/%
500, "#####\n"),file = twitter_log)
                 write.xlsx(generalmotors_dataFrame,paste("GM",nrow(generalmotors_dataFrame) %/%
500,".xlsx",sep = ""))
               }
             },
             {
               print('Invalid Security')
               cat("!!!!!!!!!!!Invalid Security!!!!!!!!\n",file = twitter_log)
             }
      )
    }
```

```
}


  #Dump Data frames to Files
  for(i in 1:nrow(securities_toCapture)) {
    data <- securities_toCapture[i,]

    switch(as.character(data$Security),
           'Ford'={
             print("##### Dumping Ford Data Frame #####")
             cat("##### Dumping Ford Data Frame #####\n",file = twitter_log)
             write.xlsx(ford_dataFrame,"Ford_Final.xlsx")
             #myjson = toJSON(ford_dataFrame, pretty = TRUE)
             #save(myjson, file=paste("Ford",Sys.Time(),".xlsx",sep = ""))
           },
           'Tesla'={
             print("##### Dumping Tesla Data Frame #####")
             cat("##### Dumping Tesla Data Frame #####\n",file = twitter_log)
             write.xlsx(tesla_dataFrame,"Tesla_Final.xlsx")
           },
           'General Motors'={
             print("##### Dumping GM Data Frame #####")
             cat("##### Dumping GM Data Frame #####\n",file = twitter_log)
             write.xlsx(generalmotors_dataFrame,"GM_Final.xlsx")
           },
           {
             print('Invalid Security')
           }
    )
    # do stuff with row
   # print(row$hastags)
  }

#Close the log file
close(twitter_log)
```

Transformation

```
`Subj` <-
function(x)
{
        if(has.Subj(x))
                return(x[,grep('Subjectivity',colnames(x),ignore.case=TRUE)])
        stop('subscript out of bounds: no column name containing "Score"')
}
```

```
`has.Subj` <-
function(x,which=FALSE)
{
        loc <- grep('Subjectivity',colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}




`Polarity` <-
                function(x)
{
        if(has.Polarity(x))
                return(x[,grep('Polarity',colnames(x),ignore.case=TRUE)])
        stop('subscript out of bounds: no column name containing "Polarity"')
}


`has.Polarity` <-
function(x,which=FALSE)
{
        loc <- grep('Polarity',colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}



`Pos_Refs_Per_Ref` <-
function(x)
{
        if(has.Pos_Refs_Per_Ref(x))
                return(x[,grep('Pos_Refs_Per_Ref',colnames(x),ignore.case=TRUE)])
        stop('subscript out of bounds: no column name containing "Pos_Refs_Per_Ref"')
}

`has.Pos_Refs_Per_Ref` <-
                function(x,which=FALSE)
{
        loc <- grep(‘Pos_Refs_Per_Ref’,colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}


`Neg_Refs_Per_Ref` <-
                function(x)
{
        if(has.Neg_Refs_Per_Ref(x))
                return(x[,grep('Neg_Refs_Per_Ref',colnames(x),ignore.case=TRUE)])
        stop('subscript out of bounds: no column name containing "Neg_Refs_Per_Ref"')
}


`has.Neg_Refs_Per_Ref` <-
                function(x,which=FALSE)
{
        loc <- grep('Neg_Refs_Per_Ref',colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}


`Senti_Diffs_Per_Ref` <-
                function(x)
{
        if(has.Senti_Diffs_Per_Ref(x))
                return(x[,grep('Senti_Diffs_Per_Ref',colnames(x),ignore.case=TRUE)])
        stop('subscript out of bounds: no column name containing "Senti_Diffs_Per_Ref"')
}


`has.Senti_Diffs_Per_Ref` <-
function(x,which=FALSE)
{
        loc <- grep('Senti_Diffs_Per_Ref',colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}


`SentVol` <-
function(x, prefix = "SEMNT")
{
        name = paste(prefix, "vol", sep=".")
        if(has.SentVol(x, prefix = prefix))
                return(x[,grep(name,colnames(x),ignore.case=TRUE)])
        stop(paste('subscript out of bounds: no column name containing"', name, '"'))
}


`has.SentVol` <-
function(x,which=FALSE, prefix = "SEMNT")
{
        name = paste(prefix, "vol", sep=".")
        loc <- grep(name,colnames(x),ignore.case=TRUE)
        if(!identical(loc,integer(0)))
                return(ifelse(which,loc,TRUE))
        ifelse(which,loc,FALSE)
}
```

Add Sentiment columns

```
addSubjectivity <- newTA(Subj, col=1)
addPolarity <- newTA(Polarity, col=4)
addPos_Refs_Per_Ref <- newTA(Pos_Refs_Per_Ref, col=3)
addNeg_Refs_Per_Ref <- newTA(Neg_Refs_Per_Ref, col=2)
addSenti_Diffs_Per_Ref <- newTA(Senti_Diffs_Per_Ref, col=5)
addSentVol <- newTA(SentVol, col=6)

`addSentiment` <-
function(fieldname, on = NA){
        lchob <- quantmod:::get.current.chob()
        x <- as.matrix(lchob@xdata)



        Volumes = NULL
        if(missing(fieldname)){
                Volumes <- Polarity(x)
                fieldname = "Polarity"
        }else{
                Volumes <- x[,fieldname]
        }


        max.vol <- max(Volumes, na.rm = TRUE)
        bar.col <- ifelse(Volumes > 0, lchob@colors$up.col, lchob@colors$dn.col)

        border.col <- ifelse(is.null(lchob@colors$border), bar.col,
                        lchob@colors$border)
        bar.col <- bar.col[lchob@xsubset]
        chobTA <- new("chobTA")

        if (any(is.na(on))) {
                chobTA@new <- TRUE
        }
        else {
                chobTA@new <- FALSE
                chobTA@on <- on
        }

        chobTA@TA.values <- (Volumes)[lchob@xsubset]
        chobTA@name <- "chartSent"
        chobTA@call <- match.call()
        chobTA@params <- list(xrange = lchob@xrange, colors = lchob@colors,
                        color.vol = lchob@color.vol, multi.col = lchob@multi.col,
                        spacing = lchob@spacing, width = lchob@width, bp = lchob@bp,
                        x.labels = lchob@x.labels, log.scale = FALSE,
                        bar.col = bar.col, border.col = border.col, time.scale =
lchob@time.scale, vol.scale=list(1, fieldname))
        chobTA@params$thin <- ifelse(lchob@type %in% c("bars", "matchsticks"),
                        TRUE, FALSE)
        if (is.null(sys.call(-1))) {
                TA <- lchob@passed.args$TA
                lchob@passed.args$TA <- c(TA, chobTA)
                lchob@windows <- lchob@windows + ifelse(chobTA@new, 1,
                                0)
                FUN <- quantmod:::chartSeries.chob
                do.call("FUN", list(lchob))
                invisible(chobTA)
        }
        else {
                return(chobTA)
        }

}


addSentimentVo <- function(fieldname, color = "blue"){
        lchob <- quantmod:::get.current.chob()
        x <- as.matrix(lchob@xdata)

        Volumes = NULL
        if(missing(fieldname)){
                Volumes <- SentVol(x)
        }else{
                Volumes <- x[,fieldname]
        }

        max.vol <- max(Volumes, na.rm = TRUE)


        bar.col <- ifelse(Volumes > 0, color, "red")

        border.col <- ifelse(is.null(lchob@colors$border), bar.col,
                      lchob@colors$border)

        bar.col <- color

        chobTA <- new("chobTA")
        chobTA@new <- TRUE
        chobTA@TA.values <- (Volumes)[lchob@xsubset]
        chobTA@name <- "chartVo"
        chobTA@call <- match.call()
        chobTA@params <- list(xrange = lchob@xrange, colors = lchob@colors,
                      color.vol = TRUE, multi.col = lchob@multi.col,
                      spacing = lchob@spacing, width = lchob@width, bp = lchob@bp,
                      x.labels = lchob@x.labels, log.scale = FALSE,
                      bar.col = bar.col, border.col = border.col, time.scale =
lchob@time.scale, vol.scale=list(1, "Text Documents"))
        chobTA@params$thin <- ifelse(lchob@type %in% c("bars", "matchsticks"),
                      TRUE, FALSE)
        if (is.null(sys.call(-1))) {
                TA <- lchob@passed.args$TA
                lchob@passed.args$TA <- c(TA, chobTA)
                lchob@windows <- lchob@windows + ifelse(chobTA@new, 1,
                             0)
                FUN <- quantmod:::chartSeries.chob
                do.call("FUN", list(lchob))

                invisible(chobTA)
        }
        else {
                return(chobTA)
        }

}


`chartSent` <-
```

```
function(x) {
        if(class(x) != "chobTA") stop("chartSentiment requires a suitable chobTA object")
        Volumes <- x@TA.values

        spacing <- x@params$spacing
        width <- x@params$width

        x.range <- x@params$xrange
        x.range <- seq(x.range[1],x.range[2]*spacing)


        color.vol <- x@params$color.vol
        log.scale <- ifelse(x@params$log.scale,"y","")

        vol.scale <- x@params$vol.scale

        if(x@new) {
                plot.new()
                plot.window(xlim=c(1, x@params$xrange[2] * spacing),
                                ylim=c(min(Volumes,na.rm=TRUE),max(Volumes,na.rm=TRUE)),
                                log=log.scale)
                coords <- par('usr')
                rect(coords[1],coords[3],coords[2],coords[4],col=x@params$colors$area)
                abline(h=axTicks(2), col=x@params$colors$grid.col, lty='dotted')
        }

        x.pos <- 1 + spacing * (1:length(Volumes) - 1)


        bar.col <- x@params$bar.col
        border.col <- x@params$border.col

        if(x@params$thin) {

                segments(x.pos,0,x.pos,Volumes,col=bar.col)
        } else {
                rect(x.pos-spacing/3,0,x.pos+spacing/3,Volumes,
                                col=bar.col,border=border.col)
        }
        legend.text <- list(list(
                                        legend=c(paste("Sentiment
Score(",vol.scale[[2]],"):",sep=''),format(last(Volumes)*vol.scale[[1]],big.mark=',')),
                                        text.col=c(x@params$colors$fg.col, last(bar.col))
                        ))
        legend("topleft",
                        legend=c(paste("Sentiment
Score(",vol.scale[[2]],"):",sep=''),format(last(Volumes)*vol.scale[[1]],big.mark=',')),
                      text.col=c(x@params$colors$fg.col, last(bar.col)), bty="n",
y.inter=0.95)
```

```
        axis(2)
        box(col=x@params$colors$fg.col)
        invisible(vector('list',2))
} # }}}
```

Trend Visualization

```
chartSentiment <-
function (xts, prefix = "SEMNT",
              sentname = paste(prefix,"polarity", sep = "."),
              volname =  paste(prefix,"vol", sep = "."),
              volcolor="blue",
              na.fill = TRUE,
              omit.na.leading = TRUE,
              type = "line",
              theme='white',
              name,
              TA = "", postTA = "", ...)
{

       counter = 1

       if(omit.na.leading){
              firstnona <- min(which(!is.na(xts[,sentname])))
              xts <- xts[firstnona:NROW(xts),]
       }

       if(na.fill){
              xts[is.na(xts)] <- 0
       }

       if(missing(name)){
              name = ""
              if(has.Cl(xts))
```

```
                name <- strsplit(colnames(Cl(xts)), "\\.", perl = TRUE)[[1]][1]
        else
                name <- strsplit(colnames(xts[,1]), "\\.", perl = TRUE)[[1]][1]

        name <- paste("Sentiment", name)
}

if(sentname != ""){
        TA <- paste(TA, "addSentiment('", sentname, "');", sep = "")
        counter = counter + 1
}

if(has.SentVol(xts)){
        TA <- paste(TA, "addSentimentVo('", volname, "');", sep = "")
        counter = counter + 1
}

if(has.Vo(xts)){
        TA <- paste(TA, "addVo();", sep = "")
        counter = counter + 1
}

if(all(has.Subj(xts),
                has.Neg_Refs_Per_Ref(xts),
                has.Pos_Refs_Per_Ref(xts),
                has.Senti_Diffs_Per_Ref(xts))){
        range <- range(na.omit(rbind(Pos_Refs_Per_Ref(xts),
                                                   Neg_Refs_Per_Ref(xts),
                                                   Senti_Diffs_Per_Ref(xts),
                                                   Subj(xts))))

        addSubjectivityAll <- newTA(Subj, col=1, yrange = range)
        assign("addSubjectivityAll", addSubjectivityAll, envir = .GlobalEnv)

        TA <- paste(TA, "addSubjectivityAll();", sep = "")
        counter = counter + 1

        TA <- paste(TA, "addPos_Refs_Per_Ref(on=", counter, ");", sep = "")
        TA <- paste(TA, "addNeg_Refs_Per_Ref(on=", counter, ");", sep = "")
        TA <- paste(TA, "addSenti_Diffs_Per_Ref(on=", counter, ");", sep = "")
}else{
        if(has.Subj(xts)){
                TA <- paste(TA, "addSubjectivity();", sep = "")#
        }

        if(has.Pos_Refs_Per_Ref(xts)){
                TA <- paste(TA, "addPos_Refs_Per_Ref();", sep = "")
                }

                if(has.Neg_Refs_Per_Ref(xts)){
                        TA <- paste(TA, "addNeg_Refs_Per_Ref();", sep = "")
                }

                if(has.Senti_Diffs_Per_Ref(xts)){
                        TA <- paste(TA, "addSenti_Diffs_Per_Ref();", sep = "")
                }
        }

        TA <- paste(TA, postTA, sep = ";")


        chartSeries(xts, type = "line", theme = "white",
                        name = name,
                        TA = TA, ...)
}
```

Corpus

```
"getField" <- function(x, ...) UseMethod("getField", x)
"getField.VCorpus" <-
function(x, fieldname){
        field <- do.call("c",lapply(x, function(y) meta(y, fieldname)))
        names(field) <- NULL
        field
}


"setField" <- function(x, ...) UseMethod("setField", x)

"setField.VCorpus" <-
function(x, fieldname, value){
        for(i in 1:length(x)){
                meta(x[[i]], fieldname) <- value[i]
        }
        x
}
```

Scrapping Relevant data

```
getRelevant <-
function(td, items, boundaries, matches.only = FALSE, fieldname = "matches"){
        if(missing(boundaries)){
                boundaries = "\\."
        }
        itempatt <- paste("\\W(",paste("(", items, ")", collapse = "|", sep = ""),")\\W",
sep = "")
        matches <- gregexpr(itempatt, Content(td), ignore.case = TRUE, perl = TRUE,  fixed =
FALSE, useBytes = FALSE)
        if(!matches.only){
                startend <- range(matches)
                delimmatches <- gregexpr(boundaries, Content(td), ignore.case = TRUE, perl =
TRUE,  fixed = FALSE, useBytes = FALSE)
                delimmatches <- as.integer(delimmatches[[1]])
                textlen <- nchar(Content(td))
                sentstart <- max(c(delimmatches[delimmatches < startend[1]], 0))
                sentend <- min(c(delimmatches[delimmatches > startend[2]], textlen))
                subtext <- substr(Content(td), sentstart+1, sentend)
                Content(td) <- substr(Content(td), sentstart, sentend)
        }
        if(any(matches[[1]] == -1)){
                meta(td, fieldname) <- 0
        }
        else{
                meta(td, fieldname) <- length(matches[[1]])
        }
        getRelevant <- td
}
```

Meta XTS

```
metaXTS <- function(corpus,
        fieldnames,
        period = "days",
        k=1,
        aggFUN = mean,
        symbol,
        prefix = "SEMNT",
        join = "inner",
        na.omit = TRUE){

        if(missing(fieldnames)){
                fieldnames <- colnames(meta(corpus))
        }

        idx <- do.call(c, lapply(corpus, meta, "datetimestamp"))
        df <- data.frame(idx, meta(corpus)[, fieldnames])

        if(na.omit){
                df <- na.omit(df)
        }

        xts <- xts(df[, -1], order.by = df[, 1])
        volxts <- xts(rep(1,NROW(xts)), order.by = index(xts))

        ep <- endpoints(xts, on = period, k = k)

        xts_agg = NULL
        if(length(ep) > 2){
                ldata <- lapply(1:NCOL(xts), function(x) period.apply(xts[,x], ep, aggFUN))
                ldata_all <- do.call("cbind", ldata)
                xts_agg <- xts(ldata_all, order.by = index(xts)[ep])
                volxts_agg <- period.apply(volxts, ep, sum)
        }else{
                #ep <- ep[ep > 0]

                xts_agg <- xts(t(as.data.frame(apply(xts, 2, aggFUN))), order.by =
index(xts)[ep])
                volxts_agg <- xts(sum(volxts), order.by = index(xts)[ep])
        }

        colnames(xts_agg) <- colnames(xts)
        colnames(volxts_agg) <- "vol"

        scorevolxts_agg <- cbind(xts_agg, volxts_agg)
        colnames(scorevolxts_agg) <- paste(prefix,colnames(scorevolxts_agg), sep=".")
        if(period == "days")
                index(scorevolxts_agg) <- as.Date(index(scorevolxts_agg))

        if(!missing(symbol)){
                names <- colnames(symbol)
                symbol <- to.period(symbol, period = period, k = k)
                colnames(symbol) <- names
                scorevolxts_agg <- merge(symbol, scorevolxts_agg, join = join)
        }

        return(scorevolxts_agg)
}
```

Pre Process the Corpus

```
preprocessCorpus <-
function(corpus, control, verbose = FALSE){

        if(missing(control)){
                control = list( removePunctuation = list(),
                                removeNumbers = list(),
                                tolower = list(),
                                removeWords = list(stopwords("english")),
                                stripWhitespace = list(),
                                stemDocument = list())

          }

          if(verbose)
                 cat("Starting Preprocessing...\n")

          for(n in names(control)){
                 if(verbose)
                        cat(n, " ...")
                 args <- control[[n]]
                 if(length(args) == 0){
                        corpus <- eval(call("tm_map", corpus, n))
                 }else{
                        corpus <- eval(call("tm_map", corpus, n, unlist(args)))
                 }
                 if(verbose)
                        cat("Done\n")
          }
          Corpus
   }
```

Score Function

```
  `score` <- function(corpus,
                                      control,
                                      scoreFUNS){

         if(missing(control)){
                control = list(
                               tolower = TRUE,
                               removePunctuation = TRUE,
                               removeNumbers = TRUE,
                               removeWords = list(stopwords("english")),
                               stripWhitespace = TRUE,
                               stemDocument = TRUE,
                               minWordLength = 3,
                               weighting = weightTf)
         }
         if(missing(scoreFUNS)){
                scoreFUNS = list(
                                polarity = list(),
                                subjectivity = list(),
                                pos_refs_per_ref = list(),
                                neg_refs_per_ref = list(),
                                senti_diffs_per_ref = list()
                )
        }

        tdm <- TermDocumentMatrix(corpus, control = control)


        res <- list()
        for(n in names(scoreFUNS)){
                args <- unlist(scoreFUNS[[n]])
                if(is.null(args)){
                        res[[n]] <- eval(call(n, tdm))
                }else{
                        res[[n]] <- eval(call(n, tdm, args))
                }
        }

        dfres <-  as.data.frame(res)
        meta <- meta(corpus)

  for(n in colnames(dfres)) {
    meta(corpus, n) <- dfres[, n]
  }
        Corpus
}

polarity <- function(x, positive, negative) UseMethod("polarity", x)

polarity.TermDocumentMatrix <- function(x, positive = posterms_GI(), negative =
negterms_GI() ){
        pos <- tm_term_score(x, positive)
        neg <-  tm_term_score(x, negative)
        all <- (pos+neg)
  ifelse(all != 0, (pos-neg)/all, 0)
}

polarity.DocumentTermMatrix <- function(x, ...) polarity(t(x), ...)

subjectivity <- function(x, positive, negative) UseMethod("subjectivity", x)

subjectivity.TermDocumentMatrix <- function(x, positive = posterms_GI(), negative =
negterms_GI() ){
        pos <- tm_term_score(x, positive)
        neg <-  tm_term_score(x, negative)
        all <- col_sums(x)
  ifelse(all != 0, (pos+neg)/all, 0)
}

subjectivity.DocumentTermMatrix <- function(x, ...) subjectivity(t(x), ...)

pos_refs_per_ref <- function(x, positive) UseMethod("pos_refs_per_ref", x)

pos_refs_per_ref.TermDocumentMatrix <- function(x, positive = posterms_GI()){
        pos <- tm_term_score(x, positive)
        all <- col_sums(x)
  ifelse(all != 0, (pos)/(all), 0)
}

pos_refs_per_ref.DocumentTermMatrix <- function(x, ...) pos_refs_per_ref(t(x), ...)

neg_refs_per_ref <- function(x, negative) UseMethod("neg_refs_per_ref", x)

neg_refs_per_ref.TermDocumentMatrix <- function(x, negative = negterms_GI()){
        neg <- tm_term_score(x, negative)
        all <- col_sums(x)
  ifelse(all != 0, (neg)/(all), 0)
}

neg_refs_per_ref.DocumentTermMatrix <- function(x, ...) neg_refs_per_ref(t(x), ...)

senti_diffs_per_ref <- function(x, positive, negative) UseMethod("senti_diffs_per_ref", x)

senti_diffs_per_ref.TermDocumentMatrix <- function(x, positive = posterms_GI(), negative =
negterms_GI()){
        pos <- tm_term_score(x, positive)
        neg <- tm_term_score(x, negative)
        all <- col_sums(x)
  ifelse(all != 0, (pos-neg)/all, 0)
}

senti_diffs_per_ref.DocumentTermMatrix <- function(x, ...) senti_diffs_per_ref(t(x), ...)

posterms_GI <- function(){
        data("dic_gi")
        data <- get("dic_gi", pos=globalenv())
        data[["positive"]]
}

negterms_GI <- function(){
        data("dic_gi")
        data <- get("dic_gi", pos=globalenv())
        data[["negative"]]
}
```