Collective Sentiment analysis for stock market prediction

Based On The Study: F. Xu and V. Keelj, "Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction," *2014 IEEE 16th Conference on Business Informatics*, Geneva

Study Followed

“Collective Sentiment Mining of Microblogs in 24-hour Stock Price Movement Prediction”

Feifei Xu *and* ***Vlado Keselj***

Faculty of Computer Science, Dalhousie University, Canada

Why Collective Sentiment is important

How investors feel about your Company

Market Sentiment

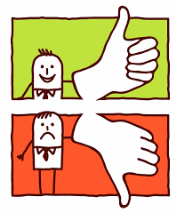
- Market sentiment is the overall attitude of investors toward a particular security or financial market.
- Market sentiment is the feeling or tone of a market, or its crowd psychology, as revealed through the activity and price movement of the securities traded in that market.
- Rising prices would indicate bullish market sentiment,
- while falling prices would indicate bearish market sentiment.

Source: Investopedia

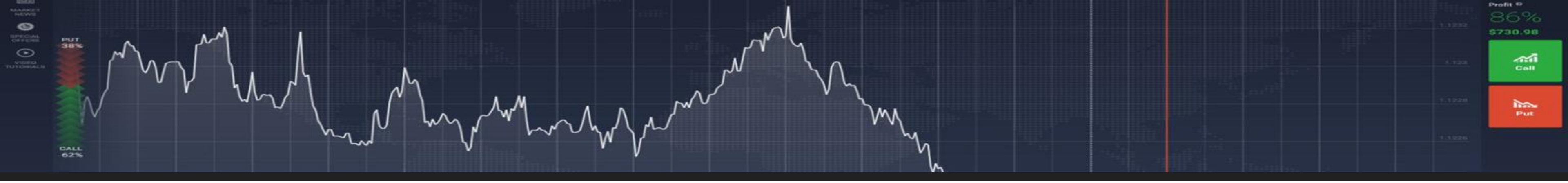
The Social Media Factor

How Twitter sentiments affect Stock Price

“Our ultimate goal is to test whether the feature of social media sentiments contributes to the stock market analysis, and to assess its predictive power.”
- Quote from study



DOWJONES
Nasda
60%
40%
20%
0%
-20%
Price
Jan '15
Mar '15
May '15
Jul '15
Sep '15
Nov '15
Jan '16
1/8/2016, 2:52:44 AM EDT powered by amigobulls.com



Our goal is 24 Stock Price Movement

An Listing

Alphabet Inc Class C
NASDAQ: GOOG

1,166.09 USD −20.78 (1.75%) ↓

Investors Discuss on StockTwits about $GOOG

A collection of tweets which show why DG Vanzara, accused in fake encounter cases involving 7 victims, was Modiji's favourite cop.

DG Vanzara

#RSS4India & IndiaWorld. RSS is ultimately for Universal Brotherhood of Mankind on Earth which is true Spirit of Hinduism since millennia.

DG Vanzara

RSS is d Fountain Head of Patriotism in India.Slogan ofRSS WITH ISIS can b coined only by Anti-nationals &Traitors who aim to ruin country

DG Vanzara

It is not sufficient to #RemoveMughalsFromBooks only, but those who faithfully claim to b their descendants shud also b removed frm India

DG Vanzara

Under d sanctifying banner of #DemocracyUnderAttack, the divisive forces are encouraging Anti National Elements to divide & destroy India.

TruthOfGujarat TruthOfGujarat

DG Vanzara

Pro-terror Traitors like Khalid & Anirban got bail within DAYS while Anti-Terror Gujarat Police got bail after 8 YEARS. Oh God save India !

DG Vanzara

Those who say no to #BharatMataKiJay r reviving d language of Pre-partition leaders of Muslim League thereby re threatening d unity of India.

DG Vanzara

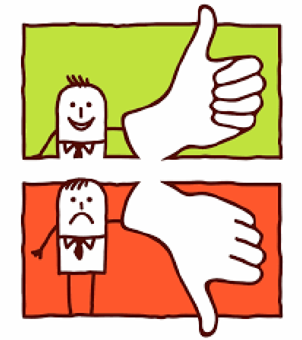
Oh Anti-national Traitors & Enemies of India, b ready to face d Last Day of Judgment from We, the People of India. Your countdown hs started.

DG Vanzara

Patriots of India, get united, stand like Himalayas & be liberated of your Mental Inhibitions inculcated by last 1200 years of Foreign Rule.

For more updates, follow us on: TruthOfGujarat TruthOfGujarat

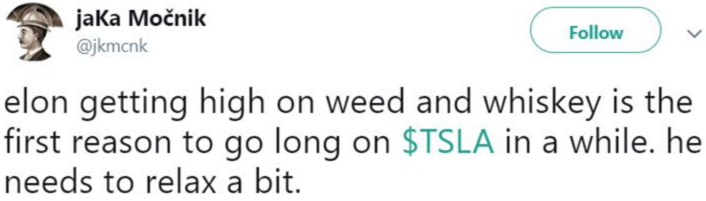
A Collective Sentiment is Formed For the Day



A scenario

CEO Smokes on LIVE interview

Twitter goes mad

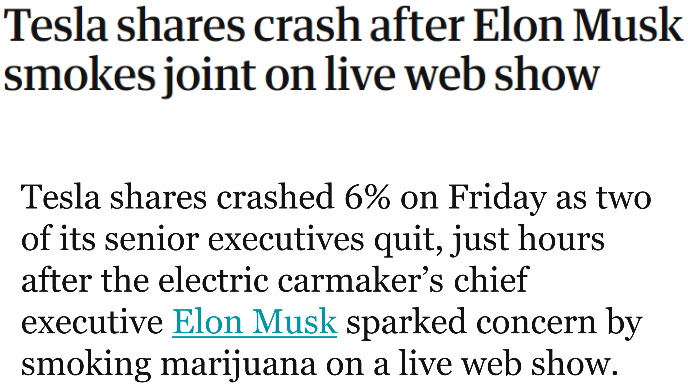


jaKa Močnik
@jkmcnk

Follow

elon getting high on weed and whiskey is the first reason to go long on $TSLA in a while. he needs to relax a bit.

Next day Stock Prices Reflect Twitter Sentiment

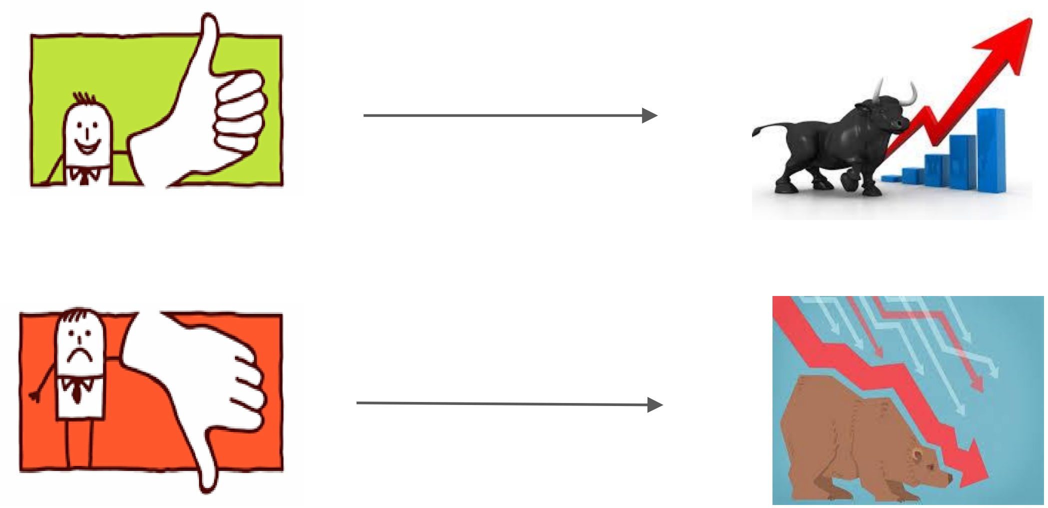


Tesla shares crash after Elon Musk smokes joint on live web show

Tesla shares crashed 6% on Friday as two of its senior executives quit, just hours after the electric carmaker's chief executive Elon Musk sparked concern by smoking marijuana on a live web show.

Source: The Guardian

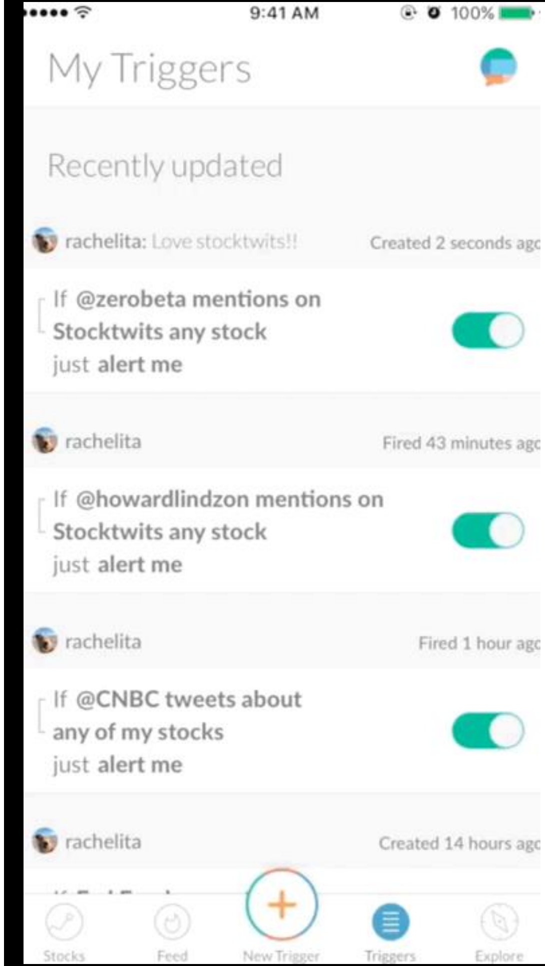
Our Hypothesis



- A positive market sentiment leads to more demand which leads to.
- A positive change in price of the stock the next day and vice versa

Scraped form StockTwits the social network for investors

Dataset



My Triggers

Recently updated

rachelita: Love stocktwits!! Created 2 seconds ago

If @zerobeta mentions on Stocktwits any stock just alert me

rachelita Fired 43 minutes ago

If @howardlindzon mentions on Stocktwits any stock just alert me

rachelita Fired 1 hour ago

If @CNBC tweets about any of my stocks just alert me

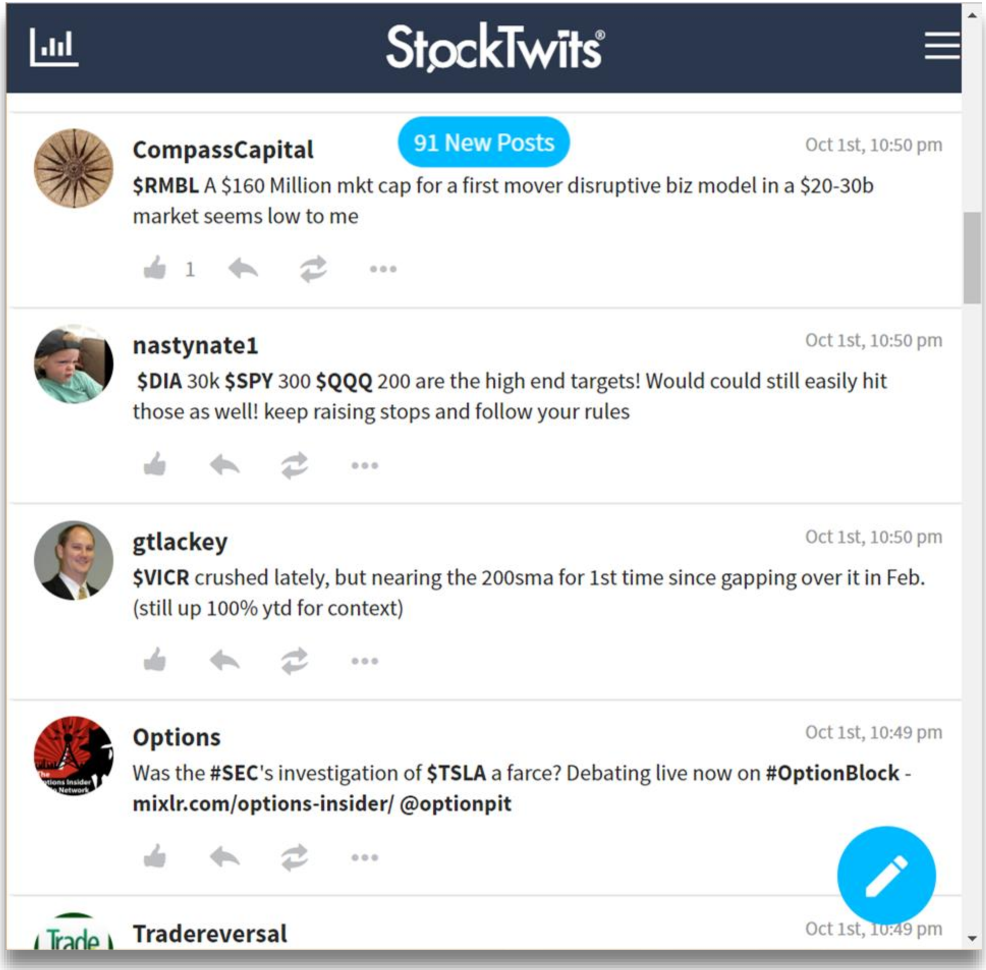
rachelita Created 14 hours ago

StockTwits

Real Investors. Real Ideas. Real Time.

StockTwits is a social media platform designed for sharing ideas between investors.

The company received the first annual Shorty Award in the 2008 finance category.



CompassCapital Oct 1st, 10:50 pm

$RMBL A $160 Million mkt cap for a first mover disruptive biz model in a $20-30b market seems low to me

nastynate1 Oct 1st, 10:50 pm

$DIA 30k $SPY 300 $QQQ 200 are the high end targets! Would could still easily hit those as well! keep raising stops and follow your rules

gtlackey Oct 1st, 10:50 pm

$VICR crushed lately, but nearing the 200sma for 1st time since gapping over it in Feb. (still up 100% ytd for context)

Options Oct 1st, 10:49 pm

Was the #SEC's investigation of $TSLA a farce? Debating live now on #OptionBlock - mixlr.com/options-insider/ @optionpit

Tradereversal Oct 1st, 10:49 pm

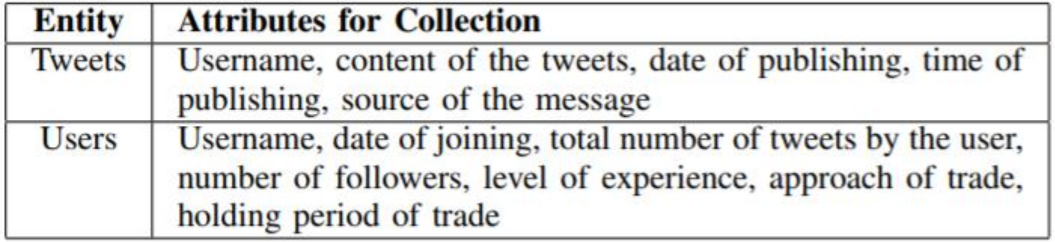
< Example Tweets

It’s an active community of investors discussing stocks

Top 16 Most Discussed Stocks

The study works on top 16 Stocks which are the most discussed ones on StockTwits.

Tweets of those stocks are collected over a period of two and a half months, along with the information of the username, date and time of publishing, source of the message, and other information which reflects the publishers profile.



Entity	Attributes for Collection
Tweets	Username, content of the tweets, date of publishing, time of publishing, source of the message
Users	Username, date of joining, total number of tweets by the user, number of followers, level of experience, approach of trade, holding period of trade

Pre-process the StockTwits

The following steps are taken to pre-process the stock tweets:

- Remove tweets on **weekends** and **public holidays**.
- Remove **duplicated tweets** by the same user.
- Replace **@ sign** in the tweets with text **“atreplace”**.
- Replace **$TICKER** for each stock in the tweets with text **“stocksignreplace”**.
- Replace **links** in the tweets with text **“linkreplace”**.
- Convert all the tweets **text** to be **lowercase**.

There are approximately 100,000 tweets for the 16 stocks included in the final analysis.

Sentiment lexicons

Sentiment lexicons are pre-trained models for example SentiWordNet, Natural Language Processing (SentiWords), Wordstat.

The vast majority of work in sentiment analysis mainly focuses on the domains of movie reviews, product reviews and blogs.

Financial Tweets don’t just fit in

The tweet “Short $AAPL @557.50”.
if regular lexicons are used, the sentiment will probably be marked as objective or neutral.
while in finance the word short is a clear sign indicating that the user expects the $AAPL stock to fall.

Liquidity is how easily you can get into and out of a stock.

Trading volume is the number of shares being traded each day – a factor that has huge implications for a stock’s liquidity.

You’re betting that the company’s stock will increase in price so that you can buy low and sell high.

This is when an investor buys more of a stock as the price goes down. This results in a decrease of the average price at which the investor purchased the stock.

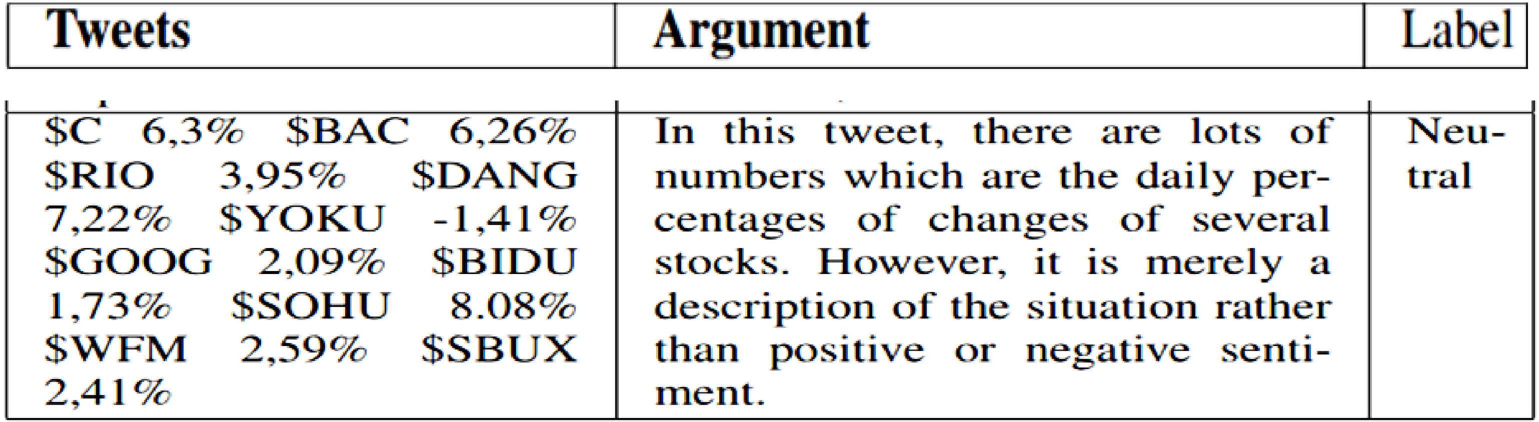
The word lists developed for other disciplines misclassify common words in financial texts.

The Solution (In the study)

- A total number of 2380 tweets on all of the 16 stocks are hand-labeled as the training and testing data.
- The tweets are labeled as positive (1), negative (-1) or neutral (0). The hand-labeled data are then randomized to generate training and testing dataset.
- For some tweets, it is difficult to label its polarity even by human efforts. In the case of vagueness, the tweets are labeled as neutral.

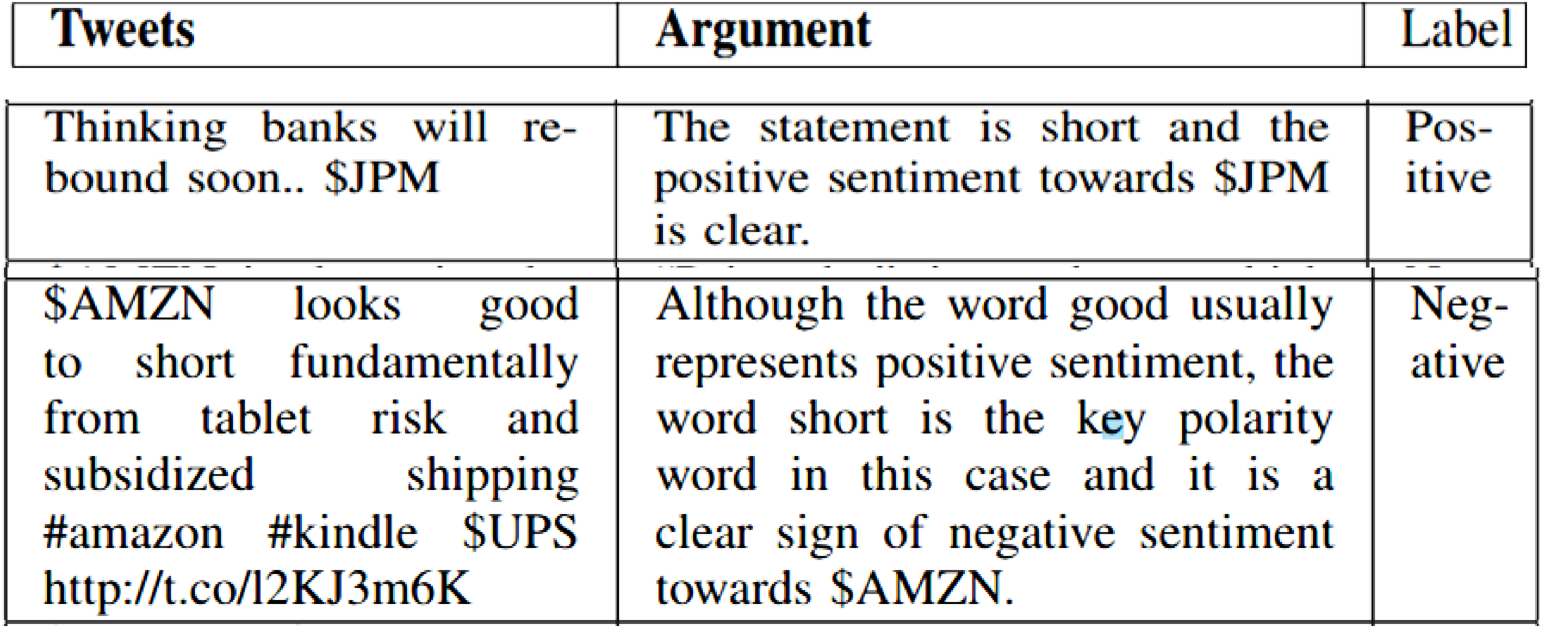
Tweets Sentiment Hand-labelling

→ **If the tweet contains external links of long articles or numerical charts about the stocks, it is generally marked as neutral.**



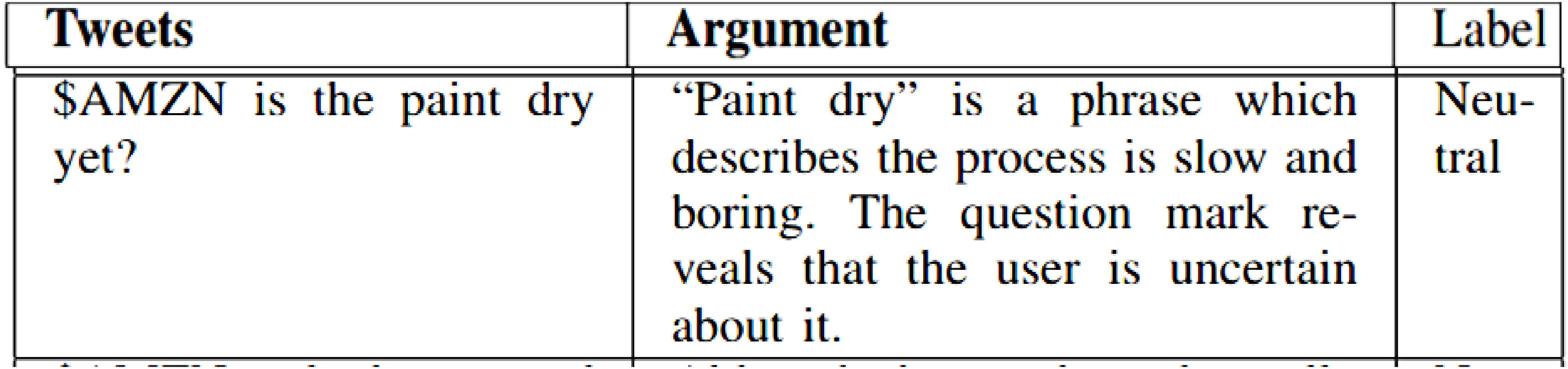
Tweets	Argument	Label
$C 6,3% $BAC 6,26% $RIO 3,95% $DANG 7,22% $YOKU -1,41% $GOOG 2,09% $BIDU 1,73% $SOHU 8.08% $WFM 2,59% $SBUX 2,41%	In this tweet, there are lots of numbers which are the daily percentages of changes of several stocks. However, it is merely a description of the situation rather than positive or negative sentiment.	Neutral

- ➔ **Positive or negative labels are only given when the sentiment can be explicitly speculated from the tweet.**



Tweets	Argument	Label
Thinking banks will rebound soon.. $JPM	The statement is short and the positive sentiment towards $JPM is clear.	Positive
$AMZN looks good to short fundamentally from tablet risk and subsidized shipping #amazon #kindle $UPS http://t.co/l2KJ3m6K	Although the word good usually represents positive sentiment, the word short is the key polarity word in this case and it is a clear sign of negative sentiment towards $AMZN.	Negative

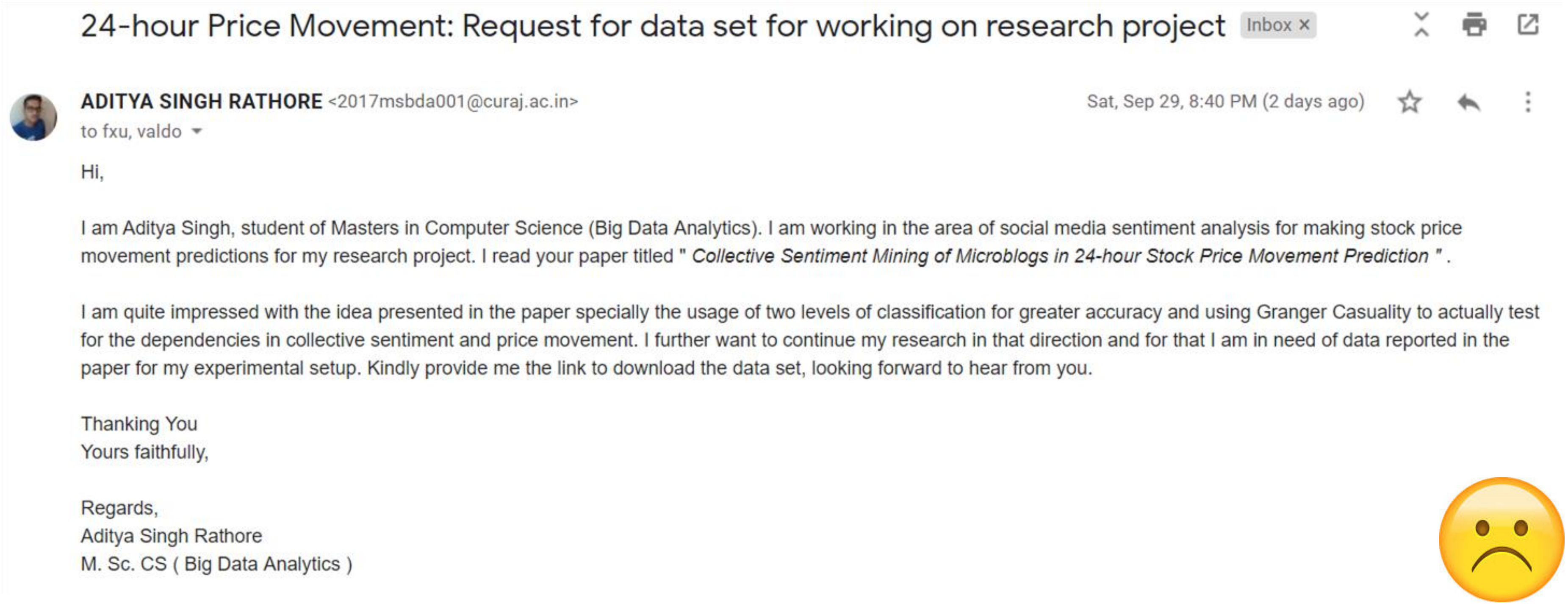
- ➔ Tweets with question marks are generally marked as neutral.



Tweets	Argument	Label
$AMZN is the paint dry yet?	“Paint dry” is a phrase which describes the process is slow and boring. The question mark reveals that the user is uncertain about it.	Neutral

Data is Money

Specifically in Banking Sector. So it’s natural to not share research data



24-hour Price Movement: Request for data set for working on research project

ADITYA SINGH RATHORE <2017msbda001@curaj.ac.in>
to fxu, valdo

Sat, Sep 29, 8:40 PM (2 days ago)

Hi,

I am Aditya Singh, student of Masters in Computer Science (Big Data Analytics). I am working in the area of social media sentiment analysis for making stock price movement predictions for my research project. I read your paper titled " *Collective Sentiment Mining of Microblogs in 24-hour Stock Price Movement Prediction* " .

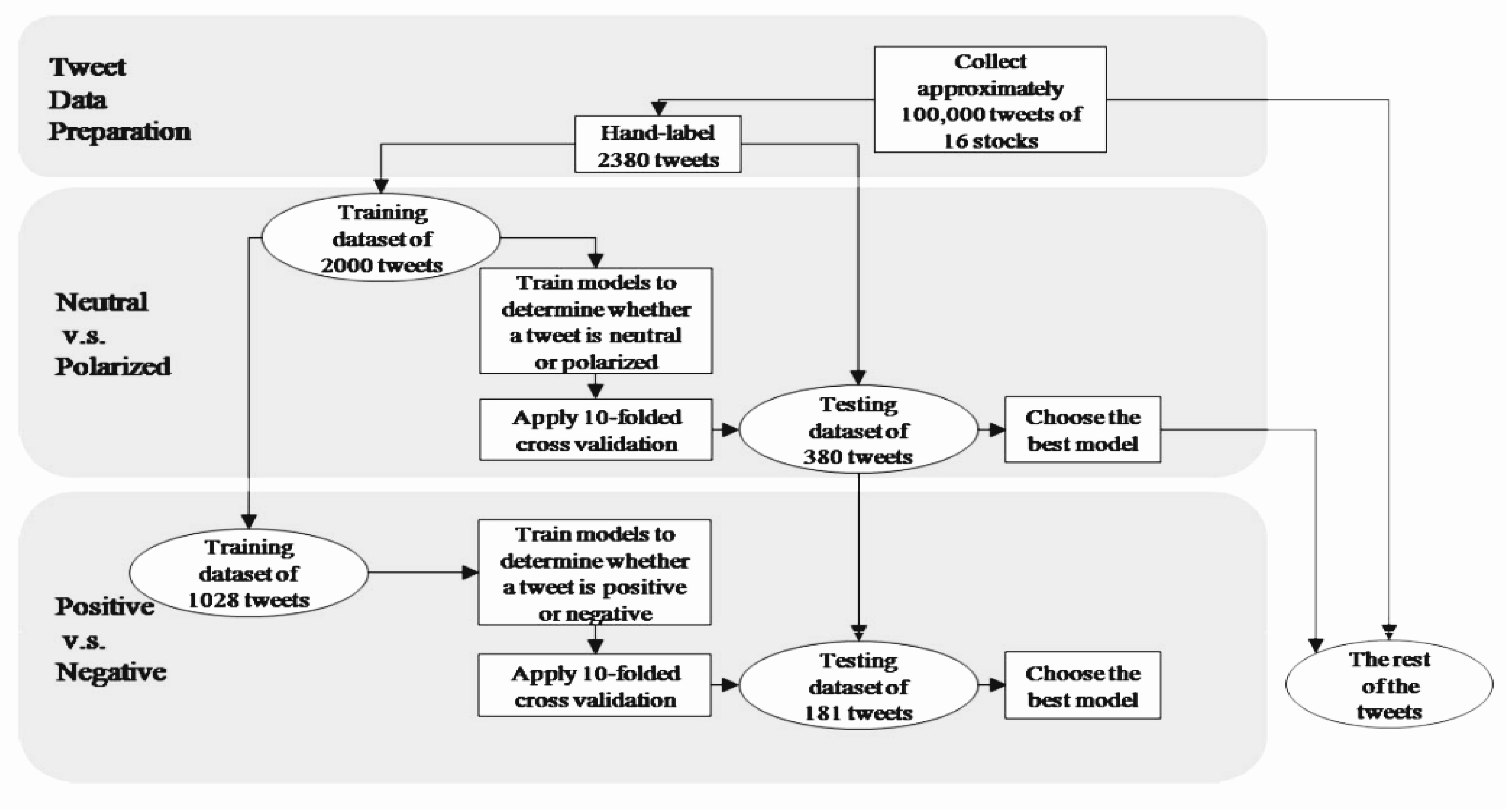
I am quite impressed with the idea presented in the paper specially the usage of two levels of classification for greater accuracy and using Granger Casuality to actually test for the dependencies in collective sentiment and price movement. I further want to continue my research in that direction and for that I am in need of data reported in the paper for my experimental setup. Kindly provide me the link to download the data set, looking forward to hear from you.

Thanking You
Yours faithfully,

Regards,
Aditya Singh Rathore
M. Sc. CS (Big Data Analytics)

☹

Authors’ Sentiment Detection Schema

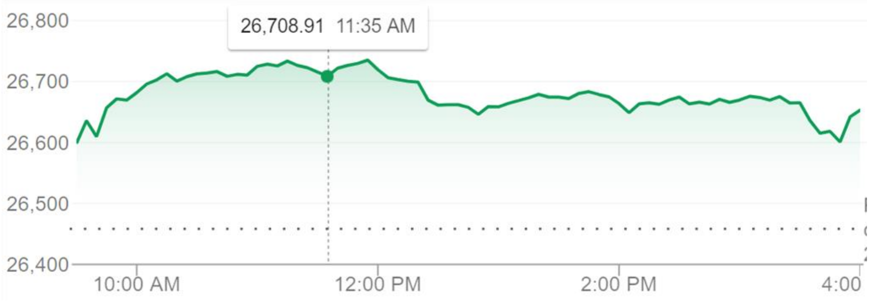

Tweet Data Preparation
Collect approximately 100,000 tweets of 16 stocks
Hand-label 2380 tweets
Neutral v.s. Polarized
Training dataset of 2000 tweets
Train models to determine whether a tweet is neutral or polarized
Apply 10-folded cross validation
Testing dataset of 380 tweets
Choose the best model
Positive v.s. Negative
Training dataset of 1028 tweets
Train models to determine whether a tweet is positive or negative
Apply 10-folded cross validation
Testing dataset of 181 tweets
Choose the best model
The rest of the tweets

Our solution

Sentiment analysis to predict the movement of the whole market.

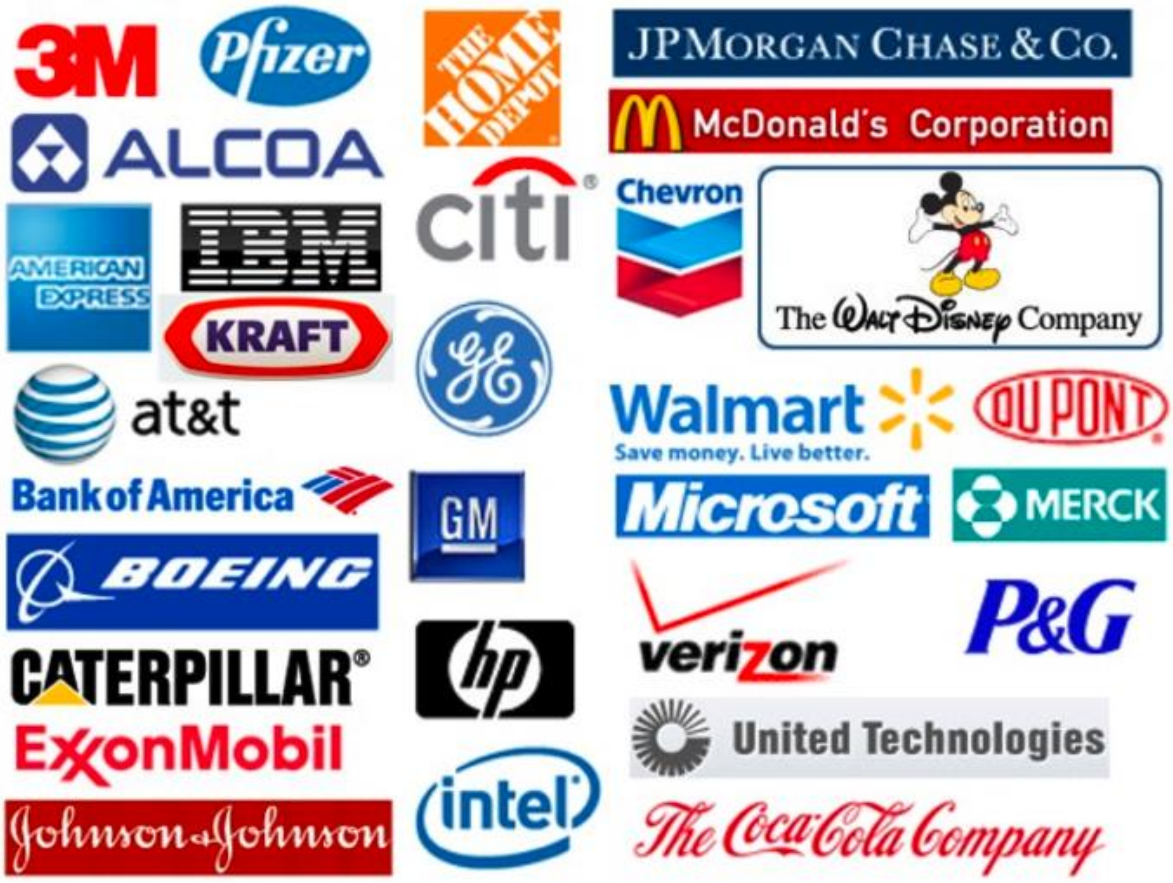
On the basis of Dow Jones Industrial Average

The Dow Jones Industrial Average (DJIA) is a stock market index that shows how 30 large, publicly owned companies based in the United States have traded during a trading session in the stock market



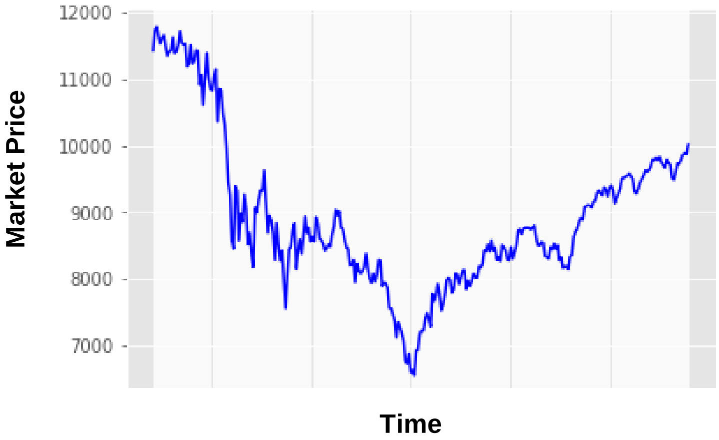
30 Companies

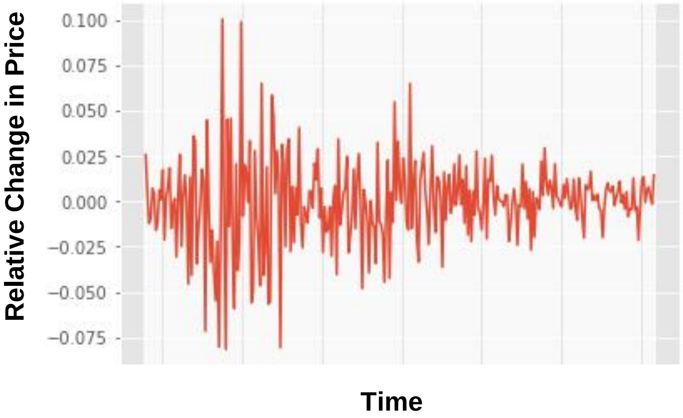
DOWJONES =



It gives an idea about how the Economy is performing as a whole. Instead of a single stock

The Daily Relative Change in DJIA (Our Dataset)





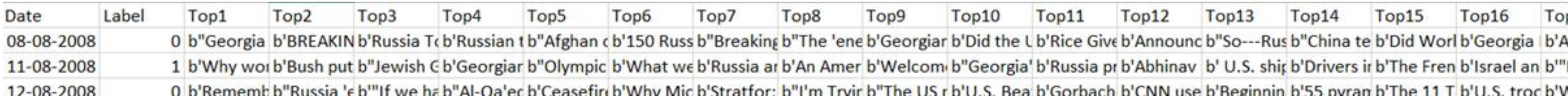
If the series of price is Pi, the daily relative change will be:

$$\frac{P_i - P_{i-1}}{P_{i-1}}$$

Our Dataset

- Top 25 Tweets about DJIA everyday from August 2008 to July 2016. With sentiment labels -1 (Neutral/Negative) or 1(Positive)

Date	Label	Tweet 1		Tweet 25
01-02-2011	-1	Historically, on an average the market declines mostly in the month of September. The three leading indicator #DJIA (Dow Jones), S&P 500 (Standard and Poor) and #NASDAQ have seemed to be performing poorly in this month.	• • •	INDU surges 0.73% closing $192.90 higher: techniquant.com/reports…Sentiment: Neutral $DJIA #DJIA $DOW #DOW



Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10	Top11	Top12	Top13	Top14	Top15	Top16	To
08-08-2008	0	b"Georgia	b'BREAKIN	b'Russia To	b'Russian t	b"Afghan c	b'150 Russ	b"Breaking	b"The 'ene	b'Georgian	b'Did the U	b'Rice Give	b'Announc	b"So---Rus	b"China te	b'Did Worl	b'Georgia	b'A
11-08-2008	1	b'Why wor	b'Bush put	b"Jewish G	b'Georgian	b"Olympic	b'What we	b'Russia ar	b'An Amer	b'Welcom	b"Georgia'	b'Russia pr	b'Abhinav	b' U.S. ship	b'Drivers ir	b'The Fren	b'Israel an	b'"
12-08-2008	0	b'Rememb	b"Russia 'e	b'"If we ha	b"Al-Qa'ec	b'Ceasefir	b'Why Mic	b'Stratfor:	b"I'm Tryir	b"The US r	b'U.S. Bea	b'Gorbach	b'CNN use	b'Beginnin	b'55 pyran	b'The 11 T	b'U.S. troc	b'V

Extracting Sentiments

Punctuation: Punctuation is treated as a feature in determining whether the tweet is neutral or polarized.

Neutral Just a question

BAD?

Length: lengthy contributions in tweet data contain elaborated explanations, which are important to detect in determining whether it is a neutral statement or a polarized declaration

Highly negative

SHAMINGLY AND blasphemously VERY BAD

Unigrams and bigrams: A unigram is a single word, and a bigram is a pair of words that appear next to one another.

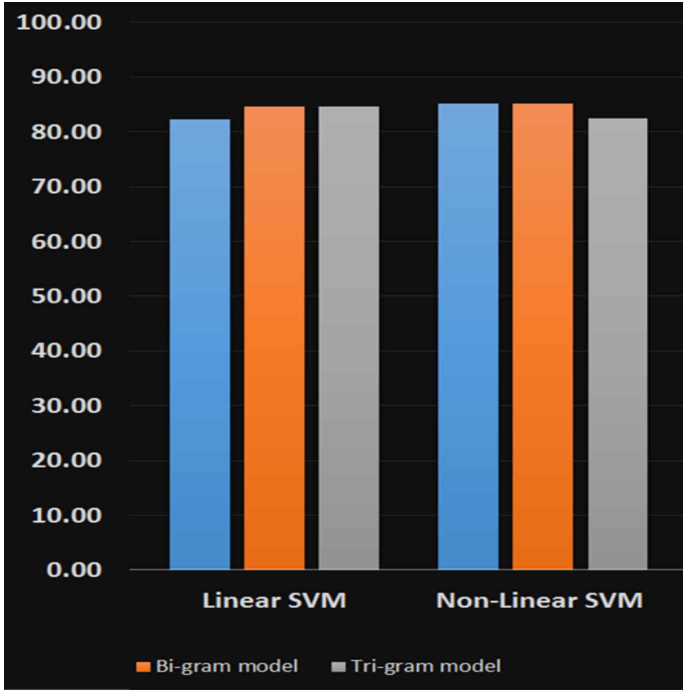
Bigram Vs Unigram

NOT BAD

BAD

Algorithms For Binary Sentiment Classification

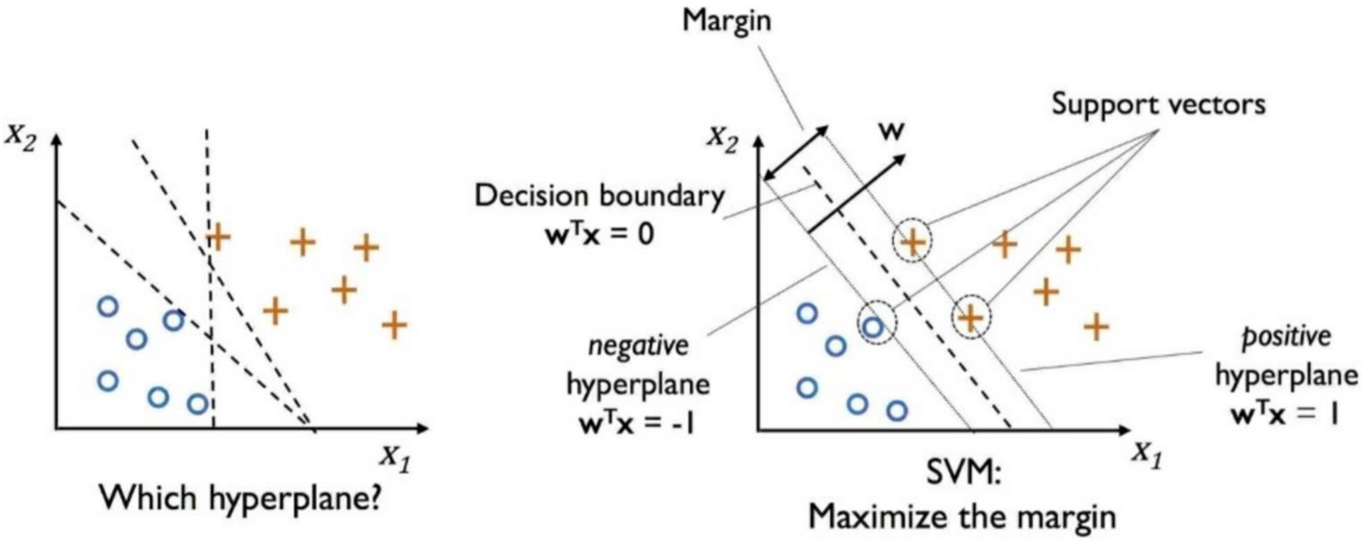
1. Linear Support Vector Machine
2. Gaussian Support Vector Machine



SVM

In SVMs our optimization objective is to maximize the margin.

The margin is defined as the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane, which are the so-called support vectors.

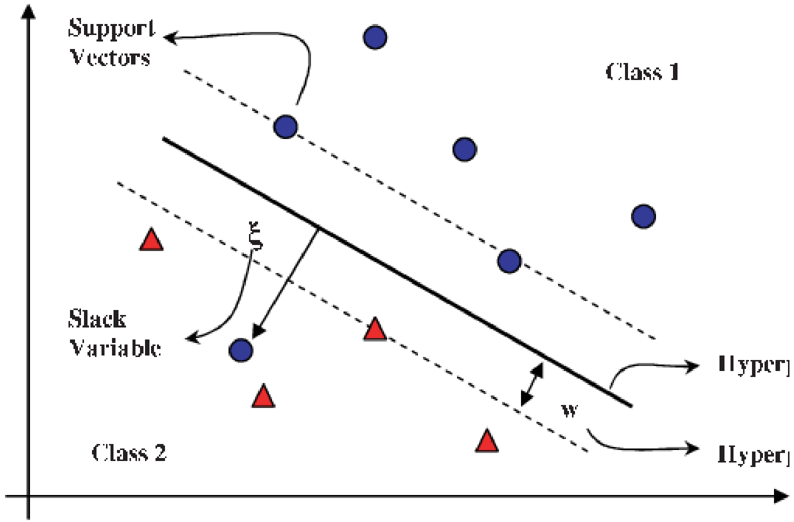


Linear SVM

The support vectors can be represented as :

$$w_0 + \boldsymbol{w}^T \boldsymbol{x}_{pos} = 1 \qquad (1)$$

$$w_0 + \boldsymbol{w}^T \boldsymbol{x}_{neg} = -1 \qquad (2)$$



SVM

Given the constraint that negative and positive should remain in their respective areas:

$$y^{(i)}\left(w_0 + \boldsymbol{w}^T \boldsymbol{x}^{(i)}\right) \geq 1 \ \forall_i$$

Margin can be represented as :

$$\frac{\boldsymbol{w}^T\left(\boldsymbol{x}_{pos} - \boldsymbol{x}_{neg}\right)}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}$$

Where :

$$\|\boldsymbol{w}\| = \sqrt{\sum_{j=1}^{m} w_j^2}$$

Our objective is to maximize the margin:

$$\frac{2}{\|\boldsymbol{w}\|}$$

Solving nonlinear problems using kernel SVM

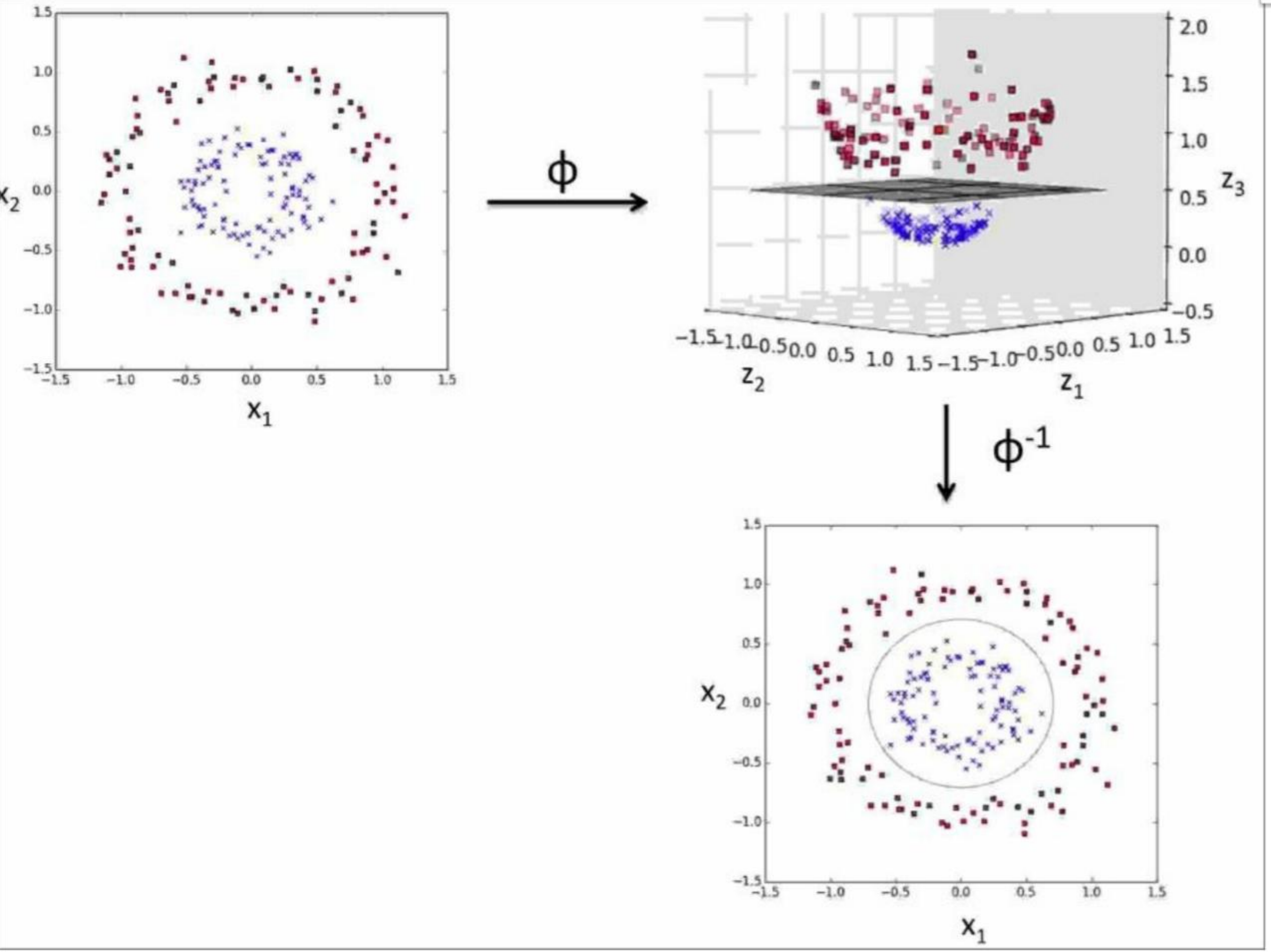
The basic idea behind kernel methods is to deal with Linearly inseparable data

We create nonlinear combinations of the original features and project them onto a higher-dimensional space via a mapping function where it becomes linearly separable.

We can transform a two-dimensional dataset onto a new three-dimensional feature space where the classes become separable:

$$\phi(x_1, x_2) = (z_1, z_2, z_3) = (x_1, x_2, x_1^2 + x_2^2)$$

eature space:



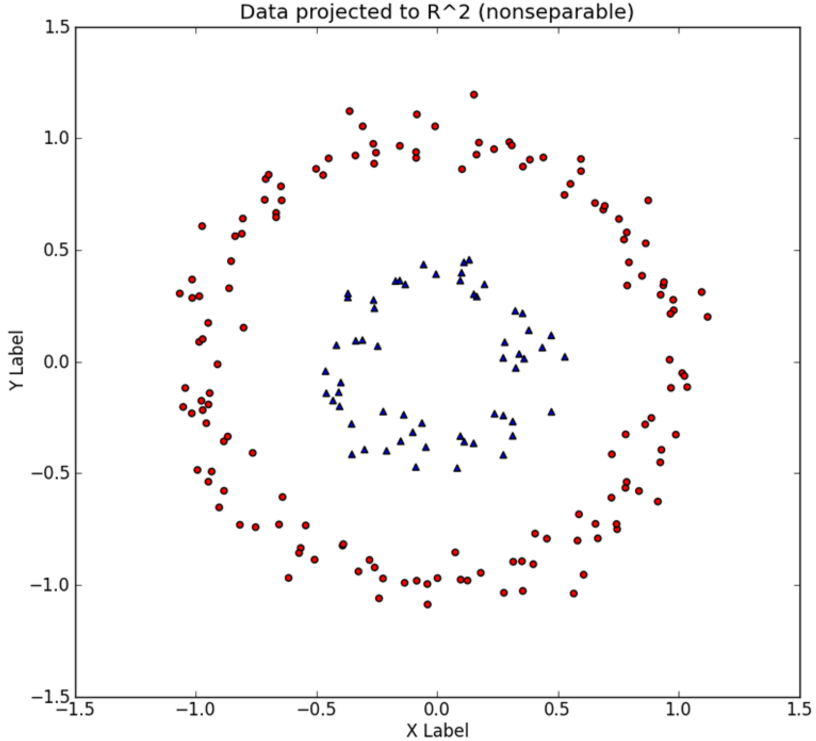
This allows us to separate the two classes shown in the plot via a linear hyperplane.

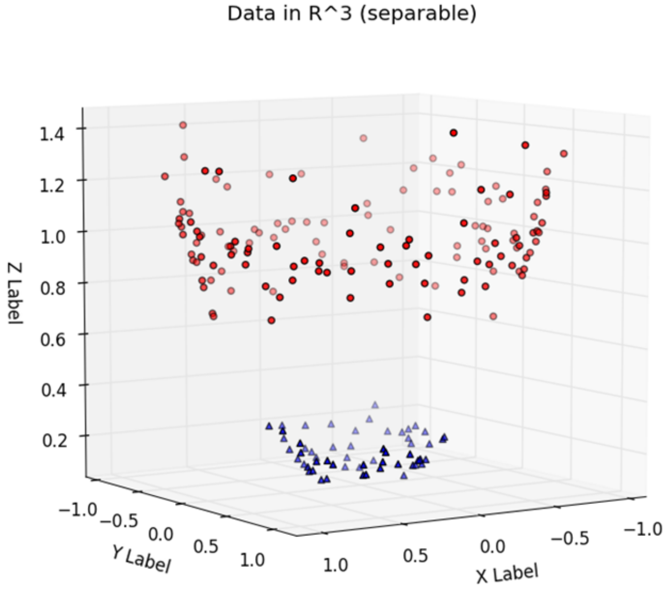
That becomes a nonlinear decision boundary.

If we project it back onto the original feature space.

One of the most widely used kernels is the Gaussian kernel:

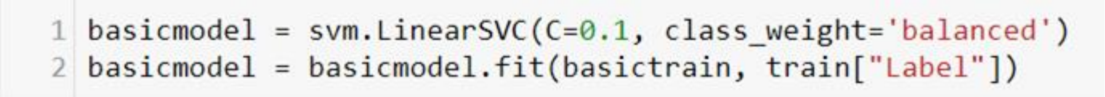
$$\mathcal{K}\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \exp\left(-\frac{\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^2}{2\sigma^2}\right)$$


Data projected to R^2 (nonseparable)

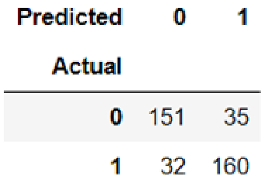

Data in R^3 (separable)

Sentiment Classifier

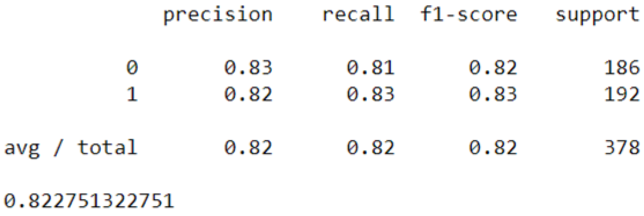
Linear (BiGram)



```python
basicmodel = svm.LinearSVC(C=0.1, class_weight='balanced')
basicmodel = basicmodel.fit(basictrain, train["Label"])
```

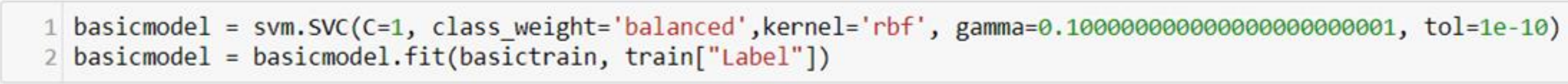


Predicted	0	1
Actual		
0	151	35
1	32	160

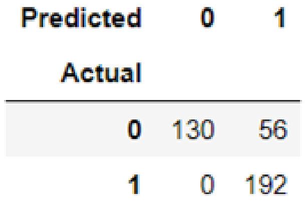


```
             precision    recall  f1-score   support

          0       0.83      0.81      0.82       186
          1       0.82      0.83      0.83       192

avg / total       0.82      0.82      0.82       378

0.822751322751
```

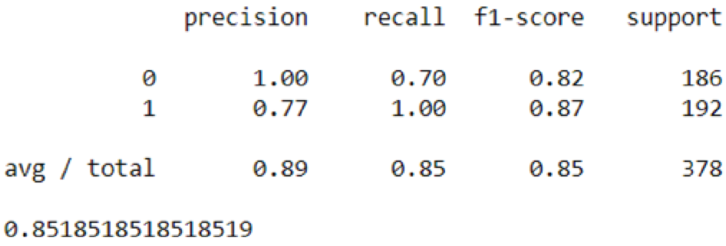
Gaussain : (BiGram)



```python
basicmodel = svm.SVC(C=1, class_weight='balanced',kernel='rbf', gamma=0.1000000000000000000000001, tol=1e-10)
basicmodel = basicmodel.fit(basictrain, train["Label"])
```



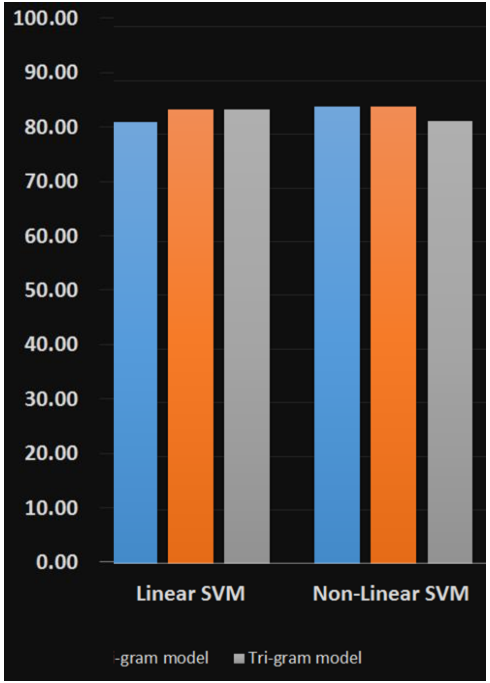
Predicted	0	1
Actual		
0	130	56
1	0	192



```
             precision    recall  f1-score   support

          0       1.00      0.70      0.82       186
          1       0.77      1.00      0.87       192

avg / total       0.89      0.85      0.85       378

0.851851851851851 9
```

Using Multiple Models

- In practice, it is always recommended that you compare the performance of at least a handful of different learning algorithms to select the best model for the particular problem.



Our process is easy

1

Stock data preparation

We will collect daily price of DJIA stock and then find the relative change.

2

Stock trade volume v.s. Tweets volume

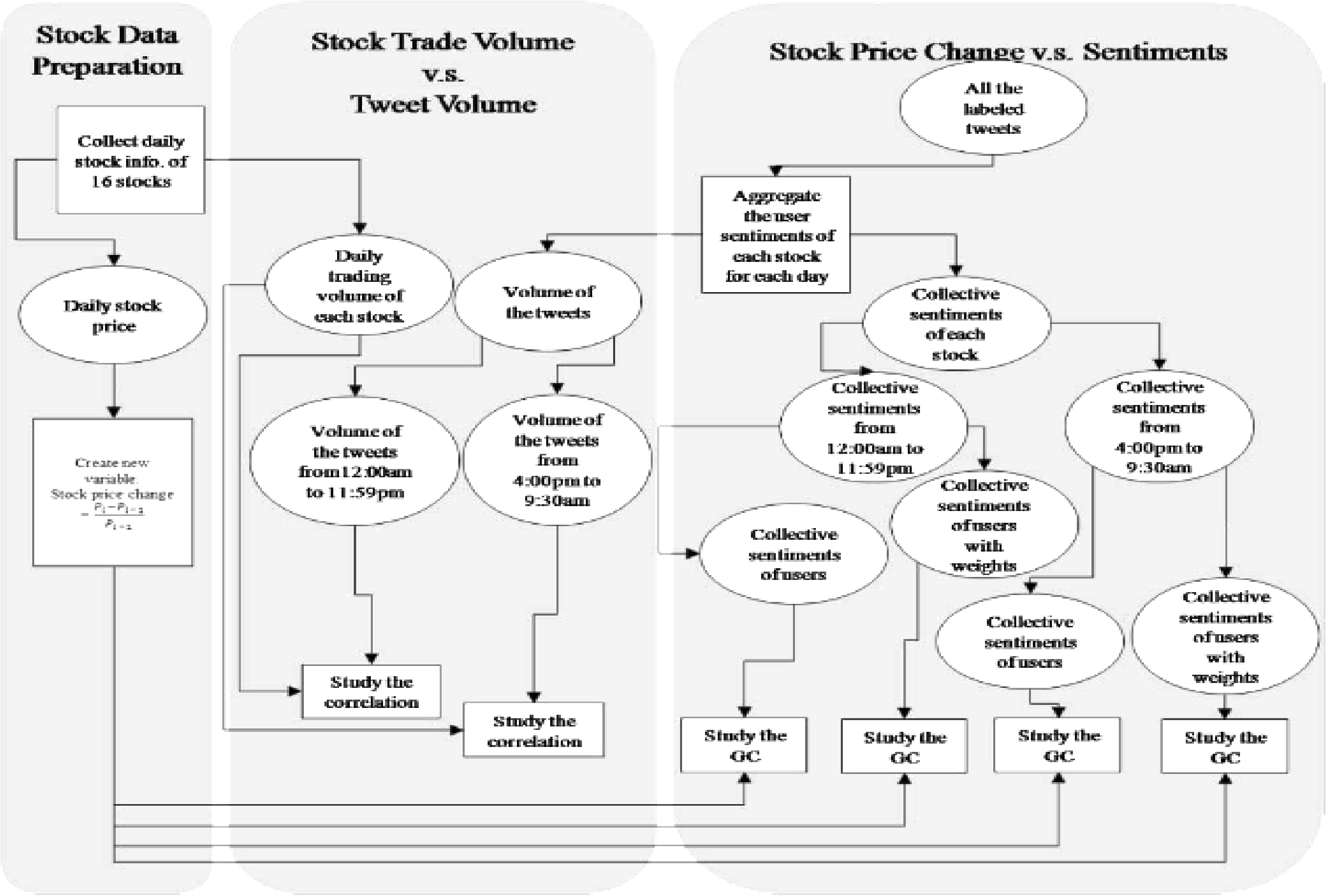
We will study the correlation between DJIA stocks relative change and tweets volume.

3

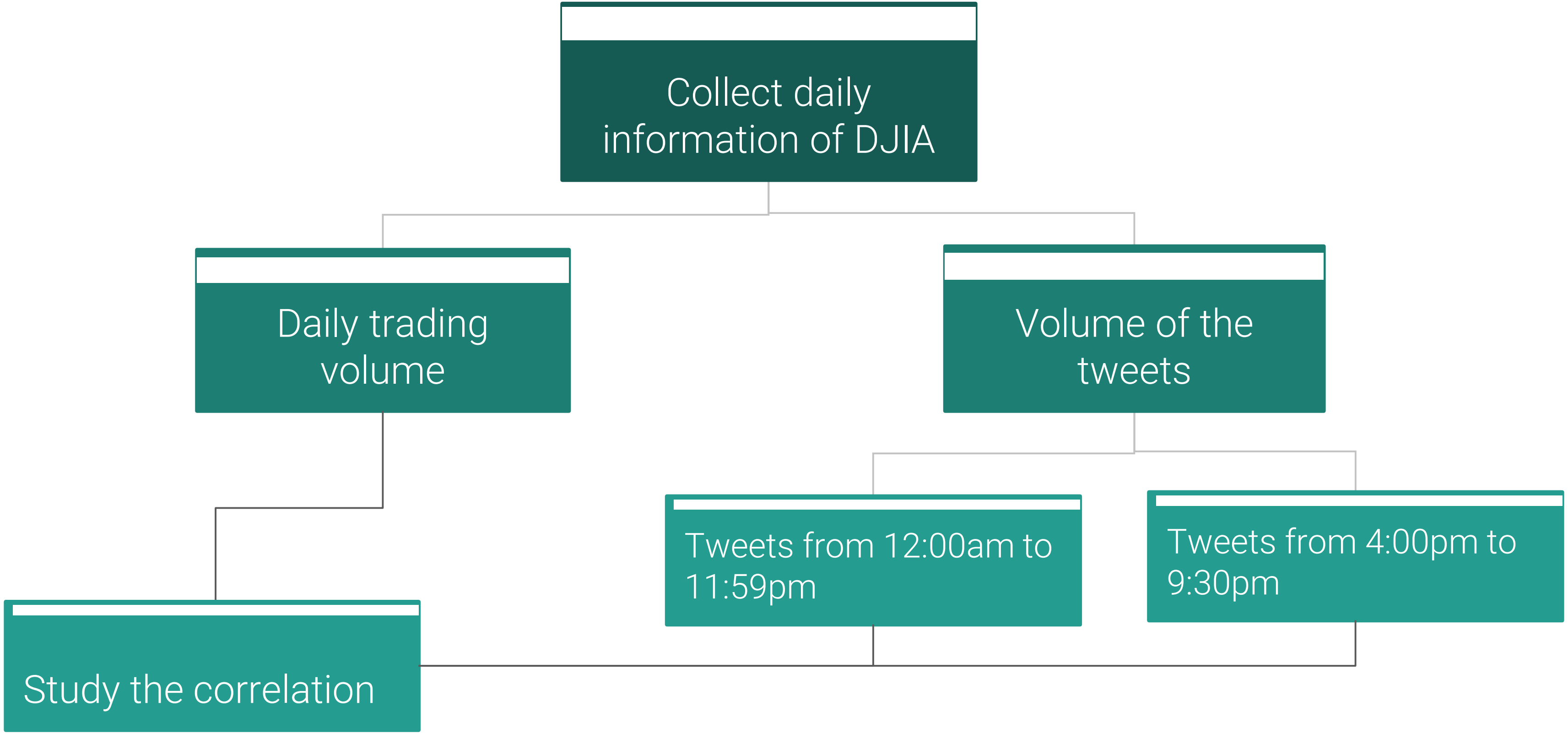
Stock price v.s. Sentiments

Apply Granger Causality test and we will see is there any effect of collective sentiments on stock price.

Reacher schema of Data Analysis


Stock Data Preparation
Collect daily stock info. of 16 stocks
Daily stock price
Create new variable: Stock price change = $\frac{P_t - P_{t-1}}{P_{t-1}}$
Stock Trade Volume v.s. Tweet Volume
Daily trading volume of each stock
Volume of the tweets
Volume of the tweets from 12:00am to 11:59pm
Volume of the tweets from 4:00pm to 9:30am
Study the correlation
Study the correlation
Stock Price Change v.s. Sentiments
All the labeled tweets
Aggregate the user sentiments of each stock for each day
Collective sentiments of each stock
Collective sentiments from 12:00am to 11:59pm
Collective sentiments from 4:00pm to 9:30am
Collective sentiments of users
Collective sentiments of users with weights
Collective sentiments of users
Collective sentiments of users with weights
Study the GC
Study the GC
Study the GC
Study the GC

Stock trade volume v.s. Tweets volume



Collect daily information of DJIA

Daily trading volume

Volume of the tweets

Tweets from 12:00am to 11:59pm

Tweets from 4:00pm to 9:30pm

Study the correlation

Approaches to Determining Collective Sentiments

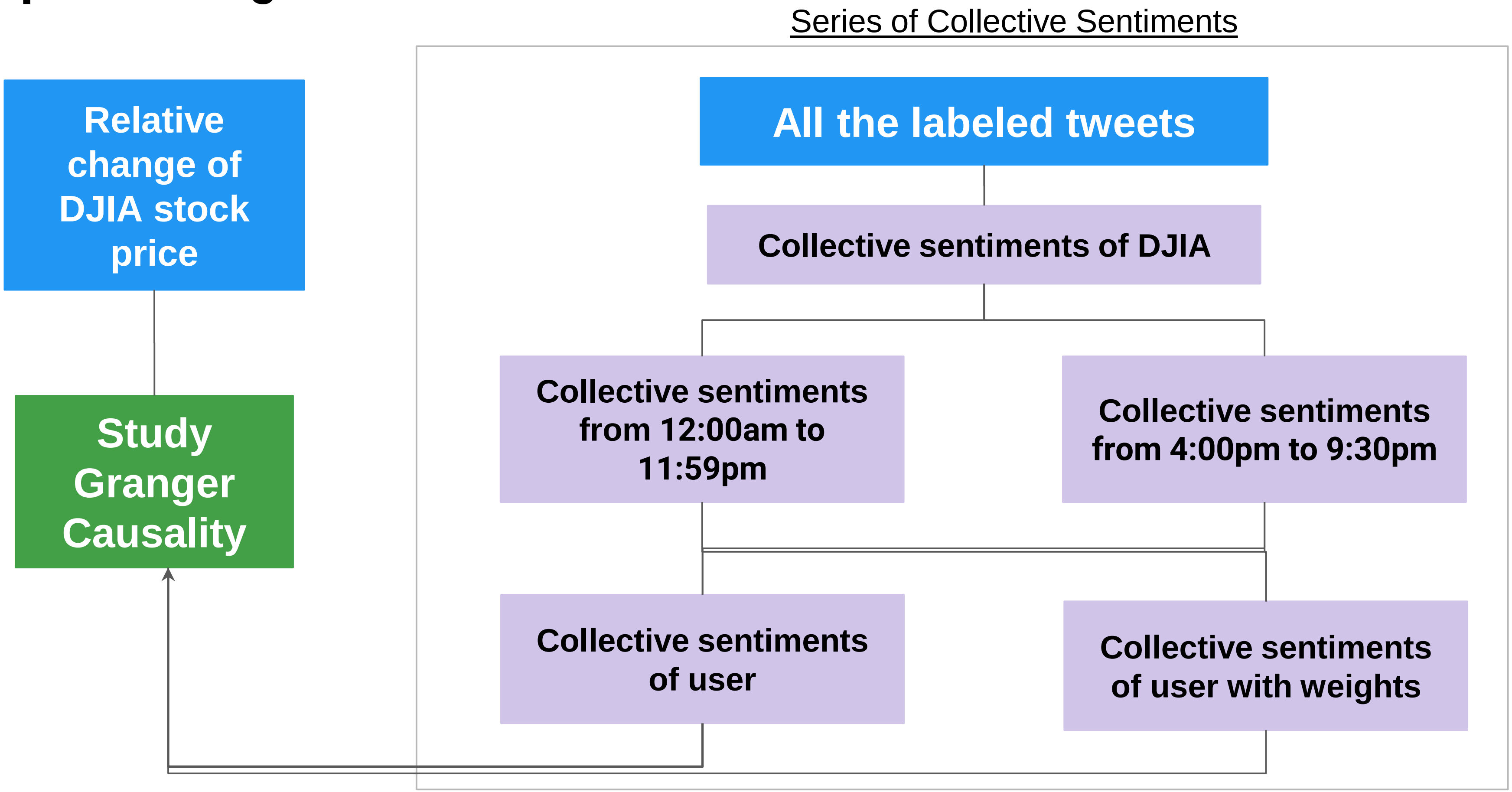
Count of tweets: If a user posted several tweets about a certain stock on the same day, those tweets are aggregated to generate the users unified sentiment, which is positive (1), negative (-1) or neutral (0). This can avoid counting the tweets with the same sentiment from the same user for multiple times.

Number of followers: Different users unified sentiments are assigned with different weights by taking into account their numbers of followers.

Time of publishing: the sentiments of the tweets posted during **12:00am and 11:59pm** on the same day are aggregated.

The sentiments of tweets posted during the period of **4:00pm (the marketing closing time) and the next day 9:30am (the market opening time)** are aggregated.

Stock price change v.s. Sentiments

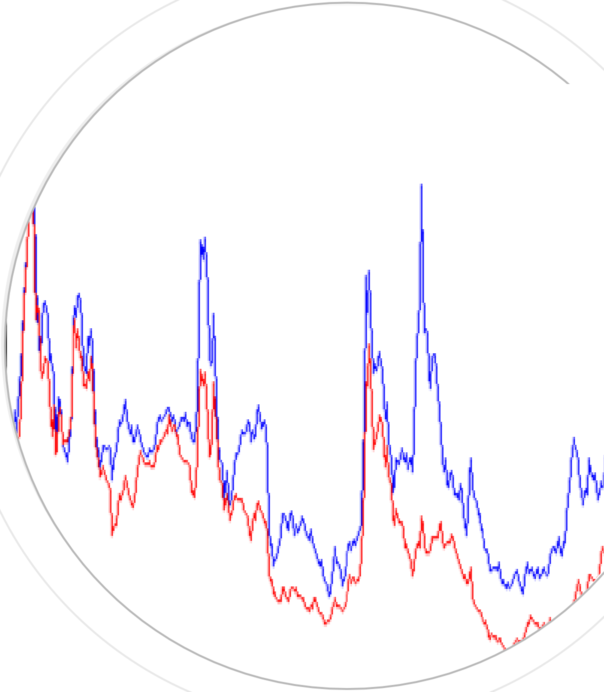

Series of Collective Sentiments
Relative change of DJIA stock price
Study Granger Causality
All the labeled tweets
Collective sentiments of DJIA
Collective sentiments from 12:00am to 11:59pm
Collective sentiments from 4:00pm to 9:30pm
Collective sentiments of user
Collective sentiments of user with weights

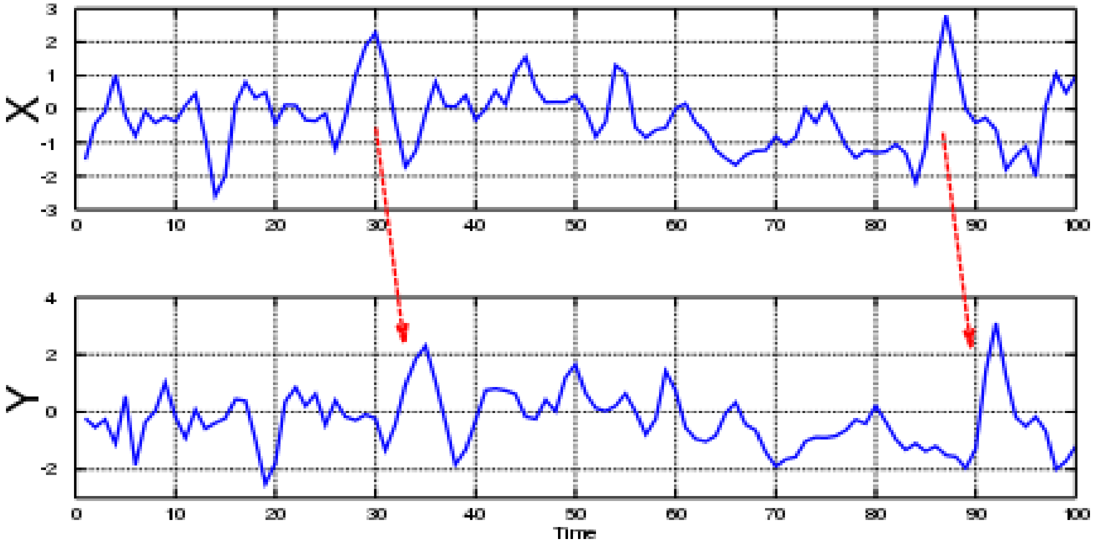
Granger Causality Test

The **Granger causality test** is a statistical hypothesis test for determining whether one time series is useful in forecasting another.

Assumptions:

1. The future can not cause the past. The past cause the present and the future.
2. The cause contains the unique information about the effect not available anywhere else.





When time series *X* Granger-causes time series *Y*,

The patterns in *X* are approximately repeated in *Y* after some time lag (two examples are indicated with arrows).

Thus, past values of *X* can be used for the prediction of future values of *Y*.

In our case:

- If X is Collective Sentiments and Y is the stock price change.

- GC test can demonstrate whether the **collective sentiment** appears to contain information helping in **forecasting the stock price change** of tomorrow that is not in the stock price change of today

Conditions :-

- Each time series must be stationary.
- If not , convert into stationary by ADF test (Augmented Dickey Fuller).

As we know day-to-day collective sentiments on the stock market is random.

AR(1) Process:

$$\mathbf{Yt = a0 + a1\ Yt\text{-}1}$$

The test is conducted at the lag(1) because the volume correlation reveals that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day.

$$\mathbf{\underset{\substack{\swarrow \\ \text{stock price} \\ \text{change}}}{Yt} = a0 + a1\ Yt\text{-}1 + a2\ \underset{\substack{\searrow \\ \text{Collective} \\ \text{Sentiments}}}{Xt\text{-}1}}$$

Hypothesis :-

H0 : a2=0 (Reject the null hypothesis)

H1: a2 ≠ 0 (Accept the null hypothesis)

T-test:-

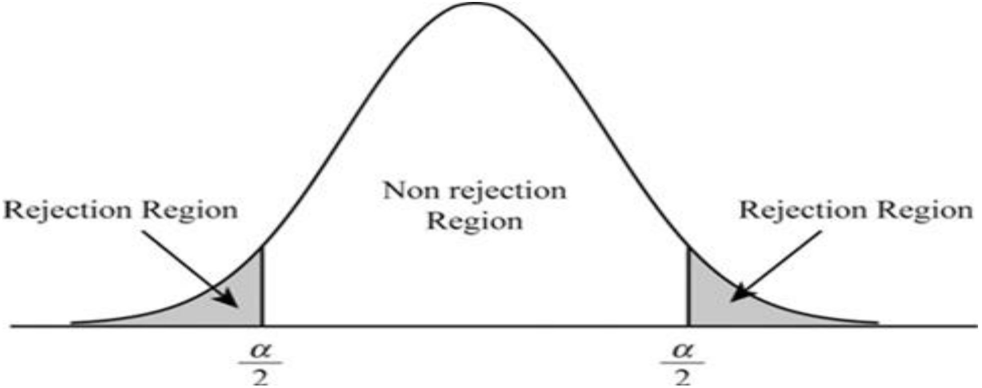
t0 (t critical value) = (X-μ)/ SE

Check the level of significance :-

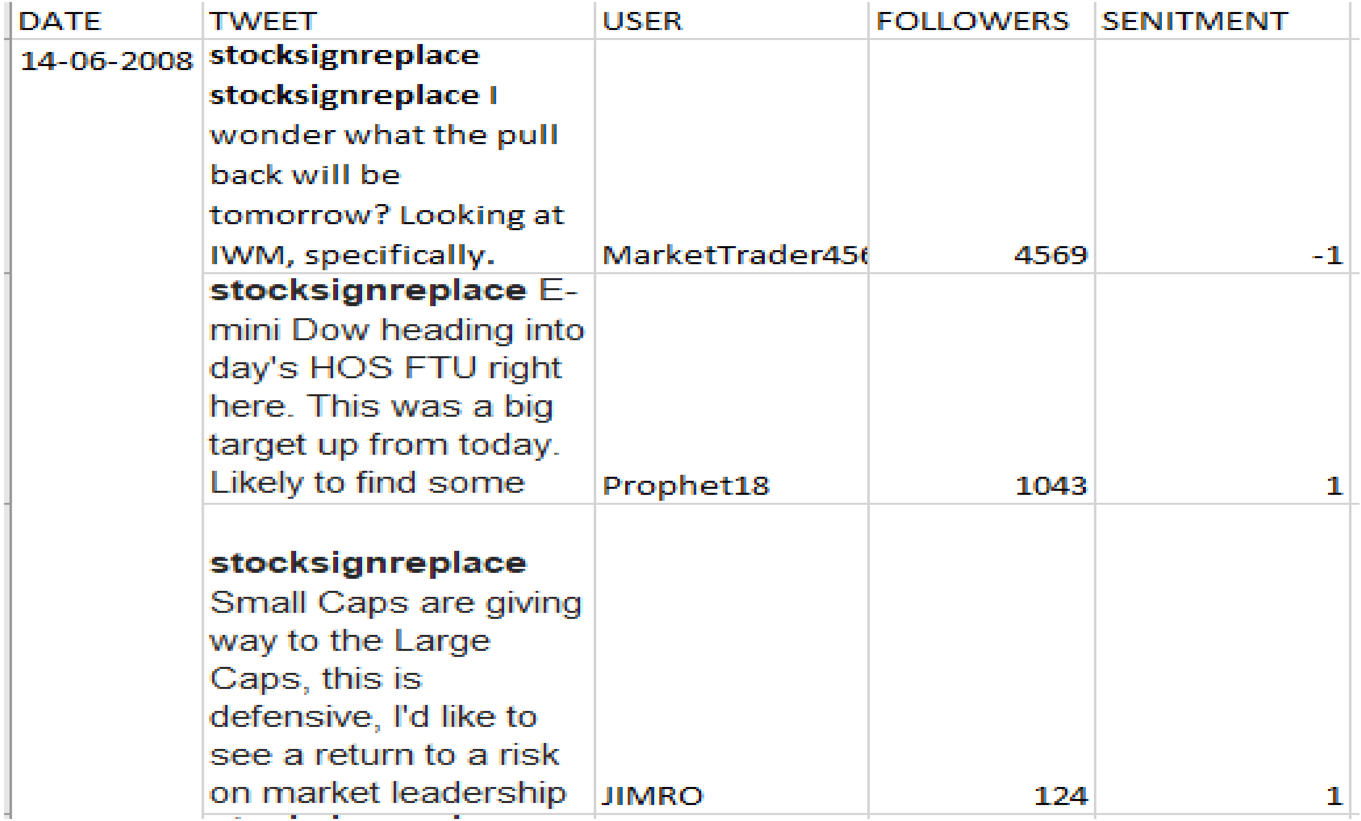
If t0 < tα , n-1 then accept the null hypothesis.

If t0 > tα , n-1 or If t0 < -tα , n-1 then reject the null hypothesis.

where , n = degree of freedom

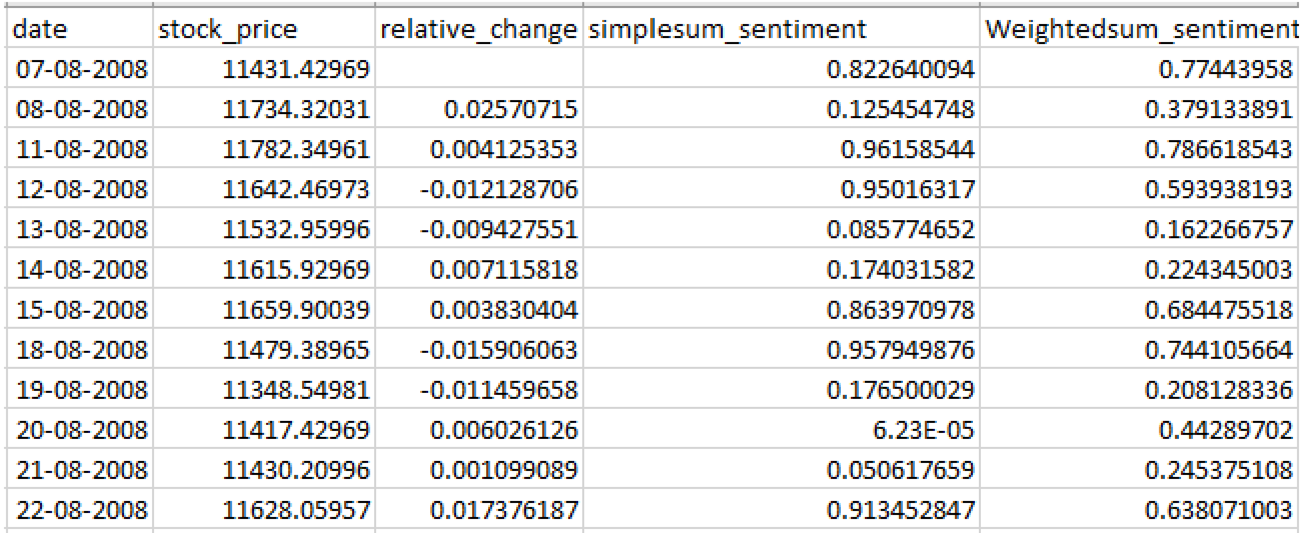


DJIA each user sentiments for each day



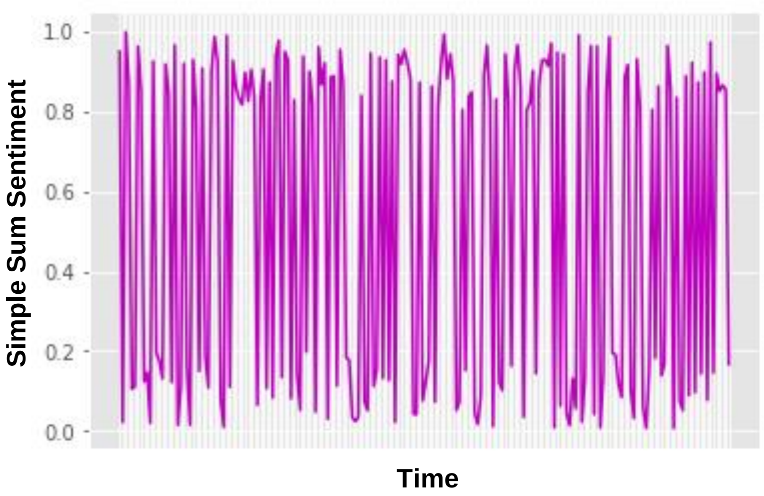
DATE	TWEET	USER	FOLLOWERS	SENITMENT
14-06-2008	**stocksignreplace stocksignreplace** I wonder what the pull back will be tomorrow? Looking at IWM, specifically.	MarketTrader45	4569	-1
	stocksignreplace E-mini Dow heading into day's HOS FTU right here. This was a big target up from today. Likely to find some	Prophet18	1043	1
	stocksignreplace Small Caps are giving way to the Large Caps, this is defensive, I'd like to see a return to a risk on market leadership	JIMRO	124	1

Collective sentiment series

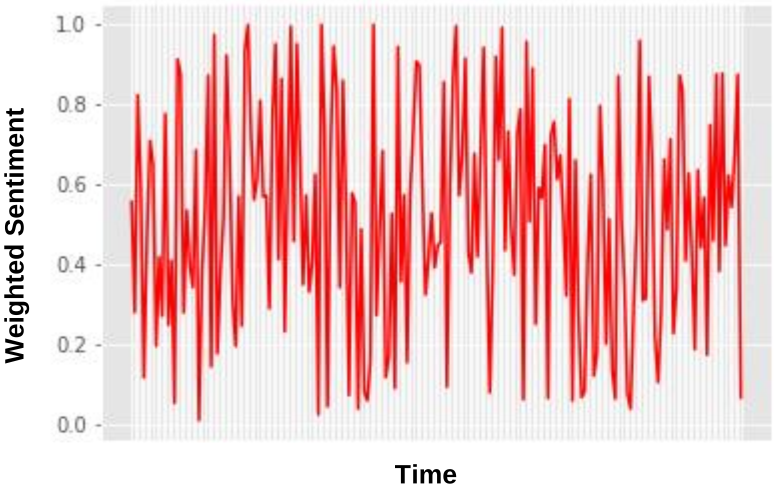


date	stock_price	relative_change	simplesum_sentiment	Weightedsum_sentiment
07-08-2008	11431.42969		0.822640094	0.77443958
08-08-2008	11734.32031	0.02570715	0.125454748	0.379133891
11-08-2008	11782.34961	0.004125353	0.96158544	0.786618543
12-08-2008	11642.46973	-0.012128706	0.95016317	0.593938193
13-08-2008	11532.95996	-0.009427551	0.085774652	0.162266757
14-08-2008	11615.92969	0.007115818	0.174031582	0.224345003
15-08-2008	11659.90039	0.003830404	0.863970978	0.684475518
18-08-2008	11479.38965	-0.015906063	0.957949876	0.744105664
19-08-2008	11348.54981	-0.011459658	0.176500029	0.208128336
20-08-2008	11417.42969	0.006026126	6.23E-05	0.44289702
21-08-2008	11430.20996	0.001099089	0.050617659	0.245375108
22-08-2008	11628.05957	0.017376187	0.913452847	0.638071003

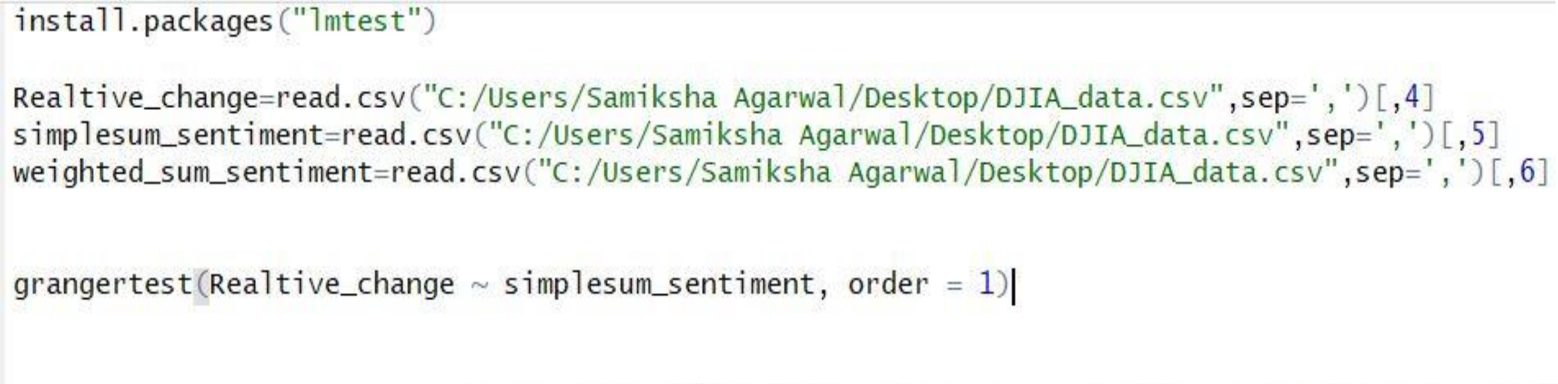
Our Series are Stationary



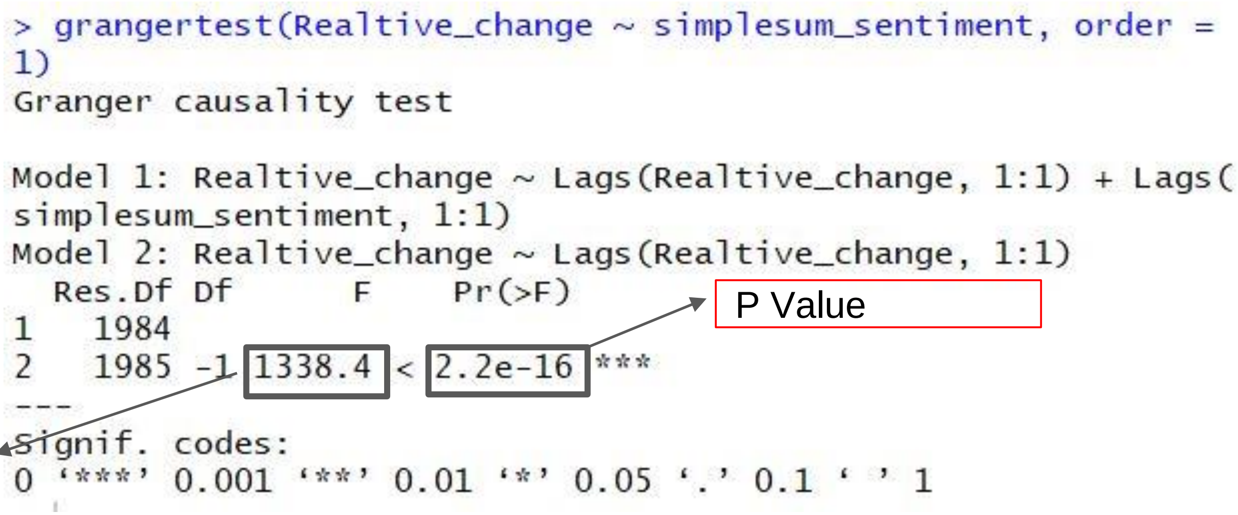
Weighted on the basis of No. of followers of the user



Implementation in R



```
install.packages("lmtest")

Realtive_change=read.csv("C:/Users/Samiksha Agarwal/Desktop/DJIA_data.csv",sep=',')[,4]
simplesum_sentiment=read.csv("C:/Users/Samiksha Agarwal/Desktop/DJIA_data.csv",sep=',')[,5]
weighted_sum_sentiment=read.csv("C:/Users/Samiksha Agarwal/Desktop/DJIA_data.csv",sep=',')[,6]


grangertest(Realtive_change ~ simplesum_sentiment, order = 1)
```



```
> grangertest(Realtive_change ~ simplesum_sentiment, order =
1)
Granger causality test

Model 1: Realtive_change ~ Lags(Realtive_change, 1:1) + Lags(
simplesum_sentiment, 1:1)
Model 2: Realtive_change ~ Lags(Realtive_change, 1:1)
  Res.Df Df      F    Pr(>F)
1   1984
2   1985 -1 1338.4 < 2.2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

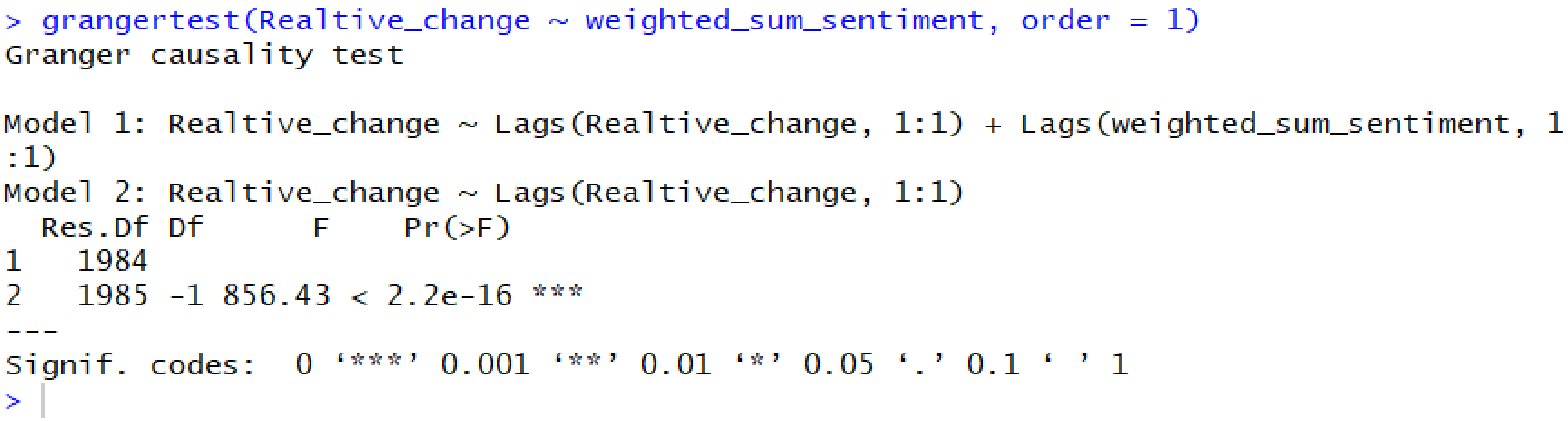
P Value

F critical

RESULT:

Reject the null hypothesis and infer that simplesum_sentiment causes Relative change

➔ grangertest(Realtive_change ~ weighted_sum_sentiment, order = 1)



```
> grangertest(Realtive_change ~ weighted_sum_sentiment, order = 1)
Granger causality test

Model 1: Realtive_change ~ Lags(Realtive_change, 1:1) + Lags(weighted_sum_sentiment, 1
:1)
Model 2: Realtive_change ~ Lags(Realtive_change, 1:1)
  Res.Df Df      F    Pr(>F)
1   1984
2   1985 -1 856.43 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The **collective sentiment** appears to contain information helping in **forecasting the stock price change** of tomorrow .

Authors Result of Granger Causality Test

Stock	Full Day Simple Sum	Afterhours Simple Sum	Full Day Weighted Sum	Afterhours Weighted Sum
$AMZN	Fail for next step (ADF test)	S-> C	X	X
$MSFT	S->C	S->C	Fail	X
$NFLX	X	S->C	X	S->C
$YHOO	X	S->C	X	S->C

With 0.05 level of significance.

Conclusion

- Demonstrated that Support Vector Machine is the best classifier with the overall accuracy rates of 71.84% and 74.3%, respectively, at the two-stage sentiment detection process

- Discovered that users activity on StockTwits overnight significantly positively correlates to the stock trading volume the next business day.

- Determined that simply summed up collective sentiments for afterhours has powerful prediction on the change of stock price for the next day by using Granger Causality test.

Thank You

Open for questions?

- Presented By

Aditya Singh Rathore
(2017MSBDA001)

Samiksha Agarwal
(2017MSBDA008)