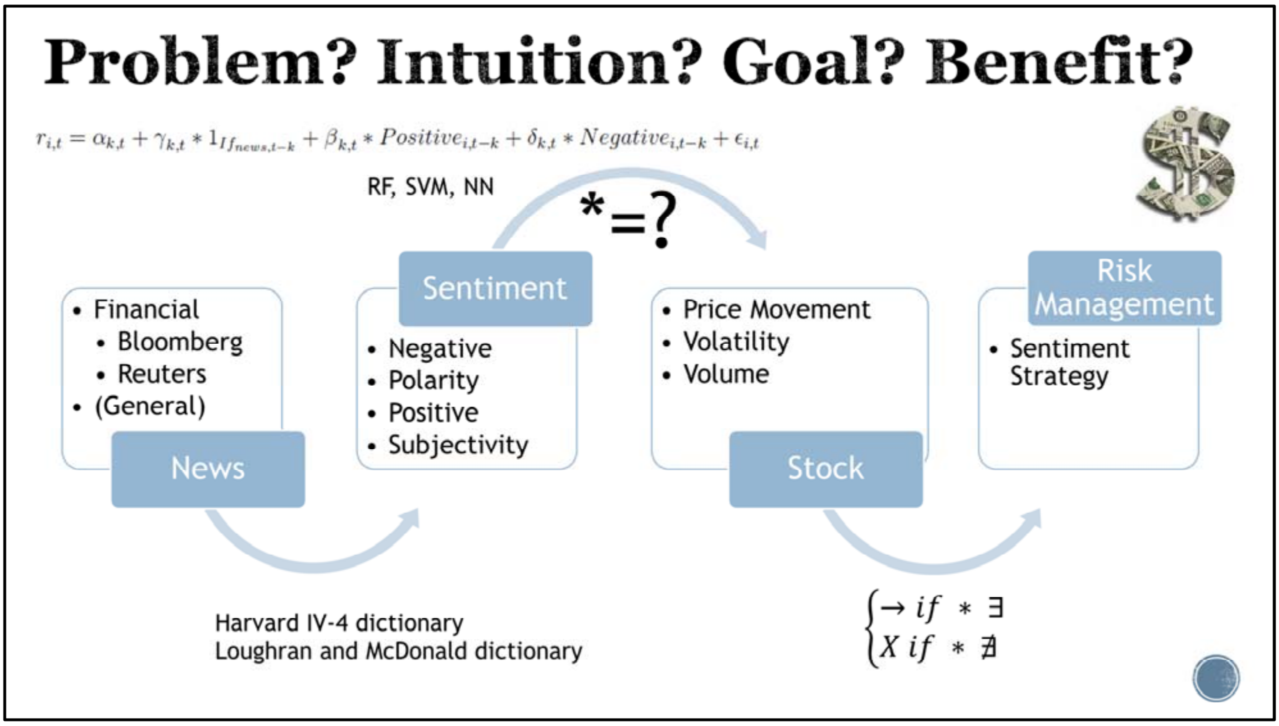
# Using Social Media Data as Early Warning Signals in Risk Management

Kehsin Su, Echo Wang

2018/03/20



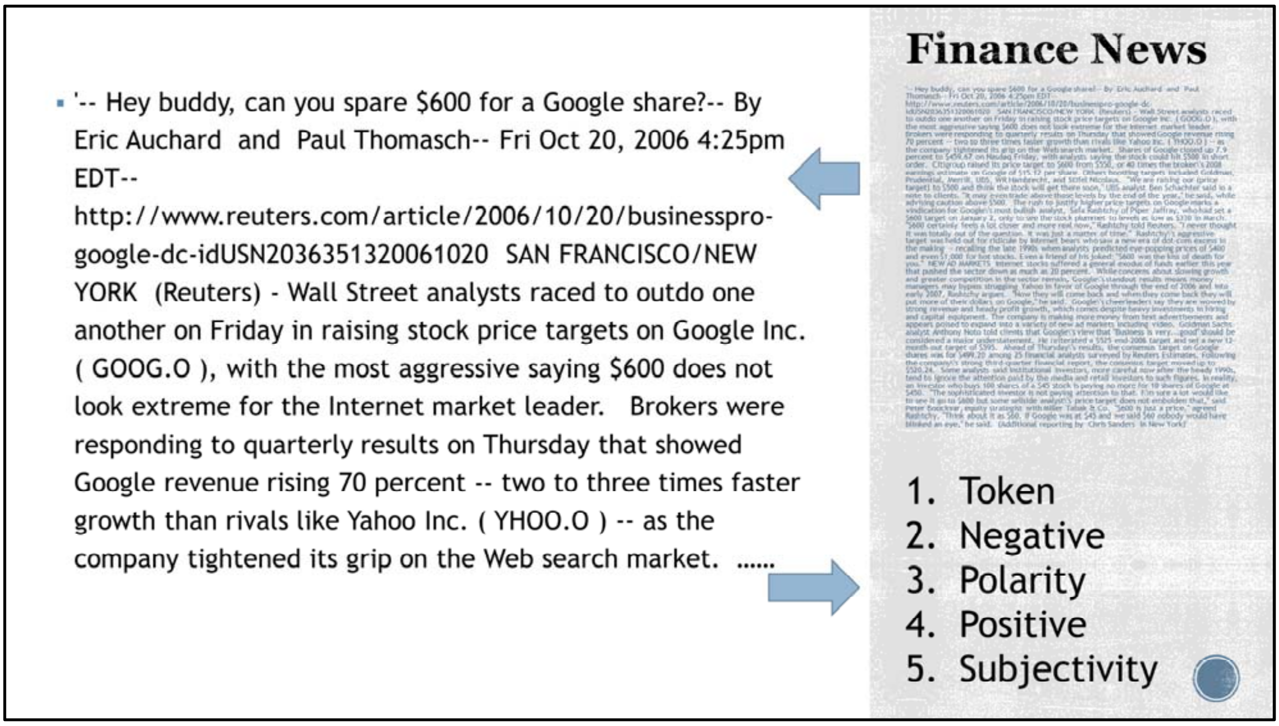

Is it possible to use financial news(or even general news) to predict the stock movement and import a sentiment strategy by our finding?

1.News is major source of market information
2. Big data era, lots of news,
3 development of NLP
4. Risk management is a hot topic after financial crisis
5. traders using technical analysis
6.everyone wants to make money.

R:return(cross-sectional regression)
1 dummy variable for firm,  k= lag,
alpha: no news mean performance,
gamma: return premium: company with published news over performance with those not

Transfer each news into a vector and then combine the vector with the same dates
Binding these vectors to a matrix and use it to predict stock movement

If it works we can import sentiment strategies and even move on general news to see is it works




Data format, plain text and we will extract the information and summary it into a single vector.

- Period: 2006/10/20~2013/11/20
- Number of news per days:

Min: 2

Max: 115

Median: 51

Mean:52.39

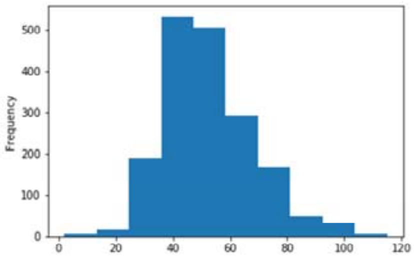
Variance: 240.3

Total: 106,520

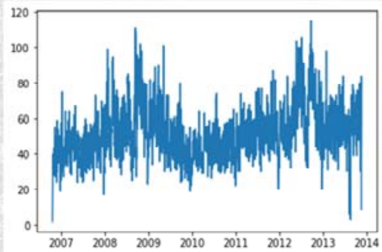
Normality? Not!

KstestResult(statistic=0.9988789227802285, pvalue=0.0)

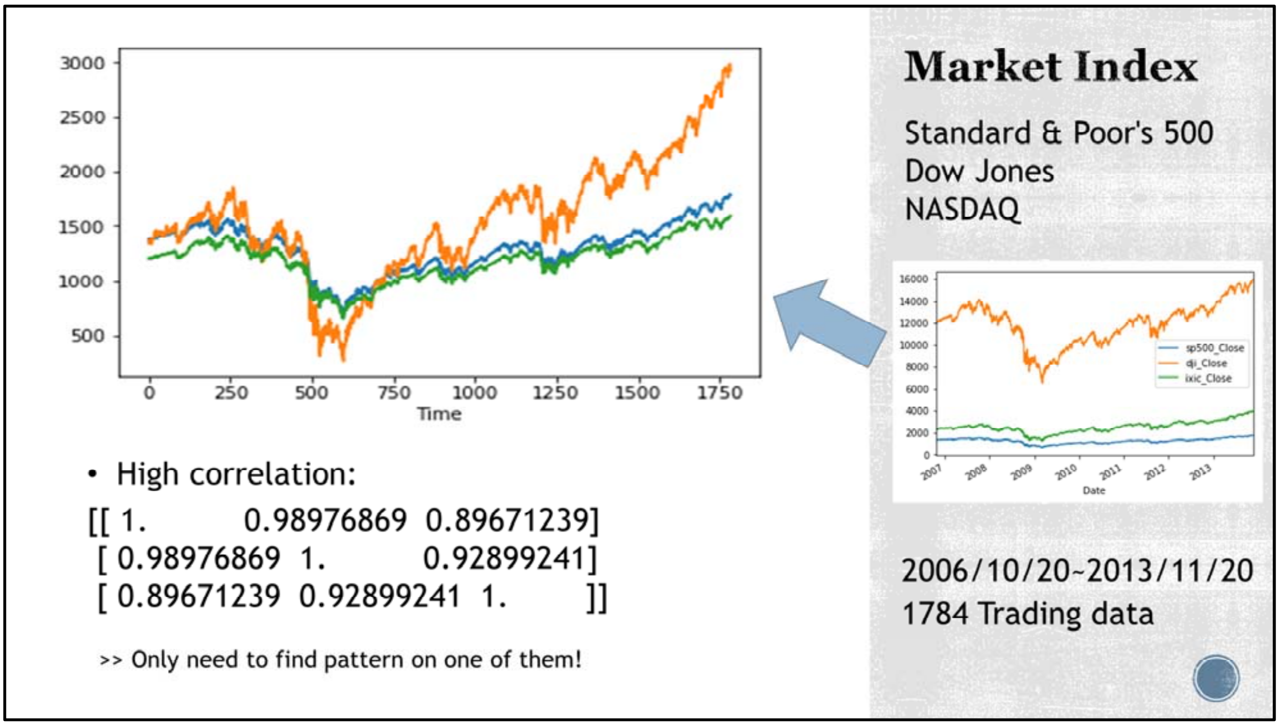
Other 446,510 financial news from Bloomberg will be included!








# one paper mentioned the number of news distributes as normal with heavy tail
# one of the paper mentioned after aggregation, it didn’t matter how many news each day




Since the market index stocks have high correlation, so we only need to find the relationship between sentiment scores and one of them.

# Metrics:

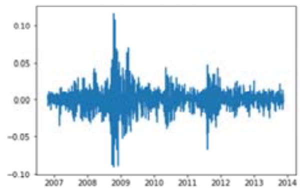
1. Movement
$$\begin{cases} 1 \; if \; Close(t) > Close(t-1) \\ -1 \; if \; Close(t) < Close(t-1) \end{cases}$$
2. Return
$$\frac{Close(t) - Close(t-1)}{Close(t-1)}$$
3. Log Return
Log(Close(t)) - Log(Close(t-1))
4. Open to Close Return
$$\frac{Close(t) - Open(t)}{Open(t)}$$
5. Volume
6. Volatility
High(t) - Low(t)

# Stocks

- Market Index:
  Standard & Poor's 500
  Dow Jones
  NASDAQ
- Individual:
  Google
  Walmart
  Boeing

The open to close return metric is refer from one of the paper. It mentioned that it can remove some season trade and traders' preference. I'll talk the detail part later.

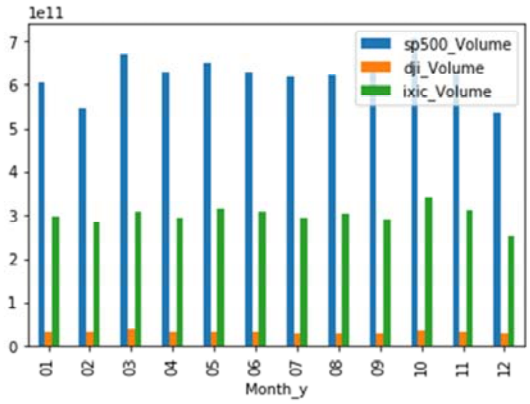
# Returns Metrics

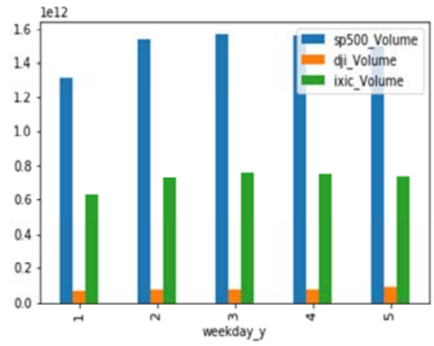


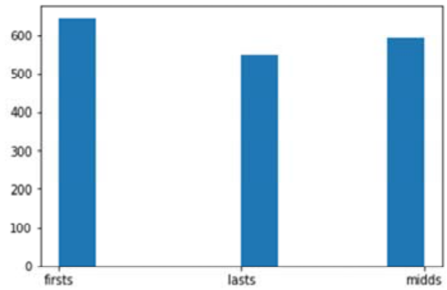

- Return:
  - ADF Statistic: -9.774654
  - p-value: 0.000000
  - Ljung-Box q (LBQ) – lag1
  - X-squared = 25.178
  - p-value = 5.226e-07
- Log return
  - ADF Statistic: -8.938593
  - p-value: 0.000000
  - Ljung-Box q (LBQ) -lag1
  - X-squared = 24.44
  - p-value = 7.667e-07
- Open to close return
  - ADF Statistic: -10.843454
  - p-value: 0.000000
  - Ljung-Box q (LBQ) -lag1
  - X-squared = 24.045
  - p-value = 9.411e-07

All stationary and independent observations

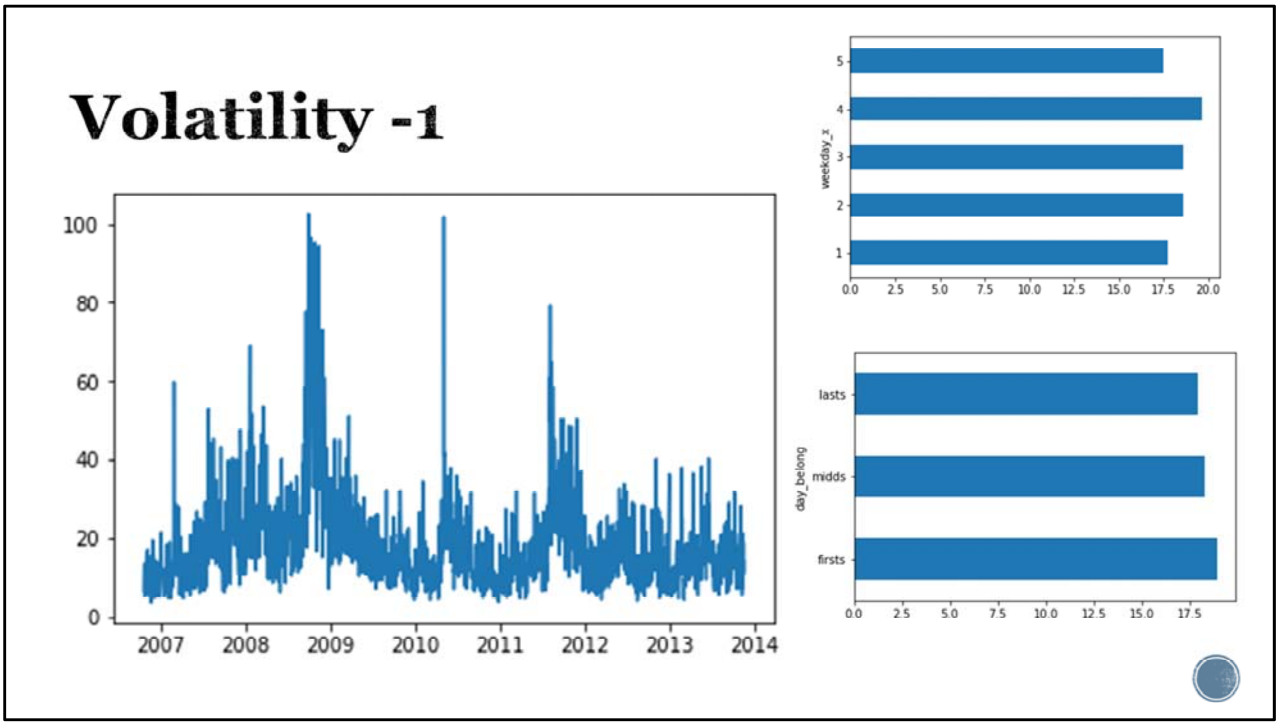
# Volume



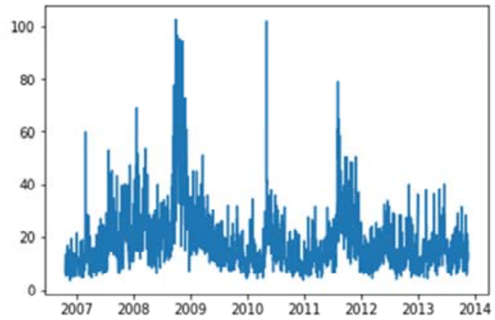



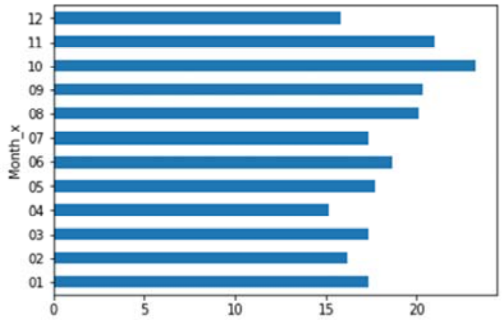

Seasonality(busy trading season:
e.g. Sep to may and less during summer,
more at beginning and end of the week)
Non-trading day gap(more time bring risks-> turn into cash before the gap)
T+0 market mechanism(avoid overnight risk)

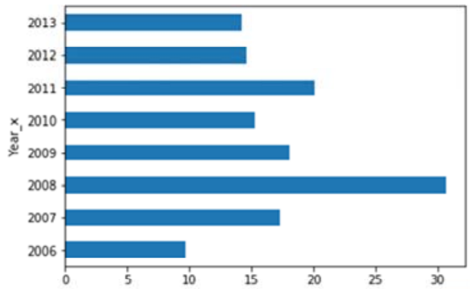



Contradict we what mentioned of the HK market, intuitively, the volatility should max on Friday and at the end of the month.

# Volatility -2




Month and weekday preference
2008 Financial crissis

# Dictionary

Harvard IV-4 categories.

| No. | Description |
| --- | --- |
| 1 | Positive vs. negative |
| 2 | "Osgood" semantic dimensions |
| 3 | Pleasure, pain, virtue and vice |
| 4 | Overstatement and understatement |
| 5 | Language of a particular "institution" |
| 6 | Roles, collectivities, rituals, and forms of interpersonal relations |
| 7 | Ascriptive social |
| 8 | Places, locations and routes |
| 9 | Objects |
| 10 | Communicating |
| 11 | Motivation-related |
| 12 | Process or change |
| 13 | Cognitive orientation |
| 14 | "I" vs. "we" vs. "you" orientation |
| 15 | "Yes", "No", negation and interjections |

10,000 words, 182 sentiment dimensions

Loughran-McDonald categories.

| No. | Description | No. of words |
| --- | --- | --- |
| 1 | Negative words | 2349 |
| 2 | Positive words | 354 |
| 3 | Uncertainty words | 291 |
| 4 | Litigious words | 871 |
| 5 | Modal words strong | 19 |
| 6 | Modal words weak | 27 |

2012 version, 3,911 words

HIV4: http://www.wjh.harvard.edu/~inquirer/
Loughran & McDonald:
https://www3.nd.edu/~mcdonald/Word_Lists.html

LM recently updated in 2015
Manually build the dictionary, so it's trustful

# Sentiment Analysis

According to the Warren Edward Buffett, the stock market could keep climbing for a year or more. It's the right time to buy the stock.

- Harvard IV-4 dictionary
- {'Negative': 0,
- 'Polarity': 0.9999996666667778,
- 'Positive': 3,
- 'Subjectivity': 0.33333329629630043}

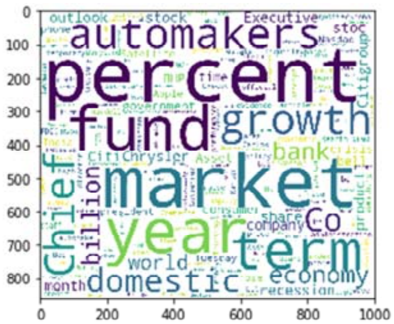
- Loughran and McDonald dictionary
- {'Negative': 0,
- 'Polarity': 0.0,
- 'Positive': 0,
- 'Subjectivity': 0.0}

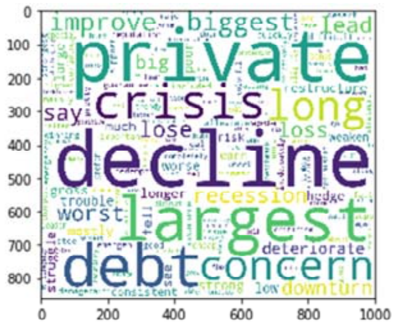
Token:
['accord', 'buffett', 'market', 'keep', 'climb', 's', 'right', 'time', 'buy']

Robert Plutchik's "Wheel of Emotions"

An example of words, the for LM is that it contains less words, so it will be quite unreliable if the words didn't contained in the dictionary. HIV contains more words.

# 2008/11/20 Financial Crisis




Able to capture the general situation of the market

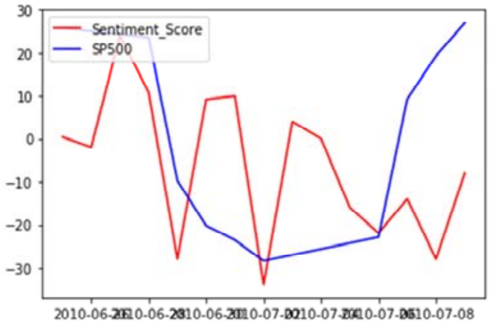
# Lowest DJIA(13-year low- 2009/03/06 )



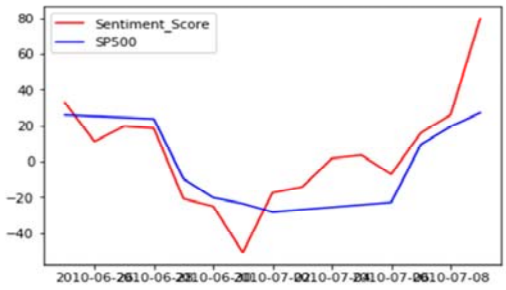
AIG suffers $62 billion loss, bailout revamped - Mar. 2, 2009

# July 2 2010

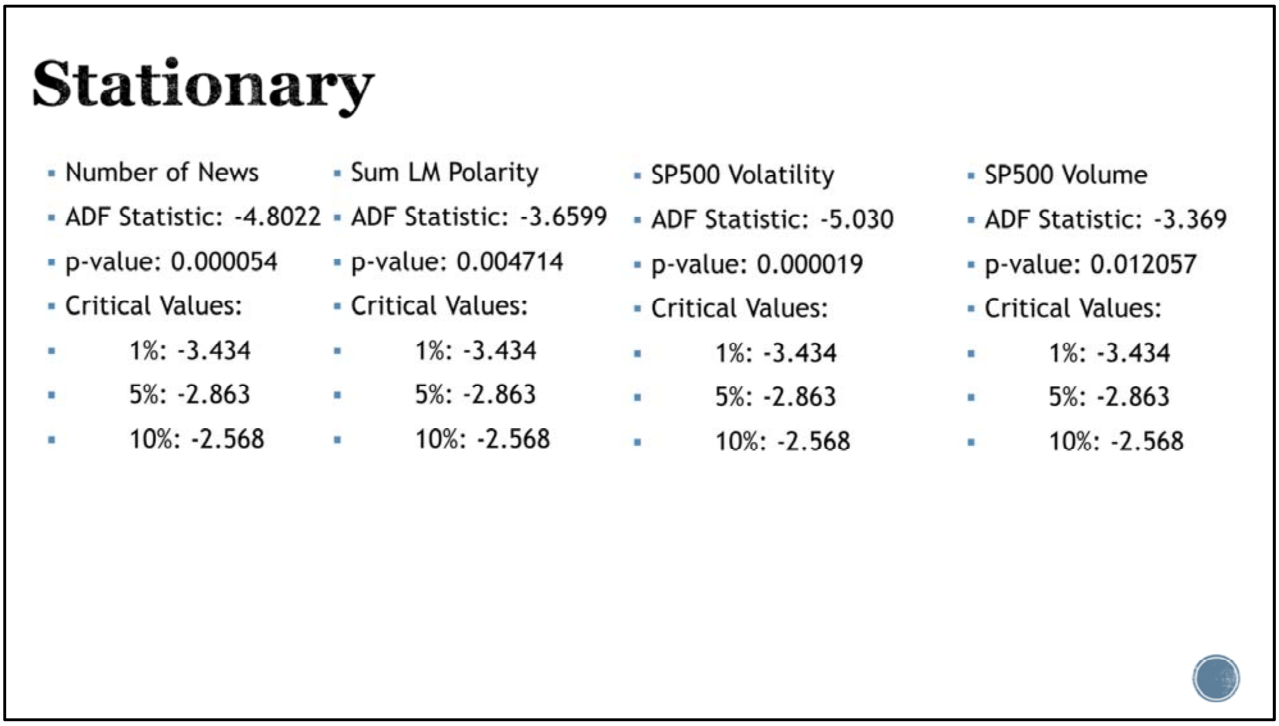
683 news




5885 News




More news more accuracy!




stationarity or trend-stationarity

# Correlation

## Return -HIV4

| | Log Return | Open to close | Sum LM Negative | Sum LM Polarity | Sum LM Positive | Sum LM Subjectivity |
|---|---|---|---|---|---|---|
| Log return | 100.0% | 99.4% | -7.0% | 3.0% | -3.0% | -5.7% |
| Open to close | 99.4% | 100.0% | -6.3% | 2.5% | -2.7% | -5.1% |
| Sum LM Negative | -7.0% | -6.3% | 100.0% | 32.2% | 95.0% | 88.8% |
| Sum LM Polarity | 3.0% | 2.5% | 32.2% | 100.0% | 53.9% | 62.6% |
| Sum LM Positive | -3.0% | -2.7% | 95.0% | 53.9% | 100.0% | 91.4% |
| Sum LM Subjectivity | -5.7% | -5.1% | 88.8% | 62.6% | 91.4% | 100.0% |

# Correction

## Return -LM

| | Log Return | Open to close | Sum LM Negative | Sum LM Polarity | Sum LM Positive | Sum LM Subjectivity |
|---|---|---|---|---|---|---|
| Log return | 100.0% | 99.4% | -8.7% | 10.3% | 1.0% | -7.6% |
| Open to close | 99.4% | 100.0% | -7.8% | 9.2% | 1.2% | -6.7% |
| Sum LM Negative | -8.7% | -7.8% | 100.0% | -86.6% | 80.0% | 92.6% |
| Sum LM Polarity | 10.3% | 9.2% | -86.6% | 100.0% | -52.8% | -87.5% |
| Sum LM Positive | 1.0% | 1.2% | 80.0% | -52.8% | 100.0% | 80.3% |
| Sum LM Subjectivity | -7.6% | -6.7% | 92.6% | -87.5% | 80.3% | 100.0% |

# Correlation

## Volume & Volatility -HIV4

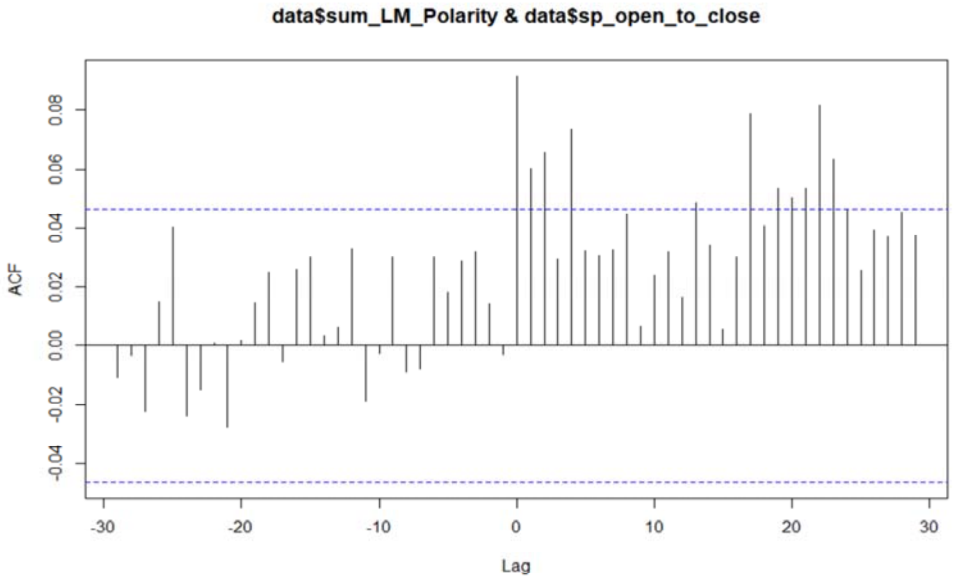
| | Count | Sum HIV Negative | Sum HIV Polarity | Sum HIV Positive | Sum HIV Subjectivity | Volume | Volatility |
|---|---|---|---|---|---|---|---|
| Count | 100.0% | 83.5% | 68.5% | 87.3% | 98.3% | 24.4% | 27.4% |
| Sum HIV Negative | 83.5% | 100.0% | 32.2% | 95.0% | 88.8% | 31.2% | 28.2% |
| Sum HIV Polarity | 68.5% | 32.2% | 100.0% | 53.9% | 62.6% | 1.6% | 7.7% |
| Sum HIV Positive | 87.3% | 95.0% | 53.9% | 100.0% | 91.4% | 26.3% | 22.9% |
| Sum HIV Subjectivity | 98.3% | 88.8% | 62.6% | 91.4% | 100.0% | 30.0% | 30.8% |
| Volume | 24.4% | 31.2% | 1.6% | 26.3% | 30.0% | 100.0% | 53.9% |
| Volatility | 27.4% | 28.2% | 7.7% | 22.9% | 30.8% | 53.9% | 100.0% |

# Correlation

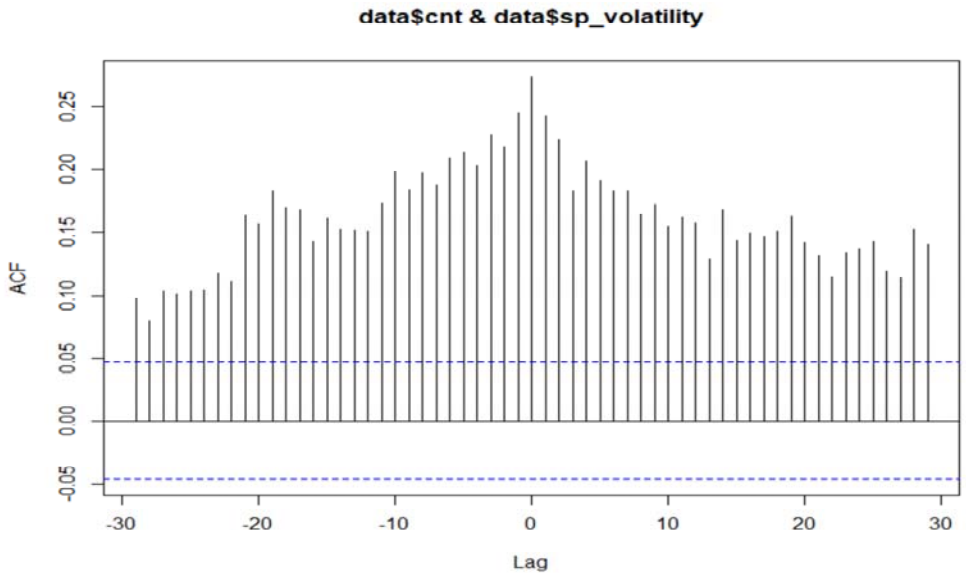
## Volume & Volatility -LM (better)

| | Count | Sum LM Negative | Sum LM Polarity | Sum LM Positive | Sum LM Subjectivity | Volume | Volatility |
|---|---|---|---|---|---|---|---|
| Count | 100.0% | 81.6% | -80.8% | 77.9% | 93.6% | 24.4% | 27.4% |
| Sum LM Negative | 81.6% | 100.0% | -86.6% | 80.0% | 92.6% | 34.1% | 32.4% |
| Sum LM Polarity | -80.8% | -86.6% | 100.0% | -52.8% | -87.5% | -34.6% | -37.0% |
| Sum LM Positive | 77.9% | 80.0% | -52.8% | 100.0% | 80.3% | 25.8% | 16.5% |
| Sum LM Subjectivity | 93.6% | 92.6% | -87.5% | 80.3% | 100.0% | 34.7% | 35.0% |
| Volume | 24.4% | 34.1% | -34.6% | 25.8% | 34.7% | 100.0% | 53.9% |
| Volatility | 27.4% | 32.4% | -37.0% | 16.5% | 35.0% | 53.9% | 100.0% |

# Cross correlation



# Cross correlation



# Movement Prediction

1~1781 training
last 178 testing

## Random Forest

- AUC Train = 0.6919403 (HIV4)
- AUC Test = 0.5546388 (HIV4)

| | Down | Up | Error |
|---|---|---|---|
| Down | 411 | 396 | 0.491 |
| Up | 270 | 704 | 0.277 |

- Matthews Correlation Coefficient(MCC) = 1.901302e-09
- AUC Train = 0.6653944 (LM)
- AUC Test = 0.5322895 (LM)

## SVM

0.6847286

0.5326143

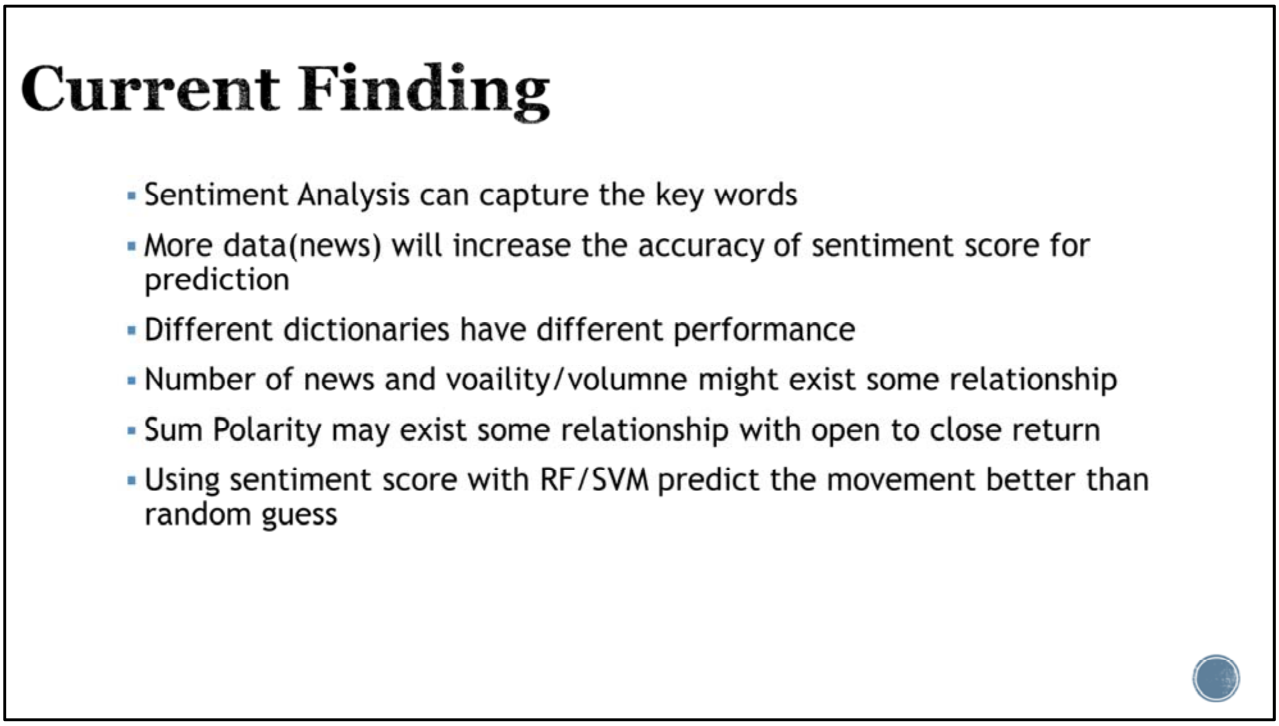
| | Down | Up | Error |
|---|---|---|---|
| Down | 184 | 72 | 0.281 |
| Up | 623 | 902 | 0.409 |

2.725109e-09

0.6720253

0.5461279

Move ~ day_belong+ weekday+ HIV_PN_diff+ sum_Positive+ sum_Polarity+ sum_Subjectivity+ sum_Negative+ var_Negative+ var_Positive+ var_Polarity+ var_Subjectivity + cnt



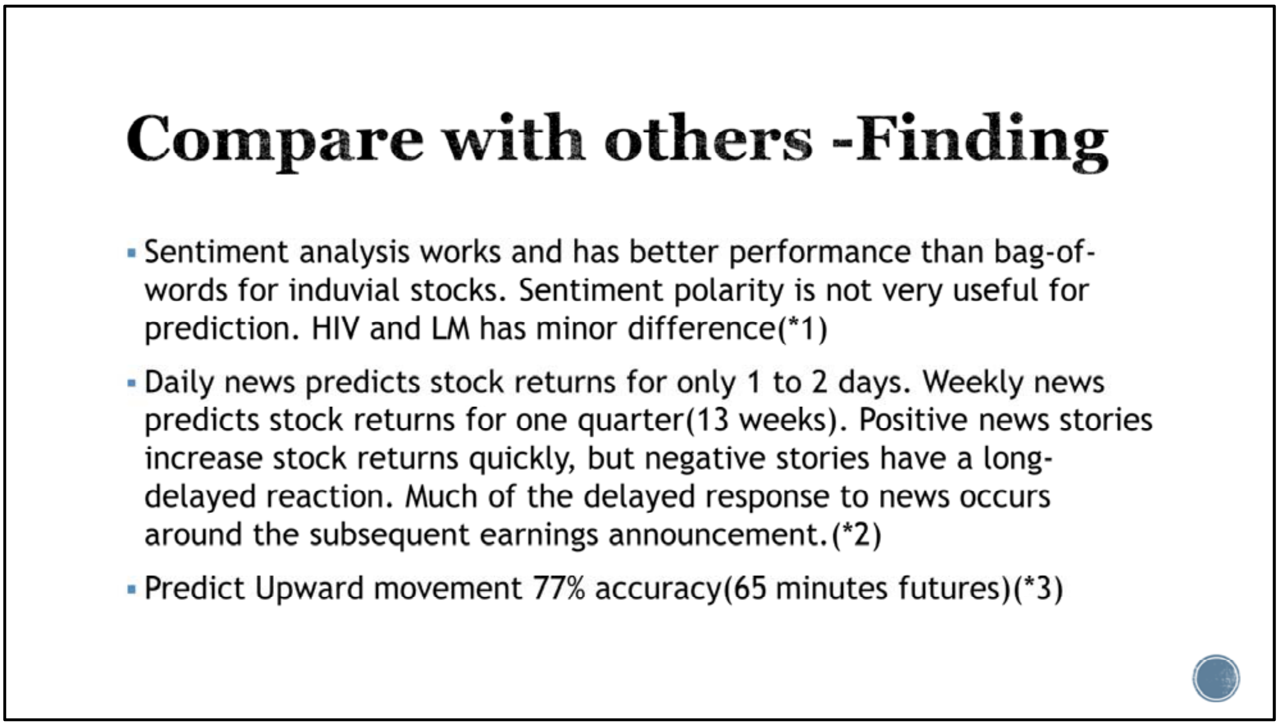

Even though LM scores have higher


1. Financial market are different
2. We have 554,915 articles without tag from 2006/11/20 to 2013/11/20 (7 years )
(use 106, 520 in the analysis part this time)
It has better quality and contains tags
3. Sp 100

# Compare with others -Approach

- Harvard psychological dictionary and Loughran-McDonald financial sentiment dictionary,  senticnet 0.3 API, RBF kernel SVM, grid search for hyper-parameters(*1)
- Thomson-Reuters neural network, cross-sectional regression, Category by firms(*2)
- Stemmed bag-of-words, TFIDF, Hierarchical agglomerative clustering algorithm with Dynamic Time Warping(DTW)/ weighted distance, Recurrent neural network (RNN)- LTSM(*3)

LTSM- long short term memory
*1 doubt check the sentiment scores (for ourselves, we can includes more dictionaries or sample partial data to do SA with free online API)
*3 upward
*3 DTW: measuring similarity between two temporal sequences. Used to calculates an optimal match between two given sequences (e.g. time series) with certain restrictions.




even for stocks with only one news event per week, it is important to gauge relative news sentiment by examining news over longer horizon rather than just one day of stories.

# Limitation

- Time consuming for large data set processing(only use 20% of the data and 50% of are still running sentiment analysis)
- More news still required
- More powerful textual analysis methods/tools required
- Hard to find the better dictionaries or expand the current ones
- Too much noise in the news
- Finance market is always changing and evolving

# Next Step

- Include more data from Bloomberg
- Working on individual stocks to see it perform better or not
- Add dictionary-Henry's Financial dictionary (Henry 2008) and mix scores from different dictionaries into the model
- Model improvement(Neutral Network, Ensemble Learning)
- Aggregate news by week/month to make better prediction
- Change the grouping criteria and set new benchmarks
- Look up more relevant papers