Predicting stock price using sentiment analysis combining Twitter, search engine and investor intelligence data

Master Thesis Defense

Rui Wu

Advisor: Dr. Shang

Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

Motivation

- Stock market is an integral part of global economy.
  - United States has a market capitalization of $18.668 trillion (2012).
- It has a profound economic impact on the economy and everyday people.
  - The stock market crash of 1929 was a key factor in causing the great depression of the 1930s

Demand

A good prediction model for stock market forecasting is always highly desirable and would of wider interest.

- Lots of studies and researches
- Yield significant profit

Social Media Power

Very early indicators can be extracted from online social media to predict changes in various economic and commercial indicators.

Twitter, which is now one of the most popular microblogging services, has been extensively used for real time sentiment tracking and public mood modeling.

Goal

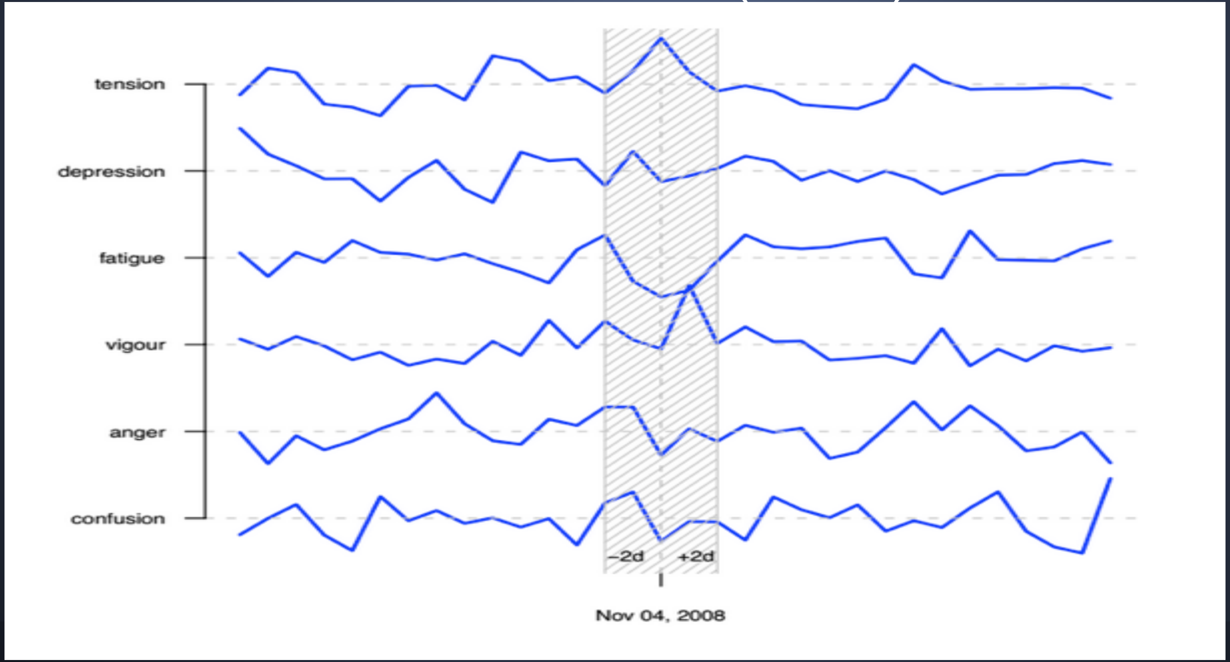
Building a stock price prediction system combining Twitter, Search Engine and Investor Intelligence data.

Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

1. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. J. Bollen (2009)
    - Extracted six dimensions of mood(tension, depression, anger, vigor, fatigue, confusion) using an extended version of the Profile of Mood States
    - Result:
        - 2008 President Selection match ‘Tension’
        - Thanksgiving match ‘vigor’

1. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. J. Bollen (2009)



2. Predicting financial markets: Comparing survey, news, twitter and search engine data. J. Bollen (2010)

- Survey a range of online data sets: Twitter sentiment, news headlines, investor survey, Google search queries.
- Correlations
  - Daily: DJIA - Trade Volume: 0.88
  - Weekly: Twitter Volume - GIS: 0.61

3. Analyzing stock market movements using twitter sentiment analysis.

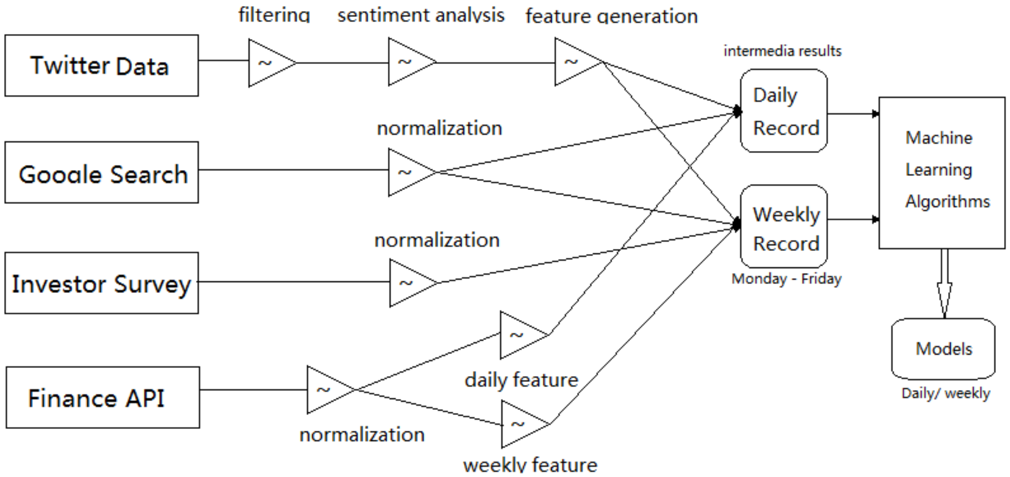
Rao, T., and Srivastava, S. (2012).

- Twitter feature generation and correlations
  up to 0.88 correlations, average value 0.5
- EMMS Prediction
  91% direction accuracy.

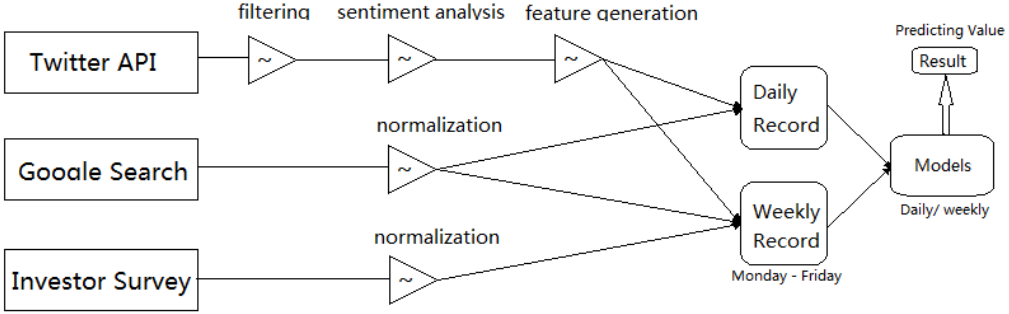
Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

Predicting Model Training



Daily/ Weekly Predicting System



Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

Methodology

- Data filtering and cleaning
- Sentiment analysis
- Feature generations
- Machine learning algorithms
- Predicting input

Twitter Dataset

- Time range: 2009/7/31 - 2009/12/31
- Size: 58.4GB, 400 million tweets
- Format: timestamp, userId, content

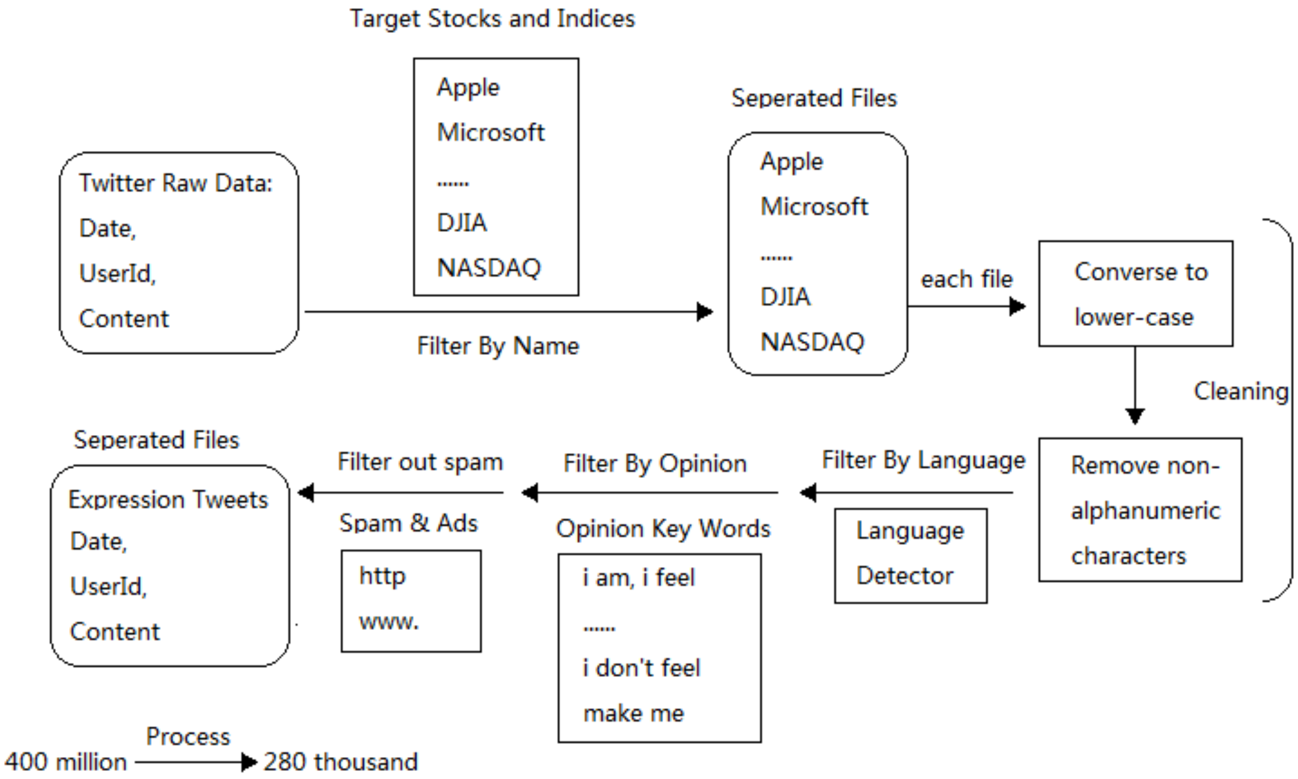
  T 2009-07-31 23:21:12

  U 214325436

  W I just got my new iPhone from Apple store.

- 20-30% of all public tweets

Data Cleaning and Filtering



Sentiment Analysis

- LingPipe from Alias-i
- Algorithm: computational linguistics
- Training set is from Internet Movie Database
  10,000 comments with labels.
- Classified in to 3 classes
  positive, neutral and negative

Feature Generations

- Twitter Sentiment features generation
- Finance features generation
- Search engine features generation
- Investor intelligence features generation

Twitter Sentiment Features

- *Mt-Positive*: total number of positive tweets
- *Mt-Negative*: total number of negative tweets
- Bullishness Bt:

$$B_t = \ln\left(\frac{1+M_t^{Positive}}{1+M_t^{Negative}}\right)$$

- Message Volume:

$$V_t = \ln\left(M_t^{Positive} + M_t^{Negative}\right)$$

- Agreement among positive and negative At:

$$A_t = 1 - \sqrt{1 - \frac{M_t^{Positive} - M_t^{Negative}}{M_t^{Positive} - M_t^{Negative}}}$$

Finance Features

- Yahoo Finance API - Historical Stock Price Data
- Close, Trade Volume, Open, High and Low

$$R_t = (\ln Close_t - \ln Close_{t-1}) \times 100$$

- Return:
- Close: Ct = ln Close t
- Trade Volume:

$$TV_t = \ln(TradeVolume_t / 10000)$$

- Volatility:

$$Vol_t = \sqrt{\frac{1}{2}\left[ln\,\frac{H_t}{L_t}\right]^2 - 2(ln2 - 1)\left[ln\,\frac{C_t}{O_t}\right]^2}$$

Example Feature set

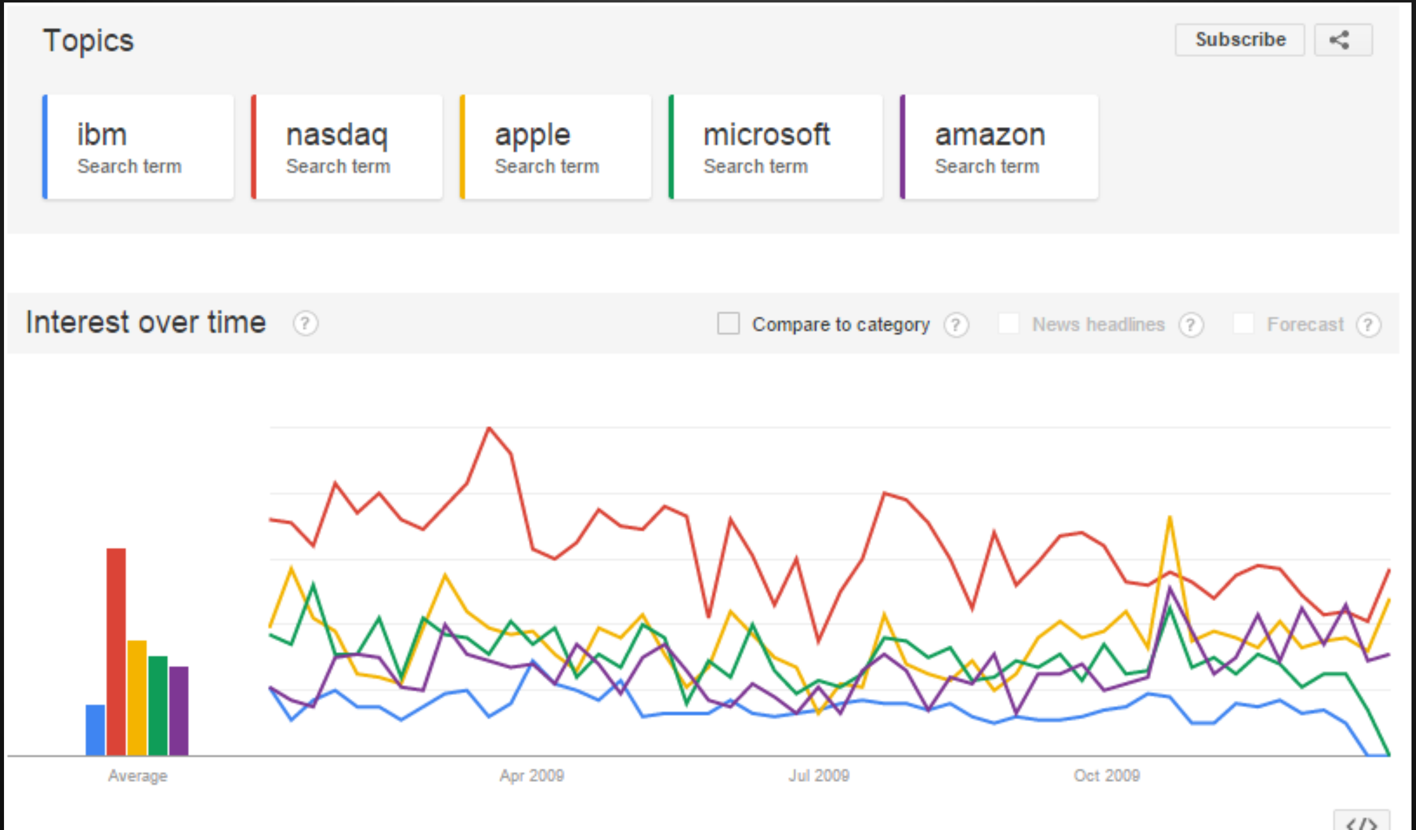
date	pos	neg	bullishness	m-volume	agreement
8/3/2009	91	381	-1.424	6.157	-0.271
8/4/2009	73	254	-1.237	5.790	-0.246
8/5/2009	79	249	-1.139	5.793	-0.232
8/6/2009	26	130	-1.579	5.050	-0.291
8/7/2009	29	118	-1.378	4.990	-0.267

date	return	close	trade volume	volatility
8/3/2009	1.844	3.169	9.193	0.00600625
8/4/2009	-0.530	3.163	9.198	0.005345883
8/5/2009	-0.266	3.161	9.265	0.013344811
8/6/2009	-0.729	3.153	9.051	0.013250721
8/7/2009	0.972	3.163	9.177	0.0076796557

Search Engine Data

- Google Insights Search
- Search volume data by given time range.
- Categories of terms
  - Investment
  - Finance
- Frequency value from 0 to 100.

Google Search Query - Investment



Investor Survey Data

- The American Association of Individual Investors
- Vote on S&P 500
- bullish + neutral + bearish = 100%
- spread = bullish - bearish
- 8 week bullish average

Example of weekly feature set

	GIS Fin	GIS In	bullish	neutral	bearish	spread	bullish 8 week average
2009-12-27 - 2010-01-02	21	21	0.3768	0.2464	0.3768	0	0.3866
2009-12-20 - 2009-12-26	20	19	0.4211	0.2947	0.2842	0.137	0.3815
2009-12-13 - 2009-12-19	24	22	0.4268	0.2195	0.3537	0.073	0.3795
2009-12-06 - 2009-12-12	24	26	0.4158	0.2475	0.3366	0.079	0.3853
2009-11-29 - 2009-12-05	30	31	0.4166	0.1666	0.4166	0	0.3772

Machine Learning Algorithms

- Decision Tree
  - Decision Stump
  - Bootstrap Aggregating
- Regression
  - Linear Regression
  - Gaussian Regression
- Neural Network
  - Radial Basis Function Network
  - Multilayer Perceptron

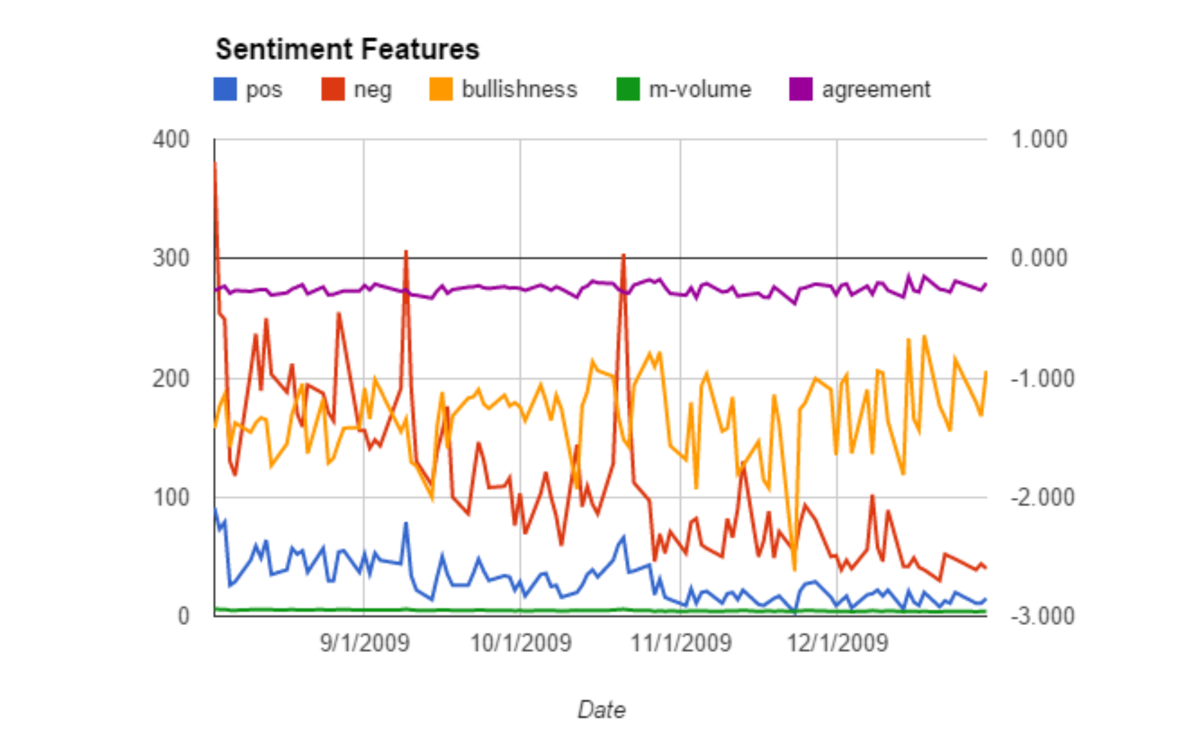
Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

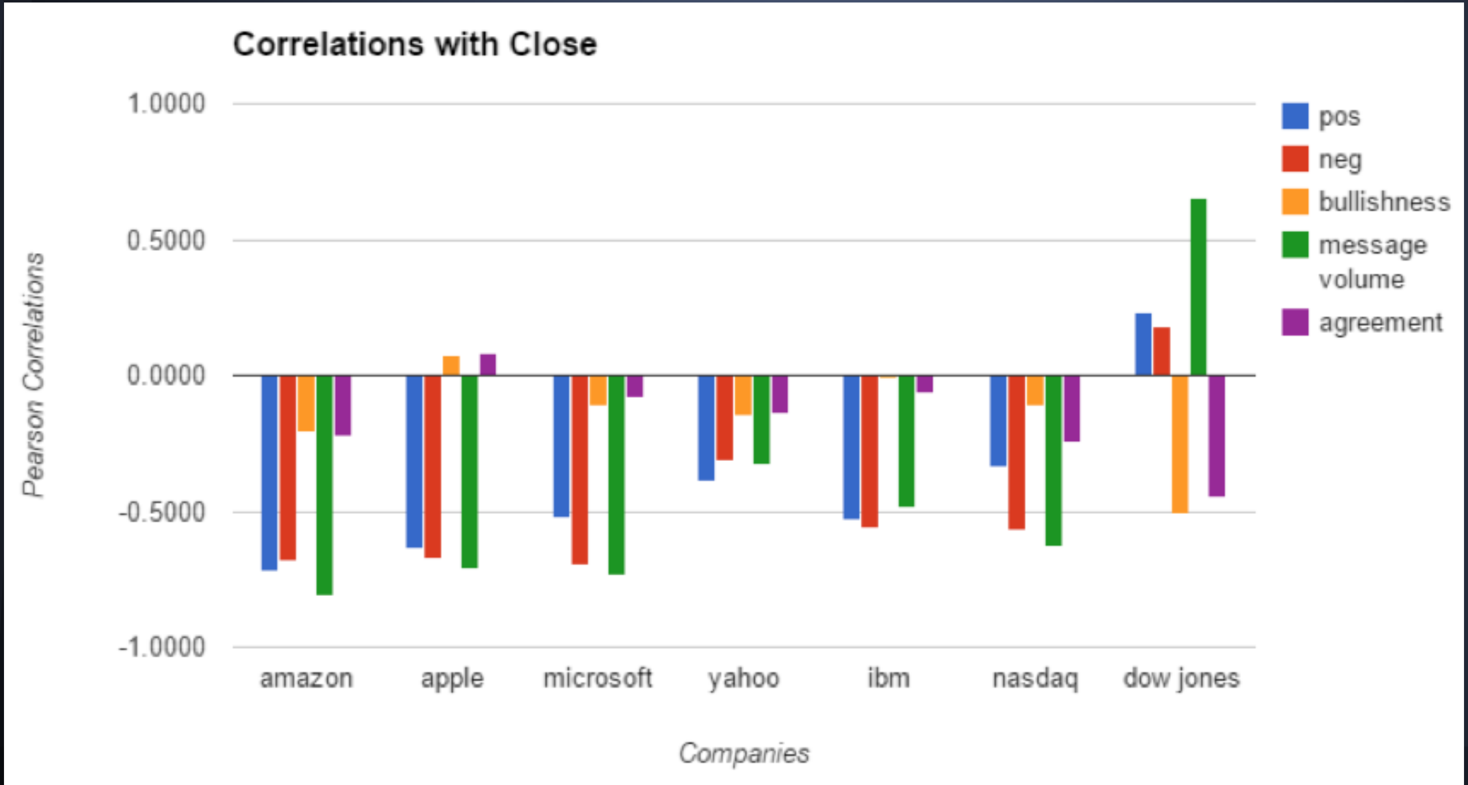
Correlations

- Daily sentiment & finance
- Weekly sentiment & finance
- Weekly GIS & finance
- Weekly AAII & finance
- Time Lag analysis

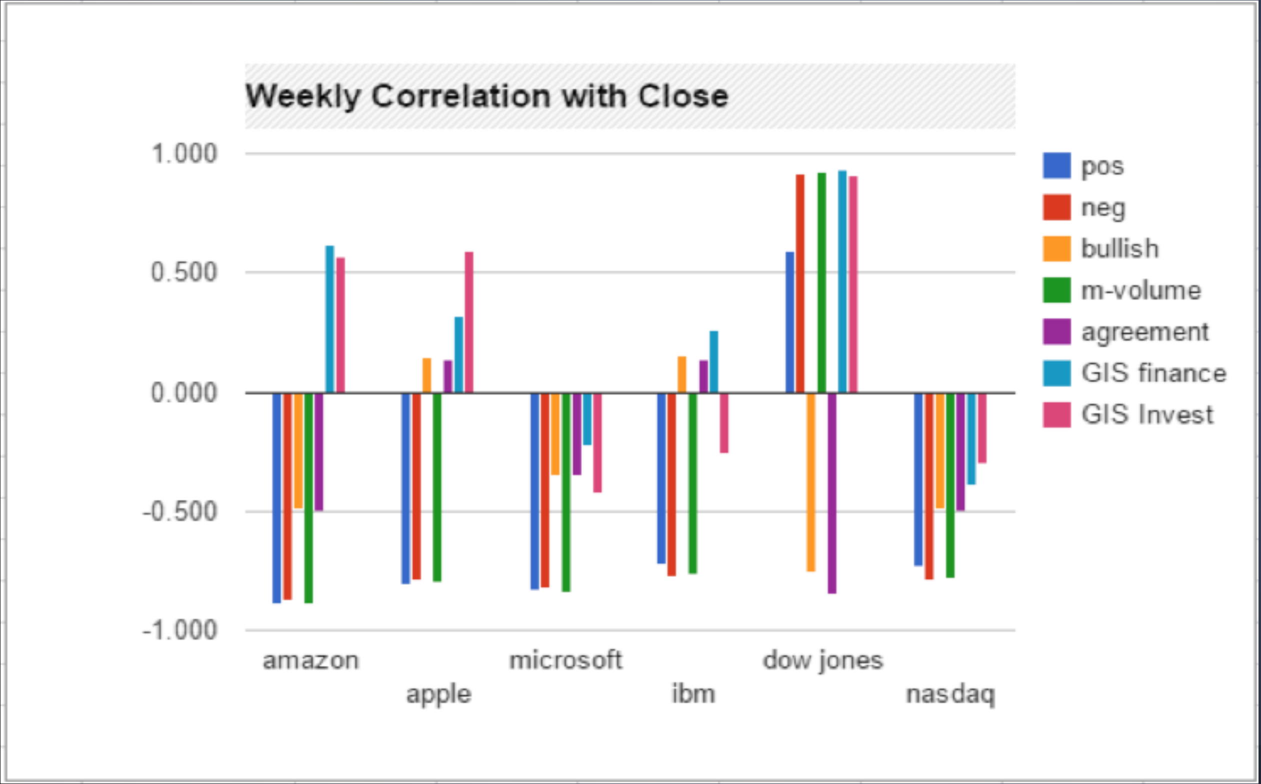
Daily feature set



Daily sentiment correlation

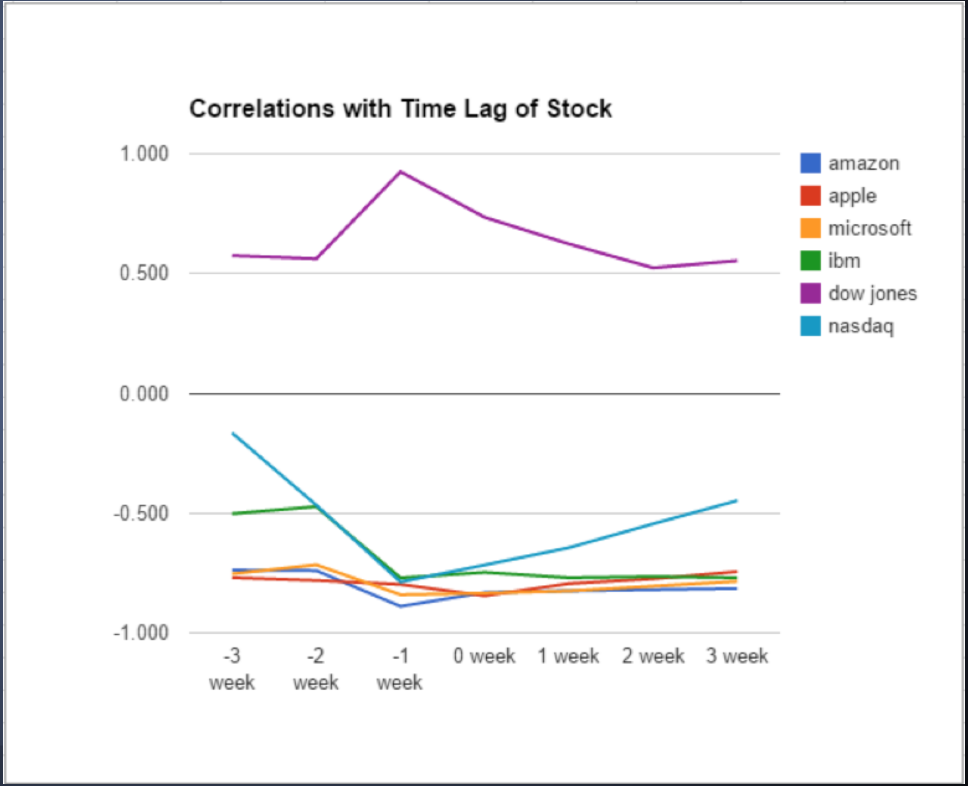


Weekly sentiment & GIS correlation

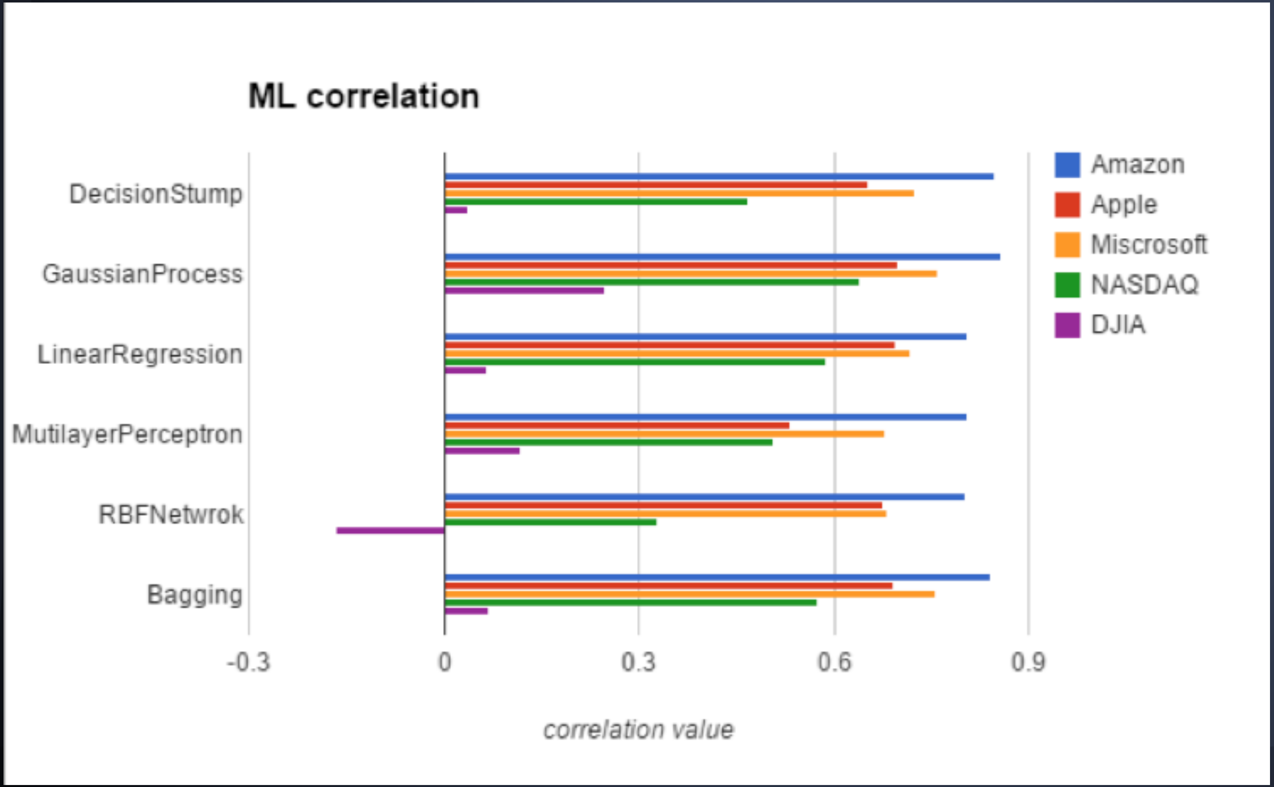


Time Lag Analysis

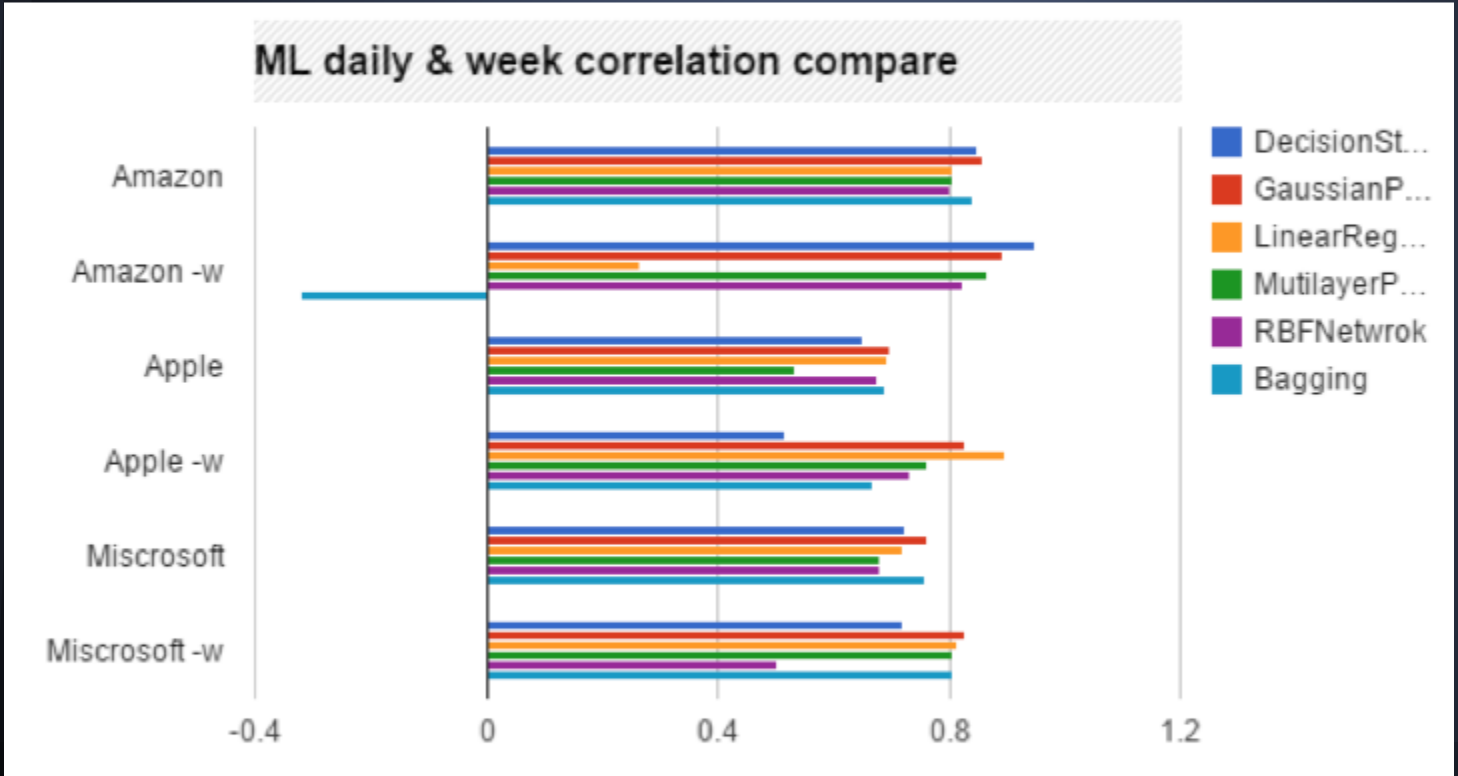
Hypothesis: Twitter sentiment can predict stock price of near future.



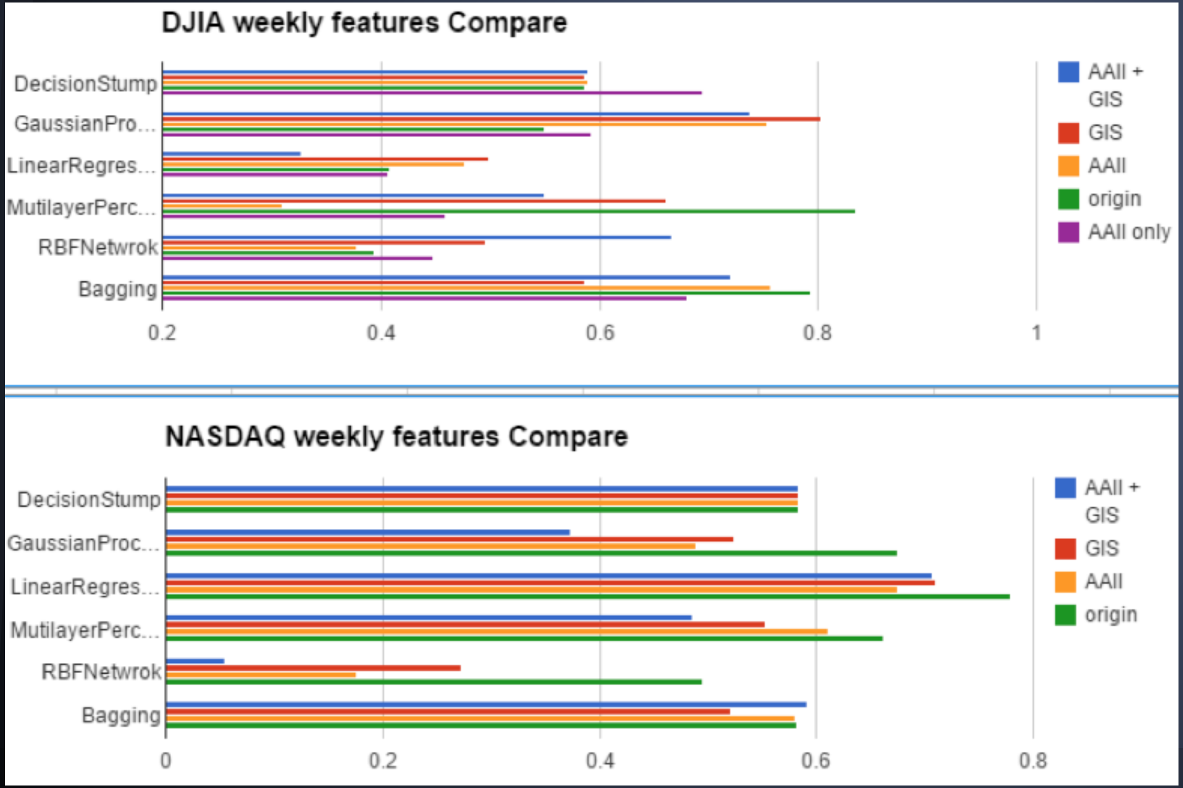
Machine Learning Result - Daily



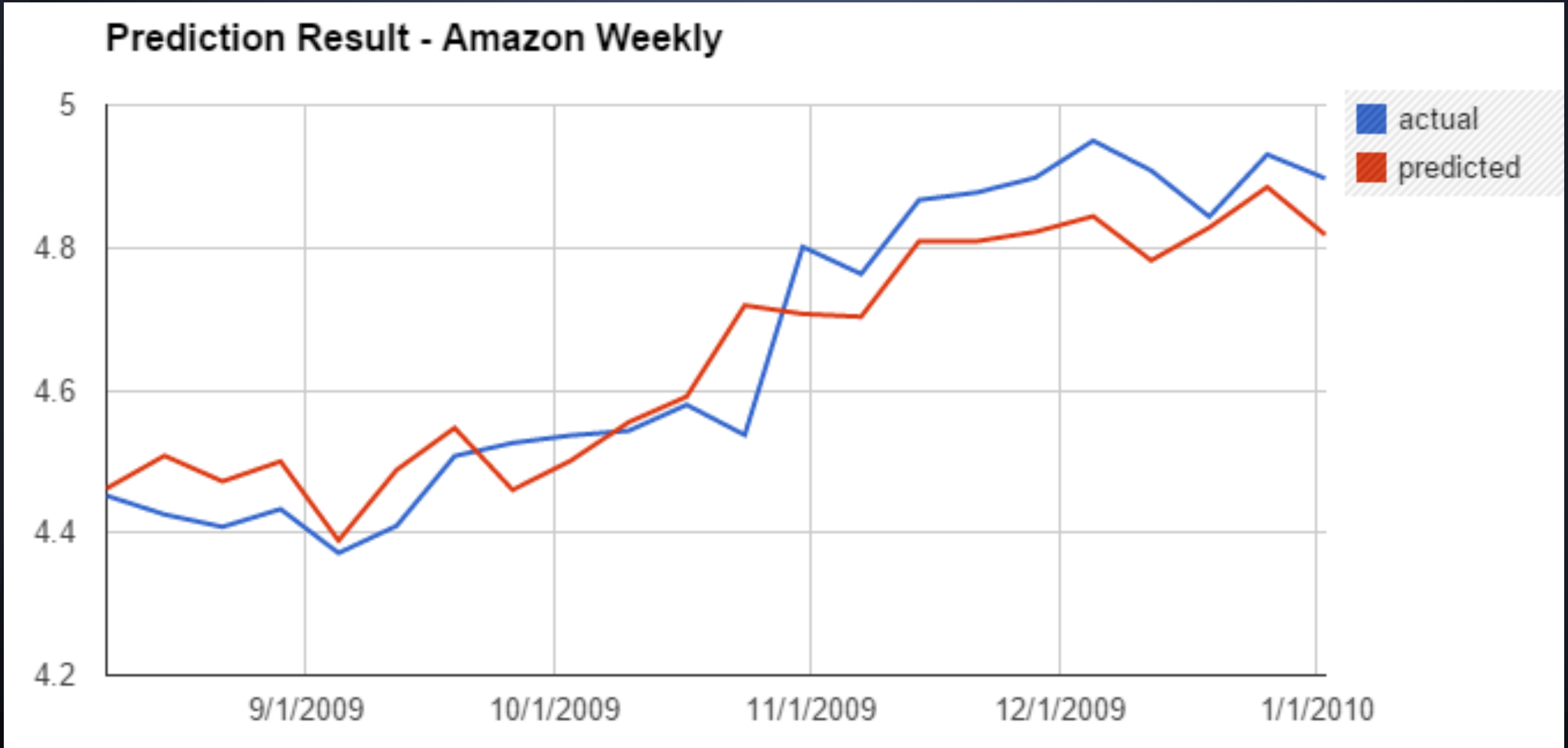
ML Daily & Weekly Comparison



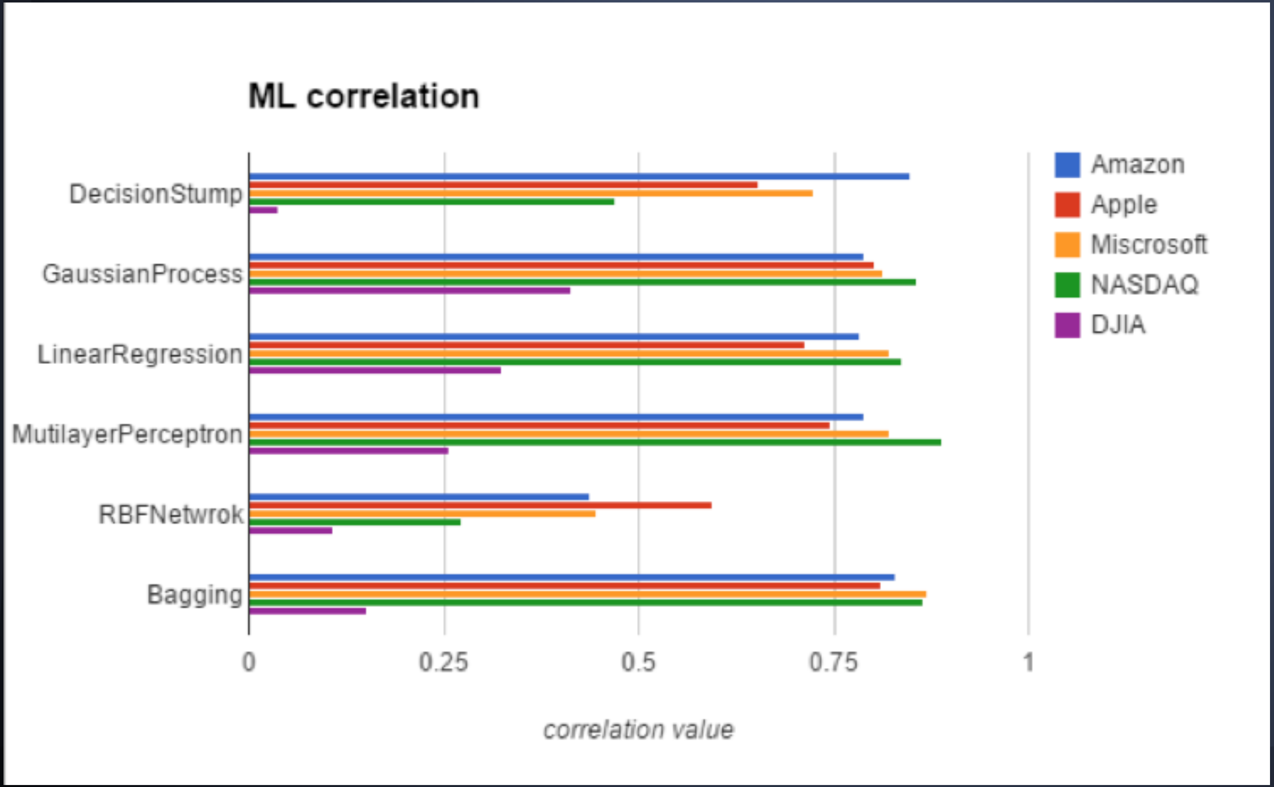
ML GIS & AAII comparison



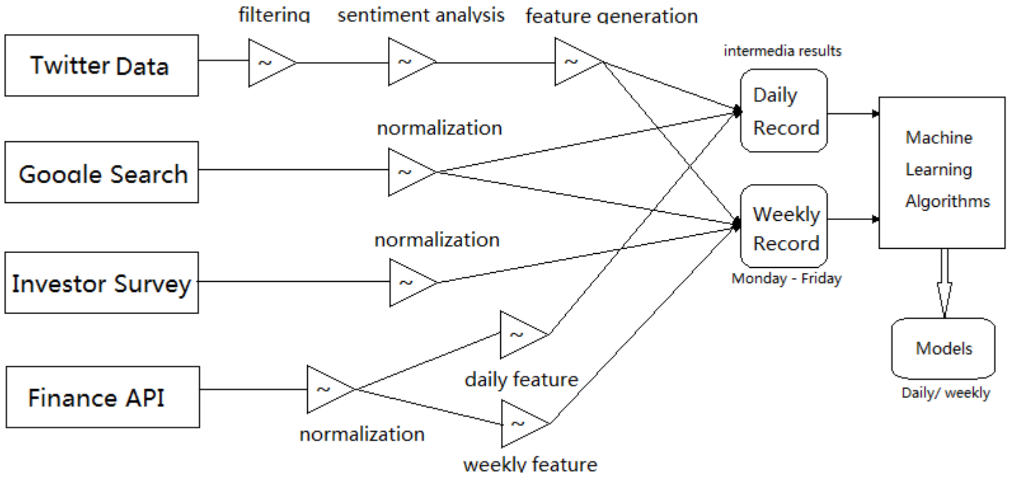
Prediction Result - Amazon Weekly



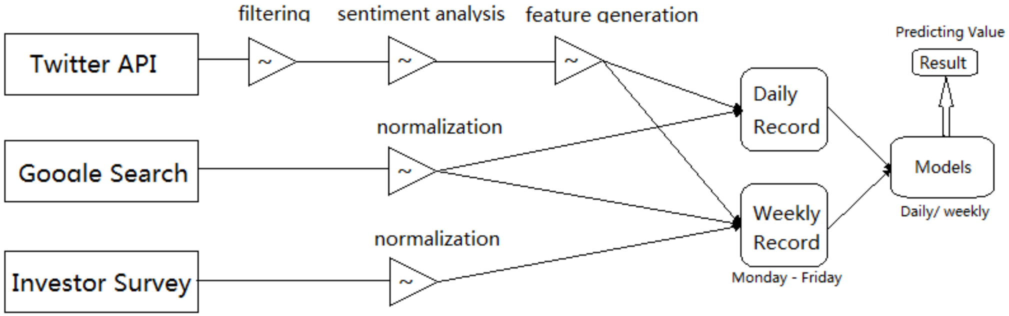
ML training 90%, test 10%



Predicting Model Training



Daily/ Weekly Predicting System



Streaming Data

```
Function GetRealTimeTweets
establishConnection();
query(names);
FOR each marketDay
        FOR each tweet IN queue
                tweet = queue.take();
                parse(tweet);
                IF tweet#language eqauls 'EN'
                        AND tweet#content CONTAINS word in opinionWords
                        AND tweet#content NOT CONTAINS 'http' OR 'www.'
                WRITE tweet TO file(name);
        END
END
```

Streaming Data

Function GetRealTimeTweets

establishConnection();

query(names);

FOR each marketDay

    FOR each tweet IN queue

        tweet = queue.take();

```
660100
Sat Nov 22
Enter to win a $750 Amazon gift card! #giveaways #totallyawesomegiveaway http:\/\/t.co\/2X42tHmPat"

660400
Sat Nov 22
You can finally watch Microsoft\u2019s \u2018E.T.\u2019 documentary on Xbox http:\/\/t.co\/m8Q0lkRll5
unny"

660700
Sat Nov 22
RT @AppIeOfficiel: Apple GLASS – the future is here \ud83d\udcf1  http:\/\/t.co\/izsxmMo08J"
```

Overview

- Introduction
- Related Work
- System Architecture
- Methodology
- Result and Analysis
- Summary and Future Work

Summary

- 3 sets of feature sets are generated and results show its strong correlations with stock price movement.
- A prediction system is built consists of the model training component and the real time data collection component.

Future Work

- Improve Twitter filter for tweets closely related to stock market.
- Train my own sentiment classifier with manually labeled dataset.
- Other algorithms

Question ?