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.
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

- 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’

- 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.
   ● 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
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

Target Stocks and Indices

Twitter Raw Data:
- Date,
- UserId,
- Content

Seperated Files:
- Apple
- Microsoft
- ....
- DJIA
- NASDAQ

Filter By Name:

Seperated Files:

Filter out spam:
- Spam & Ads
  - http
  - www.

Filter By Opinion:
- Opinion Key Words
  - i am, i feel
  - ...... i don't feel
  - make me

Filter By Language:
- Language Detector
- Remove non-alphanumeric characters

Each file:

Converse to lower-case

Cleaning

400 million → 280 thousand
Sentiment Analysis

- LingPipe from Alias-i
- Algorithm: computational linguistics
- Training set is from Internet Movie Database
  10,000 comments with labels.
- Classified into 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

Return:
\[ R_t = (\ln \text{Close}_t - \ln \text{Close}_{t-1}) \times 100 \]

Close: \[ C_t = \ln \text{Close}_t \]

Trade Volume:
\[ TV_t = \ln(\text{TradeVolume}_t/10000) \]

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

**Positivity and Agreement**

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

**Return and Volatility**

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

Topics

ibm  Search term
NASDAQ  Search term
apple  Search term
microsoft  Search term
amazon  Search term

Interest over time

Average  Apr 2009  Jul 2009  Oct 2009
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

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

Correlations with Close

Pearson Correlations

Companies

-1.0000
-0.5000
0.0000
0.5000
1.0000

amazon    apple    microsoft    yahoo    ibm    nasdaq    dow jones

pos
neg
bullishness
message
volume
agreement
Weekly sentiment & GIS correlation
Hypothesis: Twitter sentiment can predict stock price of near future.
Machine Learning Result - Daily

ML correlation

<table>
<thead>
<tr>
<th>Model</th>
<th>Amazon</th>
<th>Apple</th>
<th>Microsoft</th>
<th>NASDAQ</th>
<th>DJIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DecisionStump</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GaussianProcess</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LinearRegression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultilayerPerceptron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBFNetwork</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

correlation value

-0.3 0 0.3 0.6 0.9
ML Daily & Weekly Comparison

ML daily & week correlation compare

- Amazon
- Amazon -w
- Apple
- Apple -w
- Microsoft
- Microsoft -w

- DecisionSt...
- GaussianP...
- LinearReg...
- MultilayerP...
- RBFNetwork
- Bagging
ML GIS & AAlI comparison
ML training 90%, test 10%
Function GetRealTimeTweets
establishConnection();
query(names);
FOR each marketDay
    FOR each tweet IN queue
        tweet = queue.take();
        parse(tweet);
        IF tweet#language equals 'EN'
            AND tweet#content CONTAINS word in opinionWords
            AND tweet#content NOT CONTAINS 'http' OR 'www.'
        WRITE tweet TO file(name);
    END
END
Function GetRealTimeTweets
establishConnection();
query(names);
FOR each marketDay
  FOR each tweet IN queue
    tweet = queue.take();
    IF tweet#language equals 'EN'
      AND tweet#content contains word in opinionWords
      AND tweet#content not contains 'http' OR 'www.'
      WRITE tweet TO file(name);
  END
END
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 ?