Social Data Analysis for Stock Market Indicator

Chanon Hongsirikulkit
Hathfi B. Hakim

(hongsir2@illinois.edu)
(hbhakim2@illinois.edu)
Introduction

- People share news, ideas, and opinions on social network
- Social network becomes an unstructured data source contains valuable knowledge
- Many researches exploit the knowledge (e.g. topic, sentiment) to enhance the current applications
- Some useful applications:
  - Stock market prediction
  - Box-office revenues forecasting
- Twitter CashTag
  - Twitter provide stock ticker symbol, prefixed with a dollar sign ($)
  - Example: $AAPL for Apple, $GOOG for Google, $YHOO for Yahoo, etc.
Related Work

● Find correlation between sentiment on Twitter and DJIA (Bollen and Mao 2011)
● Use Twitter to predict individual stock price and box-office revenue of movie by constructing model based on various types of machine learning classifier (Arias, et. al. 2015)
● Propose the algorithm to extract sentiment and topic from Yahoo Finance Board for stock prediction (Nguyen, et. al 2015)
● Use social network features (e.g. verified account, numbers of follower) to cluster comovement stocks (Liu, et. al 2015)
Project Goals

- Can sentiment information on social network indicate stock price movement?
- Can topic explain the correlation between sentiment stock price movement?
Methods

- Data Collection
- Data Cleansing
- Sentiment Analysis
- Topic Mining
Methods: Data Collection & Cleansing

- **Data Collection**
  - Real Time Twitter Streaming API ([https://stream.twitter.com/1.1/statuses/filter.json](https://stream.twitter.com/1.1/statuses/filter.json))
  - Tweets collected using 3 different “cashtag” company ($AAPL, $GOOG, $YHOO) from April 1st, 2016 - April 29th, 2016.

- **Data Cleansing**
  - Filter tweets by language: english
  - AAPL: 137,560 tweets, GOOG: 51,060 tweets, YHOO: 17,360 tweets
  - Total Data: 205,980 tweets
Methods: Data Collection
Methods: Sentiment Analysis

- Construct classifier for sentiment classification
- 3 features are used
  - POS Tagging -> filter only adjective word as a feature
  - Unigram -> top 100 highest frequency words in tweet collection
  - Word Lexicon -> list of positive and negative words
- Use short product reviews to be training dataset
  - Total 10,000 reviews -> 5,000 positive reviews and 5,000 negative reviews
- Apply classifier to classify all tweets
- Calculate daily sentiment score by using sentiment polarity
  \[
  polarity = \frac{\# \text{ positive tweets} - \# \text{ negative tweets}}{\# \text{ positive tweets} + \# \text{ negative tweets}}
  \]
Methods: Sentiment Analysis

- Train 5 weak classifiers
- Use majority voting to classify sentiment label and use 100% confidence from this 5 classifiers
- Test the classifiers using 1000 positive + 1000 negative reviews

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>68.95</td>
</tr>
<tr>
<td>Multinomial Naive Bayes</td>
<td>65.9</td>
</tr>
<tr>
<td>Bernoulli Naive Bayes</td>
<td>64.6</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>63.4</td>
</tr>
<tr>
<td>Linear Support Vector</td>
<td>63.55</td>
</tr>
</tbody>
</table>
Methods: Topic Mining

- Use ToPMine to generate topics and good phrases from tweets
- Daily topic mining
  - Not many topics occur in one day
- Monthly topic mining
  - Still not many occur in one month
  - However, good phrases seems sufficient and useful
- Use ToPMine to construct good phrases from tweets in one month
- Count good phrase occurrences and select top 30 from each day
- Good phrases should represent topics in each day
Result: AAPL

Accuracy: 62%
Result: GOOG

Accuracy: 71%
Result: YHOO

Accuracy: 52%
Result: AAPL

GM https
iPhone SE
version of the iPhone
Credit Suisse
Investors are overlooking
Apple iPhone
things in tech you need to know today
things in tech
tech you need to know today
iPad Pro
Apple iPhone SE
HFT Algos
Price Target
triggered SELL
Brean Capital
billionth iPhone
AAPL PT
Apple will sell
Jim Cramer
sell its billionth iPhone
inch iPad
inch iPad Pro
Apple Music
Apple Pay
Apple Watch
price at close
Stock Chart
Stock Price
Good morning
tech news

------------- AAPL Apr 04 -------------

iPhone SE
Apple Watch
Early movers
Tim Cook
Apple CEO
CEO Tim Cook
Tim Cook is joining
Tim Cook is joining the board
joining the board
human rights
Apple CEO Tim Cook
HFT Algos
iPhone hack
tech you need to know today
bad for Apple
things in tech
things in tech you need to know today
FBI says it only knows how to hack
key Apple
Apple Watch Herm
Apple iPhone
Apple FBI
San Bernardino
sharing iPhone
San Bernardino iPhone
Price Target
Apple iPhone SE
Wall St
Wall Street

------------- AAPL Apr 07 -------------

Huge Wins
Waiting for the Fed
Apple gives us huge
billion in cash
billion left
market cap
Bears today
iPhone sales
NHL teams
NFL teams
NBA teams
MLB teams
Tim Cook
Wall St
earnings were awful
Real Madrid
st time
Wall Street
market cap shrinks
earnings decline
Dead body
Jay Z and Beyonce
rumors about the iPhone
Increased to Buy
Rating Increased
Apple shares
things in advertising you need to know today
weak earnings
conference room
Skip Bayless

------------- AAPL Apr 27 -------------
Future Work

- Perform the sentiment analysis classifier to larger twitter data (1 year)
- Real time computation for stock price indicator using sentiment analysis classification to Twitter data stream
Conclusion

- People tweet with cashtag more often in weekdays and tend to less at weekend because the stock market only open in weekdays.
- Overall the twitter sentiment polarity for specific stock symbol cashtag matches the movement of the stock price with > 50% accuracy.
- Topic can help describe the correlation between Twitter sentiment and stock price.
Thank You

ขอบคุณครับ

Terima Kasih