Twitter sentiment classification and its application in trading

SI618 Project, Xinye Xu (xinyexu)

Motivation:

Twitter is a popular microblogging service, providing online news and social networking service on which users post and interact with messages(called "tweets"). These tweets often show their attitudes towards different recent topics. From the paper by Alec Go, Richa Bhayani, Lei Huang, Twitter Sentiment Classification using Distant Supervision, they introduced a novel approach for automatically classifying the sentiment of Twitter messages. Also, they used empirical tests show that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data.

The ubiquity of data today enables investors at any scale to make better investment decisions. From Gabriele Ranco, Darko Aleksovski , Guido Caldarelli, Miha Grčar, Igor Mozetič's The Effects of Twitter Sentiment on Stock Price Returns, the sentiment polarity of Twitter peaks implies the direction of cumulative abnormal returns. The amount of cumulative abnormal returns is relatively low (about 1–2%), but the dependence is statistically significant for several days after the events. Therefore, it is worthwhile to detect the best model to predict sentiment, and relationship between sentiment and specific stock movement, combined with tweets classification for relevant stocks movement. We use stock of IBM as an example.

Data Exploration

Two Datasets:

Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter. (Twitter Dataset: http://help.sentiment140.com/for-students/)

1) Following is an example of the data file, which begins on April 7 to June 25 in minutes, containing 6 fields, 1048576 rows (tweets) in total:

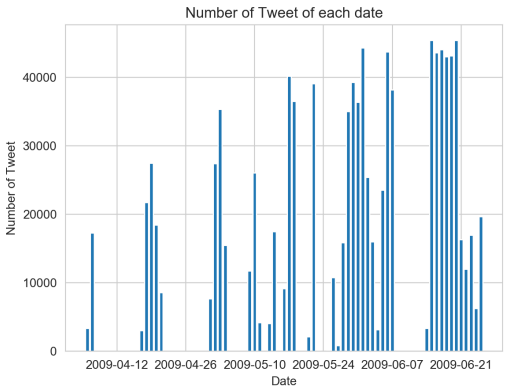
Polarity	ID of the tweet	Time in minutes	Query	User that tweeted	Text of the tweet
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D

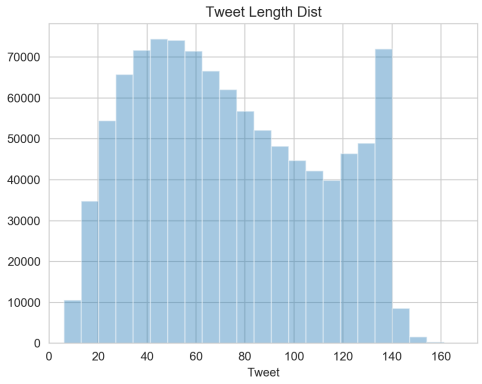
The polarity of the tweet (0 = negative, 2 = neutral, 4 = positive) is sentiment baseline. Analysis by Twittratr website using basic searching keywords method. Also, neutral sentiment was dropped by the original training data producer.

2) Stock Price: IBM price by minutes during same dates, scraping from (Kibot http://www.kibot.com/buy.aspx#free_historical_data). Cover the same time frame as the twitter datasets. And most of the data happened during 7A.M to 7 P.M. of the day, including pre-marketing trading and After-Hours Trading. Following is 3 samples from the IBM dataset, closed price is used for pricing.

Date	Time	Open	High	Low	Close	Volume
4/7/2009	09:03	76.47	76.47	76.45	76.45	2214
4/7/2009	09:05	76.39	76.39	76.37	76.37	263

From the plots, we can notice from April 6, 2009, to June 25, 2009, although the time frame is not daily continuous, meaning some dates do not have daily twitter tweets between this time interval. But given the goal is to test the intraday influence on stock market, the loss of daily continuousness will not be a big concern. Also, we can notice length of most tweets range form 20-80 lengths, and there is also quite large number of tweets have over 130 lengths texts.





Data Manipulation Methodologies

Removing dates without tweets (NA) in IBM stock price dataset. As all text and financial data are real, it is hard to consider whether they are outliers. We might keep all of them in datasets and conduct mining to figure out more useful information.

Since tweets occurred in minutes and seconds, group programming is often used to match intraday stock price in minutes, as well as daily sentiment and distribution analysis. Also, in order to find more related tweets with IBM, a "key sentence" of describing IBM's main business filed should be used for calculate the similarity with differed tweets.

Another challenge is that the tweets dataset might not be the whole dataset, as it was found that after May 28, 2009, all polarities are 0 (figure in next chapter q2), meaning negative. Therefore, in order to investigate the relationship between tweets sentiment and stock movements, that period has to be dropped. Also, during the 48 days of the whole dataset, we should check the weekdays as there is no market during weekends. Luckily, from the table below, for example, after dropped, there is 6 days Monday tweets, it contains all weekdays but they are not even distributed.

Weekdays	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Whole	9	6	5	6	6	8	8
Dropped	6	2	3	3	3	4	5

There are four main questions will be explored in this paper:

1) How to process text information into categories and label them with sentiments?

Texts can be classified by natural language processing. More accurate sentiments classification methods: Naive Bayes, Random Forest, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM).

2) Did people more prefer to show their negative attitude in the Twitter?

As neutral sentiment was dropped by the original training data producer, there are only positive and negative attitudes in this datasets. But we can still use the ratios in total dataset and dates during this time frame to solve this question.

3) Is there a correlated relationship between Twitter sentiment and the stock market?

Correlation plots, Heat plot, Basic Linear Regression with Stock index returns.

4) How to utilize the sentiment data from Twitter to predict stock index movement and test its effectiveness?

To mine more useful information, it can be found IBM has business in 29 specific industries, and we add them up as the key words, and 'IBM' itself to filter more relevant tweets in tweets datasets, and then figure out the sentiment relationship with stock price movement.

- Aerospace and defense
- Automotive
- Banking and financial markets
- Chemicals
- Construction
- Education
- Electronics
- Energy and utilities
- Government
- Government - US Federal
- Healthcare
- Insurance
- Life sciences
- Manufacturing
- Metals and mining
- Oil and gas
- Retail and Consumer Products
- Telecommunications, media and entertainment
- Travel and transportation

Modeling Analysis and Results

There are four main questions will be explored in this paper:

1) How to process text information into categories and label them with sentiments?

As the time series time, we separate the training (first 80% dates) and test datasets to evaluate the best model to estimate sentiment, so we have length of sampling twitter in training 758080; length of twitter in test 290496; ratio of total train data/test 72.3%. After separation, cross validation and other model selection methods can be used to test whether models can predict sentiments accurately compared with benchmark: Polarity.

Texts can be classified by natural language processing. More accurate sentiments classification methods: Naive Bayes, Random Forest, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM).

Methol.1: Naive Bayes Classifier

First need to Create text vectorizer by CountVectorizer function, removing the "stop_words" at the same time. Vectorization, and is an essential first step toward text analysis in order to extract features. Each property of the vector representation

is a feature. For text, features represent attributes and properties of documents—including its content as well as meta attributes, such as document length, author, source, and so on.

Methol.2: TextBlob package

TextBlob package used pre-trained Naïve Bayes Model to classify text sentiment., and it was trained on a dataset of movie reviews. Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. For example, for tweet: "@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D", the real Polarity is 0 (negative); and its estimated Polarity by the package is around 0.2167. They still have some difference regarding the sentiment conclusion.

Methol.3: Random Forest

For the Random forests, it creates bootstrapped training sets as in bagging. Grow a decision tree on each bootstrapped training data set, but consider a random subset of variables at each split. Then, it will use majority vote among all the trees to classify a new object.

Summary of model accuracy:

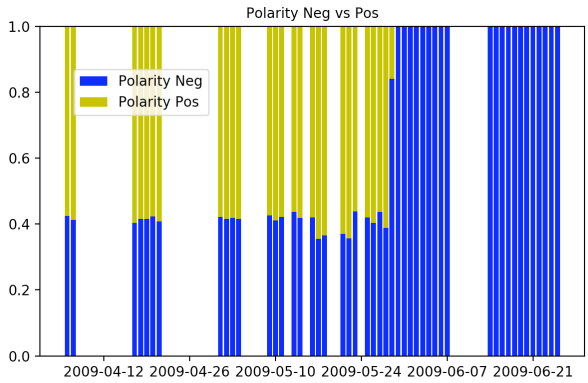
For these three methods, based on both train and test errors, we can notice that Naive Bayes Classifier is better than other methods, also obviously, TextBlob package pre-trained model based on movie reviews will not be suitable for other text mining topics.

Method\Accuracy	Train	Test
Naive Bayes Classifier	85.09%	92.43%
TextBlob package	64.60%	68.90%
Random Forest	82.42%	87.06%

2) Did people more prefer to show their negative attitude in the Twitter?

As neutral sentiment was dropped by the original training data producer, there are only positive and negative attitudes in this datasets. But we can still use the ratios in total dataset and dates during this time frame to solve this question.

With word cloud for whole tweet during this period, we can notice some common words, such as “though”, “think”, “know”, also some words with emotions, for instance, “love”, “bad”, “lol”, and so on. But it is hard to tell the major sentiment. By using original correct polarity with 0 and 4, we can clearly notice that without neutral sentiment tweets, positive tweets account for higher percentage of total tweets in each date. Unless on May 29, and after that date, positive tweets seems to be eliminated. So we have to abandon the information after May 28. Then, we will have an average about 59.09% of total tweets in one day are positive attitude during 26 days. Although facing the challenge of lack of neutral tweets, and this period is quite short, we still have some sense to say people tend to post positive tweet than negative ones.



3) Is there a correlated relationship between Twitter sentiment and the stock market?

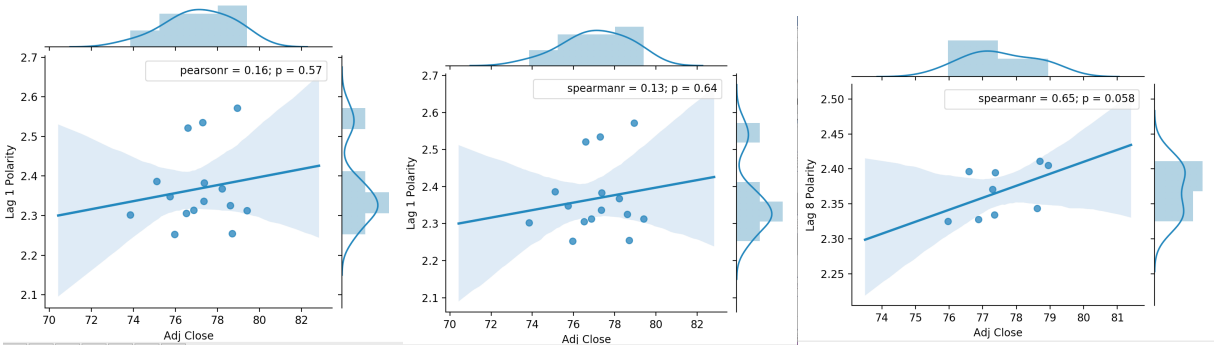
Correlation plots, Heat plot, Basic Linear Regression with Stock index returns.

Unfortunately, as we found the time of tweets data only happened during the nights, we have to give up the intraday research on stock price movement, instead, we work on relationship between daily lag polarity and IBM stock price. Following are the train dataset and lag1 data, but we only have 16 days, excluding the intraday changes. With calculated spearman correlation between1-9, we found 8 rolling windows day for polarity will become more correlated with daily IBM stock price, with maximum

spearman = 0.65. The JointGrid plot of optimal one is the third figure.

	Adj Close	Polarity
2009-04-06T00:00:00.000000000	75.94816	2.30238
2009-04-07T00:00:00.000000000	73.84682	2.34810
2009-04-17T00:00:00.000000000	75.73128	2.38651
2009-04-20T00:00:00.000000000	75.10313	2.30618
2009-04-21T00:00:00.000000000	76.50903	2.36777
2009-05-01T00:00:00.000000000	78.22901	2.31312
2009-05-04T00:00:00.000000000	79.41056	2.33654
2009-05-11T00:00:00.000000000	77.35215	2.31343
2009-05-13T00:00:00.000000000	76.87105	2.25283
2009-05-14T00:00:00.000000000	75.96148	2.32577
2009-05-18T00:00:00.000000000	78.61508	2.53487
2009-05-21T00:00:00.000000000	77.29202	2.52158
2009-05-22T00:00:00.000000000	76.59293	2.57225
2009-05-26T00:00:00.000000000	78.94581	2.38300
2009-05-27T00:00:00.000000000	77.37471	2.25446
2009-05-28T00:00:00.000000000	78.69775	2.44407

Adj Close	Polarity
75.94816	nan
73.84682	2.30238
75.73128	2.34810
75.10313	2.38651
76.50903	2.30618
78.22901	2.36777
79.41056	2.31312
77.35215	2.33654
76.87105	2.31343
75.96148	2.25283
78.61508	2.32577
77.29202	2.53487
76.59293	2.52158
78.94581	2.57225
77.37471	2.38300
78.69775	2.25446



4) How to utilize the sentiment data from Twitter to predict stock index movement and test its effectiveness?

In order to find more relevant tweets with IBM, we extract industries of IBM as a key list. Then, compared with Google pre-trained word2vec model, and my own traing model based on tweets dataset, my model seems to have better distinct.

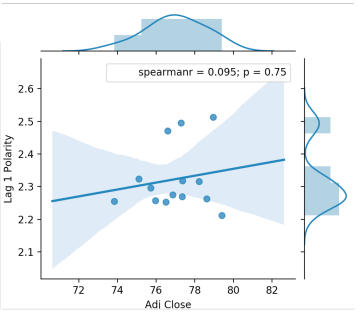
Avg Similarity w.r.t Key words	Tweet 1	Tweet 2	Tweet 3
Google pre-traine	0.03710	0.03041	0.03274
My train model	0.05059	0.05649	0.03685

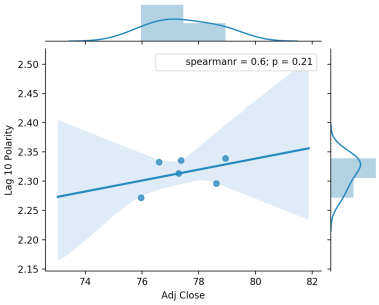
Then, for each tweets, (#388401 in total) tweets average similarity score with respect to IBM key words are calculated. The maximum is 0.33, the minimum is -0.11, mean is 0.04932, which is used as the benchmark to select tweets that are more relevant to IBM. So the half number of tweets are kept.

Following are the train dataset with related tweets average polarity. With calculated spearman correlation between1-12, we found 10 rolling windows day for polarity will become more correlated with daily IBM stock price, with maximum spearman

= 0.6. The JointGrid plot of optimal one is the third figure. But after extracting more related tweets, it seems there is no much improvement of tweets polarity with future 1 day IBM stock price movement. However, with larger datasets, and even smaller time interval, there might be some relationship.

	Adj Close	Polarity
2009-04-06T00:00:00.000000000	75.94816	2.25460
2009-04-07T00:00:00.000000000	73.84682	2.29534
2009-04-17T00:00:00.000000000	75.73128	2.32394
2009-04-20T00:00:00.000000000	75.10313	2.25305
2009-04-21T00:00:00.000000000	76.50903	2.31636
2009-05-01T00:00:00.000000000	78.22901	2.21223
2009-05-04T00:00:00.000000000	79.41056	2.26970
2009-05-11T00:00:00.000000000	77.35215	2.27447
2009-05-13T00:00:00.000000000	76.87105	2.25702
2009-05-14T00:00:00.000000000	75.96148	2.26270
2009-05-18T00:00:00.000000000	78.61508	2.49498
2009-05-21T00:00:00.000000000	77.29202	2.47109
2009-05-22T00:00:00.000000000	76.59293	2.51322
2009-05-26T00:00:00.000000000	78.94581	2.31849
2009-05-27T00:00:00.000000000	77.37471	2.28415





Reference

1. Alec Go, Richa Bhayani, Lei Huang, Twitter Sentiment Classification using Distant Supervision
2. Gabriele Ranco, Darko Aleksovski , Guido Caldarelli, Miha Grčar, Igor Mozetič's The Effects of Twitter Sentiment on Stock Price Returns
3. Stefan Jansen, Hands-On Machine Learning for Algorithmic Trading