Predicting Stock Trading Volume through Social Media Data

Yeqing Chen
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Abstract

Social media plays a big role and can profoundly affect individual behavior and decision-making. This project is to predict the stock trading volume trend from multiple online sources. I chose Facebook to demonstrate the relationship between the stock trading volume and social media data. I used Twitter API to obtain the number of tweets, the number of "retweet" and "favorite" of twitter users. Support vector machine Model(SVM) is used in this project. The daily stock volume data is from Yahoo! Finance, Standard & Poor's 500 indexes (S&P 500) during the period of July 2015 to December 2015. Through social media data, I analyze the tweet amount and the response form public. This way, I generated the percentage of impact rate on trading volume through social media data. It would be helpful to investment companies to predict the trend of stock market.

Problem Definition

A few years ago, the Journal of Computational Science published a paper which predicts stock market through twitter mood by using Google Profile of Mood States (GPOMS) and a Self-Organizing Fuzzy Neural Network. They found an accuracy of 86.7% in predicting the stock price trend and the "calm" mood out of the six dimension mood plays an key role in the prediction.
This project mostly focused on the number of tweets to analyze the correlation between the stock trading volume and twitter data. Based on the number of tweets, number of retweets and favorites on each tweet of the target company, generate the impact rate. Furthermore, I chose certain amount of tweets content to implement sentimental analysis. This way, we can see the comparison of the two results. Naïve Bayes and decision tree were used as the classification method.

Design and Implementation

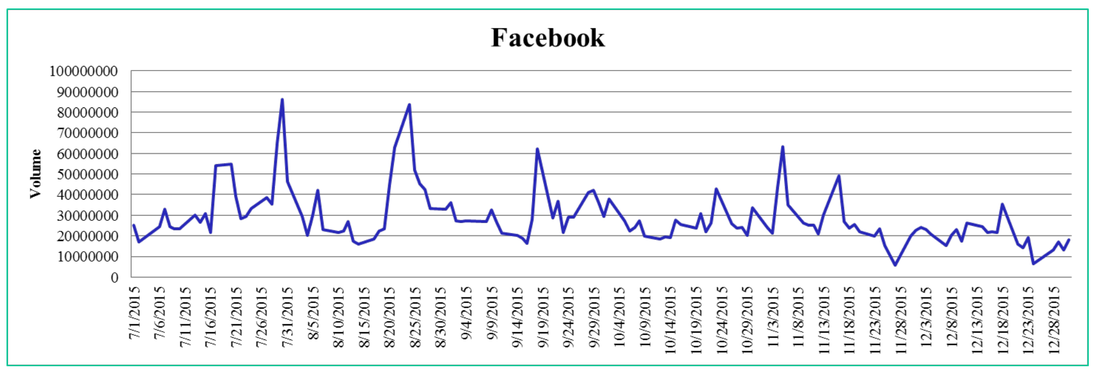
The relation between stock market and social media data based on Linear Regression.

$$Y_t = \alpha + \beta X_t + \varepsilon_t;$$

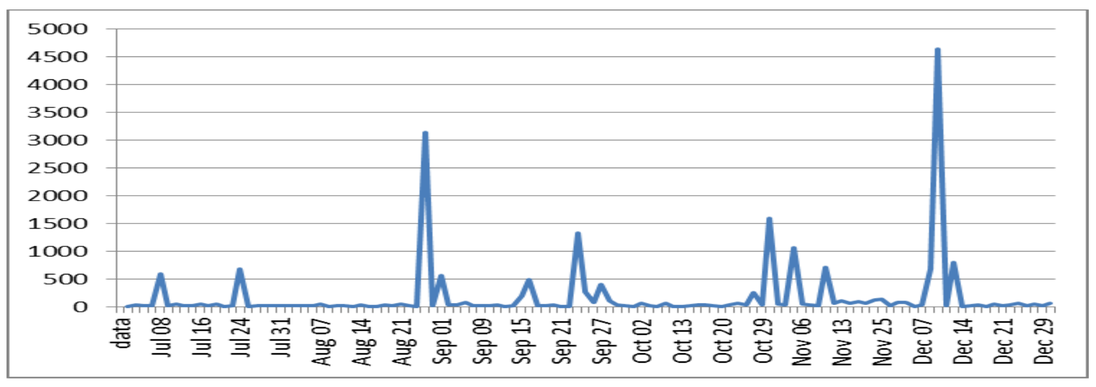
Where Y_t represents a stock indicator on day t , X_t represents the related Twitter predictor on day t, α is the intercept, β is the slope, and ε_t is a random error term for day t. α and β are the regression parameters need to be determined.
Here is the step. Firstly, I applied that the Twitter predictors are correlated with the stock market indicators; then I would find out whether and how well we can predict the stock market trading trend by using Twitter data. In order to manage this problem, I apply a linear regression to the Twitter predictors and the stock market indicators.

Result

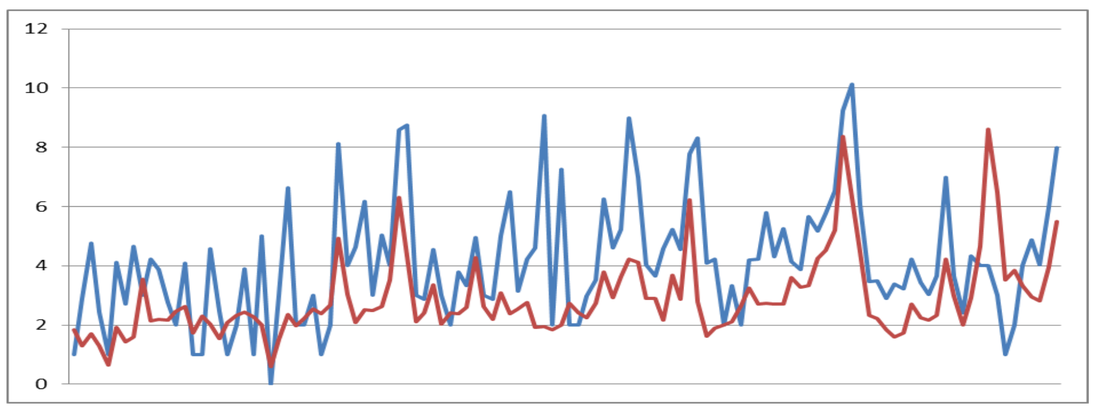
In this phase, I first investigate the correlation between the Twitter predictor and stock market indicators. Then I added sentiment analysis by analyzing tweet moods on some selected tweets . From the charts we can see that the number of tweets, the number of retweets and favorites dose not indicate a very positive correlation. Yet, the tweet index and the positive mood together improved the result.



Dow Jones Facebook daily volume between July 1, 2015 and December 31, 2015.



Tracking the number of daily tweet, the number of retweets and favorites for each tweet by Facebook between July 1, 2015 and December 31, 2015.



The DJIA, the positive mood and the number of tweets chart to show correlation

Performance Evaluation

To evaluate the accurate of the prediction, Confusion Matrix is used. As presented in the table below, the accuracy of the prediction is around 50% when using the data in June as testing data set.

Reality\Prediction	Volume Rises	Volume Falls
Volume Rises	17.30%	36.84%
Volume Falls	14.69%	31.17%

Confusion Matrix

Conclusion and Future Work

In this project, I have investigated whether or not the daily number of tweets is correlated with S&P 500 stock trading volume. On some extent, my result demonstrate that daily number of tweets including its responses from public users is correlated with the stock market trading trend. It seems that Twitter data can be useful to predict stock market. However, the methodologies and algorithms which apply in the data set are very important. For a more accurate result, more factors should be considered, and the other social media resources such as news should also be investigated. This project provides a basic data mining concept. In the future, I will cover some other companies and public news will be investigated in order to find more valuable and interesting patterns and provide a higher accuracy result.