COMP 451 FINAL YEAR PROJECT

Can Twitter Sentiment analysis predict the stock market?

Proposal

Student Name: He Jiqiong

Student ID: 11804425D

Program Stream: 61032

Program Title: Double Degree in Computing and Management

Supervisor: Maggie Li

Co-Supervisor: QI Laurel

Table of Content

Problem Statement

The topic of my project is to use Twitter sentiment analysis to predict the trend of stock market. As one of the largest social networking platform, Twitter allows users to post their interests, moods and opinions within 140 words. With such functions, there are plenty of tweets every day that reflect users' opinions and feelings, which tend to be clustered and coincident towards a specific problem. To some degree, the information in Twitter represents public opinions, which is a potential reflection of stock market prediction. Stock market trend, on the other hand, is hard to predict because of the theory of efficient market. In this theory, the asset prices already reflect all available information, so it is impossible to "forecast" prices or market trend according to any information. However, in fact the fully efficient market is hard to achieve, that is, there are always some delays in the reflection of the information on the stock market, so we can still use some tools and information to predict prices and trends.

By utilizing information from Twitter, we can get information about public opinions about the whole market or a single stock. When majority of users hold similar attitudes towards the whole market or a single stock, the corresponding trend may be influenced by those expressions in Twitter. For example, if large portion of users talking about stock A express positive attitudes in Twitter, the trend of stock A in the future may be positively influenced. Therefore, by examining the sentiment words or expressions in Twitter, we can match and analyze the relationship between sentiment and stock trends to draw conclusions.

Objectives and Outcome

There are two levels of goals of this project. Firstly, the basic goal is to establish a model to predict the increase or decrease of prices of individual stock or index by measuring the representative or majority tweet sentiment words or phrases related. Secondly, my stretch goals include two aspects. On one hand, since the stock market has different sectors (e.g. industrials and transportation) and each sector may have different trends of movement with the same sentiment words or phrases, so I would like to make estimation on separated sectors to increase the efficiency and accuracy of my project. On the other hand, my basic goal is about deciding the direction of changes, but I hope in my stretch goal to upgrade my model to predict the possible range of increase or decrease.

Project Methodology

In my project, I will focus on the stocks listed NASDAQ, because NASDAQ is one of the largest US stock market and its major indices, such as S&P100, are of great significance in US market or even the world market. With such scope, there would be more users in Twitter participating into the discussion of the market and stocks, so the data is more accessible. Meanwhile, NASDAQ divides listed stocks according to sectors and each sector has its index, so it is more convenient for me to conduct stretch goals during my project.

There are several steps in my project:

1. Twitter and tweet collection
This step is to collect sentiment words or phrases from Twitter. Actually there are many types of tools and perspectives to collect Twitter sentiment. Up to now, I have three choices: using Twitter's streaming API to collect data continuously; using OpinionFinder to gather positive and negative moods; as well as Google-Profile of Mood States to collect data in 6 emotional dimensions.

2. Stock information collection
Besides collecting sentiment data from Twitter, I also need to collect stock prices or index prices. Data would be collected from NASDAQ.com and Yahoo Finance.

3. Data Analysis using data mining algorithms
In this step, all required data is collected from Twitter, and I need to hire some algorithms to further analysis the relationship between prices and sentiment words. Basically, Naive Bayes Classifier and Apriori algorithm are necessary. In detail, representative sentiment words or phrases would be selected, and the probability of existence of increase or decrease (or the range according to stretch goals) corresponding to the same time period would be calculated. With the calculated probability, Naive Bayes Classifier is utilized to classify the increase or decrease (or range) of stock prices. Moreover, Apriori algorithm can provide the amount of support and confidence of assertion in the estimation. Besides those basic algorithms, some advanced algorithms which could improve accuracy or efficiency would be analyzed or developed.

4. Observation and conclusions
With the results from step 3, some frequent or important patterns can be mined. According to those results, a model of determining the increase or decrease (or the range) could be developed. Meanwhile, with the results, the benefits or drawbacks of different algorithms can be compared and improved for better accuracy or efficiency.

Project Schedule

Time	Arrangements
September to November	Identify references to understand the topic and find out steps, algorithms to carry out the project; preparing for the proposal of the project.
December to February	Implement the project with tools, algorithms decided in the previous steps; achieve the basic goal; preparing for the midterm report of the project.
March to April	Refine and improve the accuracy and efficiency; try to achieve the stretch goal; preparing for the final presentation and final report.

Resource Estimation

Twitter sentiment collection: Streaming API; OpinionFinder; Google-profile of Mood States.
Stock prices collection: NASDAQ.com and Yahoo! Finace
Data mining tools: Matlab or PASW or VBA (undecided)
Data mining algorithms:

Basic: Naive Bayes Classifier; Apriori
Advanced (from papers and combining by me): Dirichlet Process Mixture Model; Granger Causality Analysis and Self-organizing Fuzzy Neural Network; Linear Regression; or others

Literature Review

1. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013, August). Exploiting Topic based Twitter Sentiment for Stock Prediction. In ACL (2) (pp. 24-29).

Based on sentiments from Twitter, the writers predict the S&P100 stocks. Firstly, they utilize a continuous Dirichlet Process Mixture model to learn the daily topic set from Twitter. Then, for each topic, they derive its sentiment according to its opinion words distribution to build a sentiment time series with VAR. In the following, they regress the stock index and the Twitter sentiment time series to predict the market. According to their results, experiments on real-life S&P100 Index show that their approach is effective and performs better than existing state-of-the-art non-topic based methods.

2. Davies, A., & Ghahramani, Z. (2011, August). Language-independent Bayesian sentiment mining of Twitter. In Workshop on Social Network Mining and Analysis.

This paper presents a new language-independent model for sentiment analysis of short, social-network statuses. The writers demonstrate this on data from Twitter collected by Streaming API and Search API, modelling happy versus sad sentiment, and show that in some circumstances their model outperforms similar Naive Bayes models, which they use as a baseline for comparison, by more than 10%. They also propose an extension to allow the modelling of different sentiment distributions in different geographic regions, while incorporating information from neighboring regions. Finally they raise their considerations when creating a system analyzing Twitter data and present a scalable system of data acquisition and prediction that can monitor the sentiment of tweets in real time.

3. Chen, R., & Lazer, M. (2013). Sentiment analysis of twitter feeds for the prediction of stock market movement. stanford. edu. Retrieved January, 25, 2013.

In this paper, the authors investigate the relationship between Twitter feed contents and stock market movement. Their target is to see if, and how well, sentiment information extracted from these feeds can be used to predict future shifts in prices. They construct a model which contains classification and regression based on linear regression, estimate its accuracy with k-fold cross-validation, and put it to the test on real market data using a mock portfolio. Their results indicate that the model is successful in generating additional profit.

4. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.

In this paper, the writers investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. Firstly, they analyze the text content of daily Twitter feeds by two mood tracking tools, OpinionFinder that measures positive vs. negative mood, and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). Then they cross-validate the resulting mood time series by comparing the ability to detect the public's response to the presidential election and Thanksgiving Day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Finally, their results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. They also find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.