Correlations between emotional factors and the stock market

Max Xiong
Data Science Research Laboratory. Mentor: Dr. S. Samsonau
Princeton International School of Mathematics and Science

Address and Contact Information:
19 Lambert Dr, Princeton NJ 08540
e-mail: Max.Xiong@prismsus.org

Abstract

When people talk about stock market, they most likely relate it to the pure number analysis such as operation profitability, debt equity ratio, dividends for investors, etc. Nowadays, computer algorithm controls more than fifty percent of the trading volume so humans seem less important in the process than they were in the past. However, the large amount of capital invested by normal investors are still considerable. In order to analyze what may possibly influence the decisions made by traditional investors, this research studies the emotion factors in investment, as information is the fundamental criteria for normal investors, by quantifying the emotional levels

Introduction

The main factor that will influence investors' emotion is the information they received about the stock market. The information can be conveyed from traditional media such as newspaper, broadcast, and television. It can also be received through modern social media including Facebook, Twitter, blogs, and numerous websites.

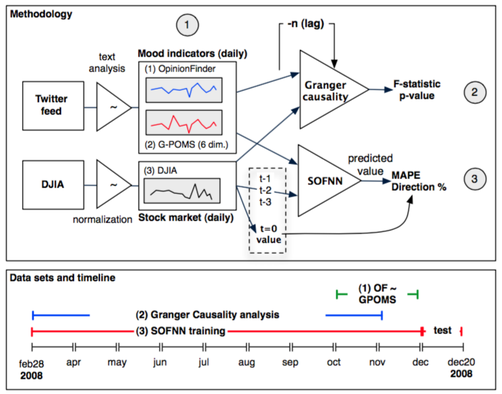
Motivation

In order to have concrete evidence to test the potential relationship between investors' emotion and stock price, it is necessary to find a way to use numbers to represent the emotional level. In sentiment analysis, scientists go through the lexicon and build different versions of word lists to describe the emotions when people use certain words. They conclude each word to be either positive and negative sometimes with different degrees of positivity and negativity. The simplest version would be dividing words into positive words and negative words, which are assigned to the values 1 and -1. The sum of score of all words in one sentence represents the negativity or positivity of the emotion. After having a way to quantify emotion, statistical analysis can be run to verify the potential relationship.

Background

In [1] Johan Bollen et al. implemented :

- Data:9.8 million tweets from 2/28/2008 to 12/19/2008
- Analysis tools: Opinion Finder and Google-Profile of Mood States (can analyze emotions in 6 dimensions)
- Granger-causality test
- Prediction interval: 2-6 day
- Results: 87.6 % predicting the daily up and own changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6 percent.



Main Objectives of this project

1. Capture and Clean Tweets
2. Capture Stock Price Data
3. Score Tweets and Merge by Time
4. Granger-Causality Test and Analysis

Implementation

Chosen Social Media: Twitter The Twitter API allows user to collect tweets based on different types of filters. I choose to search for tweets that contain keyword, ŝtock market, which is the independent variable in this research. After collecting tweets from API, I proceed data cleaning by converting all files into txt form in order to be scored according to the sentiment analysis wordlist. Then, the stock price in the interval of 1 minutes. The last section is the Granger-causality test and running cross validation to verify the relationship between emotions and stock price.

Capture and Clean Tweets

- The first step in capturing tweets or collecting data is connecting the local user with the Twitter API in order to use the searching function online to capture tweets. Twitter API provides certain functions for public study. After registering an Twitter App online, authorization (handshakes) will request user to allow the access, and then the OAuthFactory function connects the API and store the authorization information to the local file my_oauth.Rdata.

RT @philstockworld: Mondays Oil Mess: Rent-A-Rebel Jacks up Prices into the Holiday USOAAPL Earnings – https://t.co/cGHB3WDKA8 https

- Then the data capturing process first read the authorization information and stream tweets that contain keywords entered by user in the function filterStream. It stores the data in a local variable called "tweets.json" as the format of streaming is json. Name can be changed while data will be stored additionally with the same name. ParseTweets function convert all tweets collect into a data frame with time, content, user, location etc.
- In order to process the tweets we collect, mainly scoring the emotion of each tweet. In the scoring function, we get rid of all special symbols, starting users name, punctuation, digit, link, capital letter. It will be easier for judging whether one string contains on string.

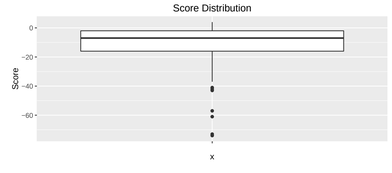
Capture Stock Price Data

- Google provides last 15 days of data with one minutes interval. We only have to construct the url string to access the txt format of table and then format them together. We first go to the url, get organize the data into columns. We then search for the time stamp, which is in POSCIX format. Then we convert it to EST and get two columns from this data frame: price closed and time for the future analysis

Time2	summary.score	CLOSE	close_num
2016-05-17 09:15:00	-14	7896.85	7896.85
2016-05-17 09:16:00	-19	7905.8	7905.80
2016-05-17 09:17:00	-20	7901.6	7901.60
2016-05-17 09:18:00	-29	7902.75	7902.75
2016-05-17 09:19:00	-7	7909.1	7909.10
2016-05-17 09:20:00	-27	7908.25	7908.25

Score Tweets and Merge by Time

- In the scoring function, we use a for loop to search for the positive or negative word in one specific string. It is fairly easy to achieve by lookup each word in the word list and record number of positive words and negative words and calculate the difference between them. However, we also have to get rid of the repetitive count when two positive words or two negative words contain one of the other. The for loop firstly run through negative words and delete repetitive terms and executes in the same way for positive words.



- After we put the tweets into a data frame, we then run the clean and score functions to process the information we have. Then merge the score and the price tables by time. Meanwhile, we also make the smallest units to be 1 minutes instead of 1 second because Google data does not provide the instantaneous price.

Granger-Causality Test and Analysis

- Granger Causality Test shows that whether one factor is the predictive factors of another. It uses one factor at first then use two factors to see the correlation between two polynomials by F test. The result will show R square, the percentage of data or correlation that can be explained. It will tell whether a factor is significant predictive factor of the other. In this test, we want to see whether emotion is the deterministic factor of the stock price. We only have to assign our previous two columns to be time series data and run the Granger-causality test code.

ftest	p.value	R.squared
40.351	6.56e-10	0.261

Results and Conclusions

After running Granger causality test the 10 days interval data, the p-value is $9.03 * 10^{-11}$ and the R^2 is around 0.26. It indicates that there is an moderate correlation between emotional factors and stock market price from the perspective of this research. There is still room for many improvements.

Forthcoming Research

The next stage of the research will be improving dictionary by adapting a wordlist that differentiates the emotional level of the words. Or using machine learning to analyze the emotion by Google Opinion Finder. The keywords search also needs improvement as "stock market" is not the only thing that is related to the stock market. A specific company analysis may also be run considering the public exposure of each company is different. The program is also expected to run the analysis test and prediction at the same time for potential contribution to the trading strategy

References

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

Acknowledgements

I am thankful to PRISMS for providing time and support for my endeavors. The algorithm was implemented in R language [2].