Group 1: Qixu Cao, Xiaoman Dong, Yigao Li, Tao Yu

Professor: Ami Gates

ANLY 501

October 6, 2017

The Impact of Social Media on Stock Prices

**Data Science Problem:**

The goal of this project is to predict stock prices based on the influence of twitter posts by investigating whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can indeed to predict the stock market. This project grabs more than 60,000 rows of data from both Twitter and finance stocks. To be specific, this project analyzes both the total number of tweets as well as the amount of positive and negative words from the tweets about a certain stock in order to predict its stock price in future. Taking Apple Inc as an example, the data collection process includes searching tweets with #AAPL, as well as the hourly stock prices of Apple from the same week. By analyzing these two datasets, this project may able to predict whether the influence of Twitter posts of a certain company could affect the stock price.

In terms of the reason choosing this topic, this project desires to know the impact of social media on the stock market. Many stock market predictors or professionals like to post their opinions and prediction on Twitter, such as Twitter users “Harmongreg,” “Danzanger,” “Fousalerts,” and “Peterlbrandt.”Apart from these individual users, we also consider media sources like Bloomberg and The Wall Street Journal, which like to post stock information on Twitter. Moreover, many popular figures, such as celebrities, frequently comment about important issues on Twitter. Therefore, if this study can explore the relationship between those tweets and the stock market, it may be revealing of whether or not social media would impact the stock market.

Bollen and Mao, from Indiana University - Bloomington, did similar research on Twitter moods and the stock market. They collected millions of tweets and stock prices in a 10-month period and reached the result that considering public moods (as given by tweets) could improve the accuracy of predicting the Dow Jones Industrial Average [Bollen and Mao 1].

**Potential Analysis:**

1. Discovering the relationship between social media and stock price requires at least two datasets, hourly stock price and hourly twitter text. Since stock prices vary wildly over the course of a given day, hourly stock prices are collected instead of daily ones. There are 10 variables in the Twitter data and 7 variables in the stock data.

The data we collect are from Twitter and Alpha Vantage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable names of Twitter Data** | **Explanation of variable** | **Example** | **Data Type** |
| Ticker | Stock quote | AAPL | Categorical |
| user\_id | Id of Twitter user | 1356604729 | String Data |
| user\_name | Name of Twitter user | Patrick Barks | String Data |
| user\_followers | Number of this user’s followers | 180 | Count Data |
| user\_verified | Whether verified by Twitter | FALSE | Binary Data |
| user\_location | Location of twitter user | Staffordshire, UK | String Data |
| user\_description | The description of this user | My life into 140 characters at a time. Mad on football and music. Indie/rock/metal and anything else which takes my fancy.<https://t.co/IPbk1R7QPY> | String Data |
| Text | Text content of tweet including #AAPL | What can we expect from #AAPL today? | String Data |
| created\_at | The time and date of this tweet | Tue Oct 03 13:36:12 +0000 2017 | Time and Dates Data |
| Source | Tweet source | <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a> | String Data |

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable names of Stock** | **Explanation of variables** | **Example** | **Data Type** |
| index | Stock price at exact time | 9/28/2017 12:00:00 PM | Time Data |
| open | The open stock price at 11 am | 12.636 | Numeric Data  (Floating point) |
| high | The highest stock price from 11 AM to 12 PM on 9/28/17 | 12.65 | Numeric Data  (Floating point) |
| low | The lowest stock price from  11 AM to 12 PM on 9/28/17 | 12.61 | Numeric Data  (Floating point) |
| close | The close stock price at 12 PM on 9/28/17 | 12.6285 | Numeric Data  (Floating point) |
| volume | The number of shares or contracts traded in a security market during 11 AM to 12 PM on 9/28/17 | 800120 | Count Data |
| symbol | Ticker | AMD | Categorical Data |

2. Twitter user information can offer a brief assumption about users’ social impact. The Twitter dataset has variables such as “Ticker,” “user\_followers,” “user\_verified,” “user\_location,” “Text” and “created\_at” to describe the user’s tweets that include certain stock names. For example, a twitter user with large number of followers could have a greater social impact; a twitter user with verified information could offer more reliable information, and so on. The target stock market of this project is within the United States, therefore collecting Twitter users’ location from within the U.S. is sufficient. When collecting data from Twitter, this project aims to find tweets with stock index hashtags, like #AAPL, because those tweets with stock indices should be related to the stock market rather than just grabbing data with “APPLE.” The tweets including “APPLE” may talk about the company culture or any number of things, while using the exact stock index “AAPL” can find the tweets more precisely.

The stock market dataset includes variables “index,” “open,” “high,” “low,” “close,” “volume,” and “symbol” to represent the key information of stocks, and it tends to focus on its hourly price rather than the daily price. Hourly price tracks the changes of stocks precisely. For example, there were over 1000 tweets with #AAPL during a single hour (10 am to 11 am) due to the new iPhone release, and that may cause the stock price of Apple to change significantly.. Then, by analyzing this data, it is able to draw a conclusion whether social media can affect the stock market and to predict the stock price’s trend.

3. In terms of the possible directions, this study considers sorting the text from Twitter users by finding texts with positive and negative attitude words. As mentioned above, this project focuses on the American stock market, hence it is also critical to find the users that are in the U.S.. Given the tweets from U.S. locations, the text with attitude words cross-referenced with changes in stock prices can predict future stock prices. This project’s hypothesis is social media can impact stock prices. For example, if Apple releases its newest product and a well-followed Twitter user tweets about it, it could increase the hype around Apple, thus influencing its stock price. One difficulty is determining whether this is simply coincidence.

4. The direction to investigate used is text mining. By putting text search into certain positive terms, such as “increase” and “perform well,” and trying to find out whether those positive words are correlated with a given stock’s price change. However, the text mining process needs to analyze the text contents from Twitter very carefully. Another direction of this project is to sort the users of Twitter and to rate them according to the number of their followers and their status of verified or not. This is important since a high ranking Twitter user’s tweets carry more weight than a lower ranked individual.

**Data Issue:**

The purpose of the study is to find whether stock price is influenced by social media. There are two datasets being collected for this project: Twitter information with hashtagged stock symbols and hourly stock prices. The stock data is from Alpha Vantage, which is pre-collected and cleaned by the website, is relatively clean. The Twitter data is incomplete, noisy, and inconsistent so it requires more effort to clean. Many variables are missing because users did not fill out all their information; some users’ descriptions contain emojis that are hard to be transferred into strings and many users did not even tweet in English, which cannot be found in the attitude dictionary (which is planned to be added to project 2).

Finding the influence between social media and stock price requires tracking the time accurately. Thus, collecting hourly stock price data is crucial. It is relatively easy to collect time information on Twitter, but most stock price data is daily data.

After two datasets have been combined, a new issue appears: the time format is not the same in both data sets. In order to find the relationship between stock price and Twitter content, conforming the time format of both datasets is necessary. Therefore, data cleaning process involves manipulating the time column. A new column, “newtime,” which contains a standardized time format is introduced to the datasets.

**Data Cleaning:**

To measure how clean the two datasets are, two types of rates have been introduced: Objective (naive) score and Subjective (logic) score.

Objective Score calculates the completeness of entire dataset, which is taking the rows with complete information and dividing it by the total number rows, even if the completed information is invalid. This is defined as:

The Objective Score should be equal to 1 after all the NA rows are dropped. This method is very straightforward and intuitive, but its lack of efficiency and logic makes it “naive.”

Since the project mainly focuses on text mining, the removed columns are the columns with NA

in the “text” column.

The Objective Score and Subjective Score screenshots are followed:

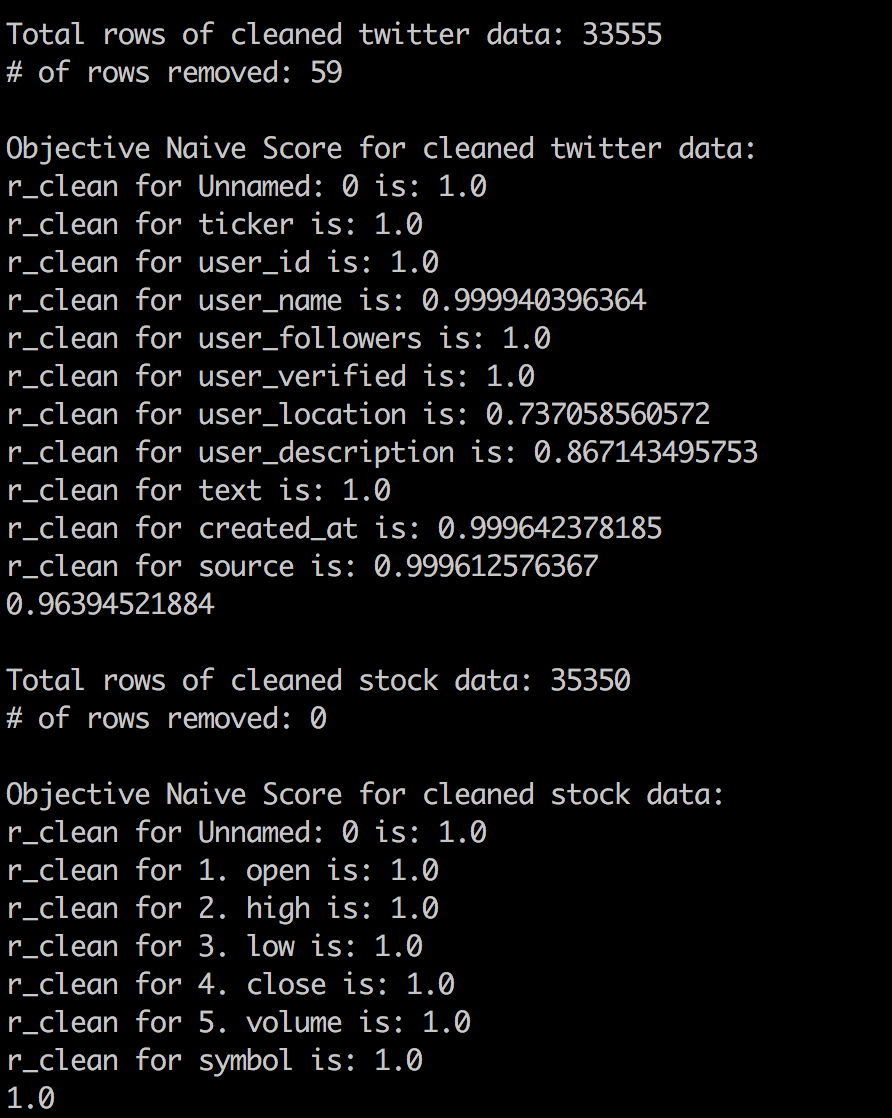
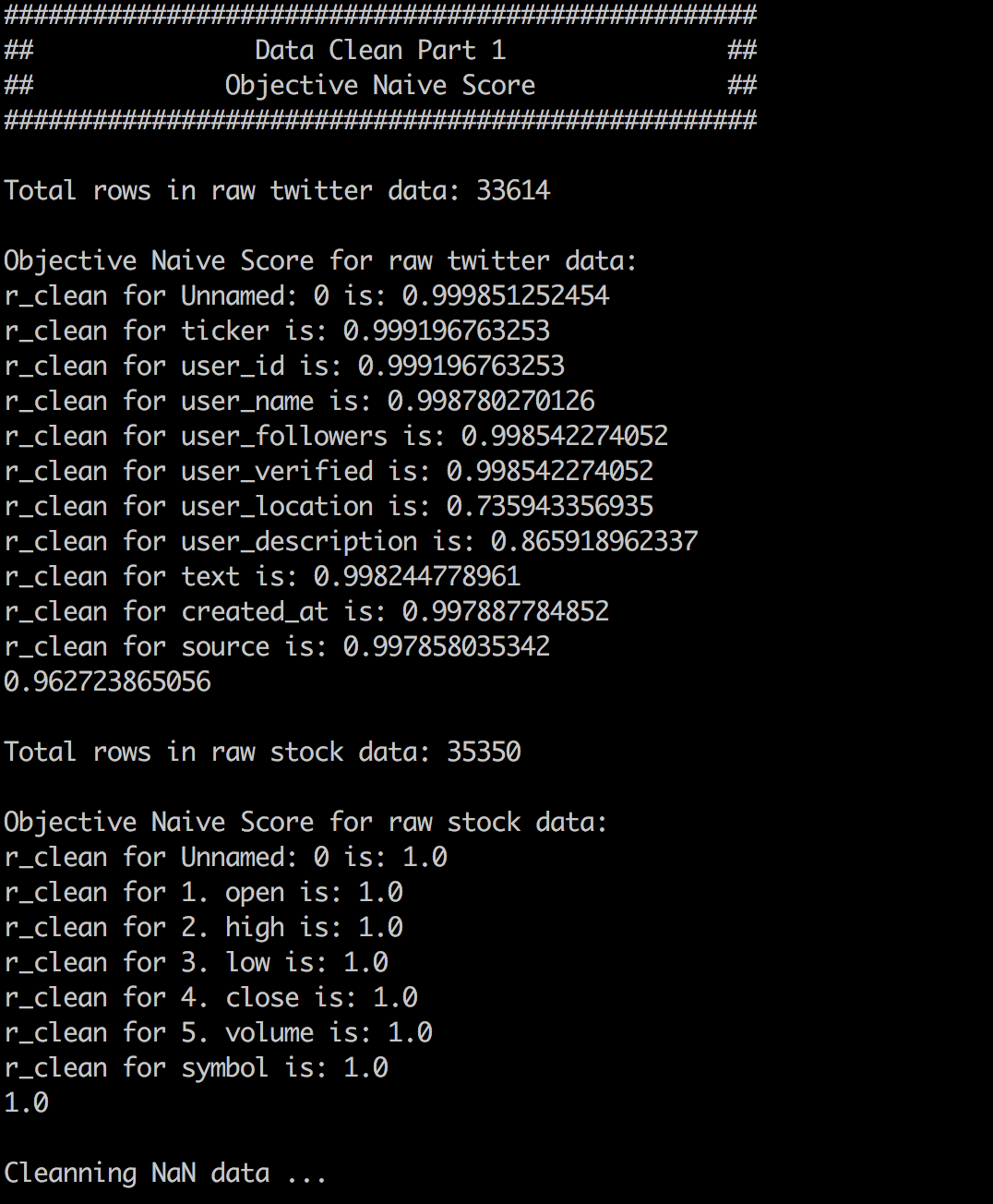


Fig 1. Objective Naive Score before removing rows with missing values [‘text’]

Fig 2. Objective Naive Score after removing rows with missing values in [‘text’]

From the result, it is obvious that the overall score has been increased after removing the rows with missing [‘text’] information. And it is also noticeable that stock raw data is very clean.

However, only dropping the rows with incomplete data is not enough. So the second metric, which is called Subjective(Logic) Score, has been introduced in order to calculate the percentage of invalid data. To increase Subjective Score, the raw twitter data needs to be filtered by some conditions such as: whether [‘text’] is formed by another language; whether [‘text’] is too short to be considered spam, whether a column is all numerical, whether a column is all text, etc.. After all the cleaning process has been done, all the rows with invalid data will be filtered out. Subjective Score should be calculated by:

After removing the rows with invalid data, the Subjective Score should also be equal to 1.

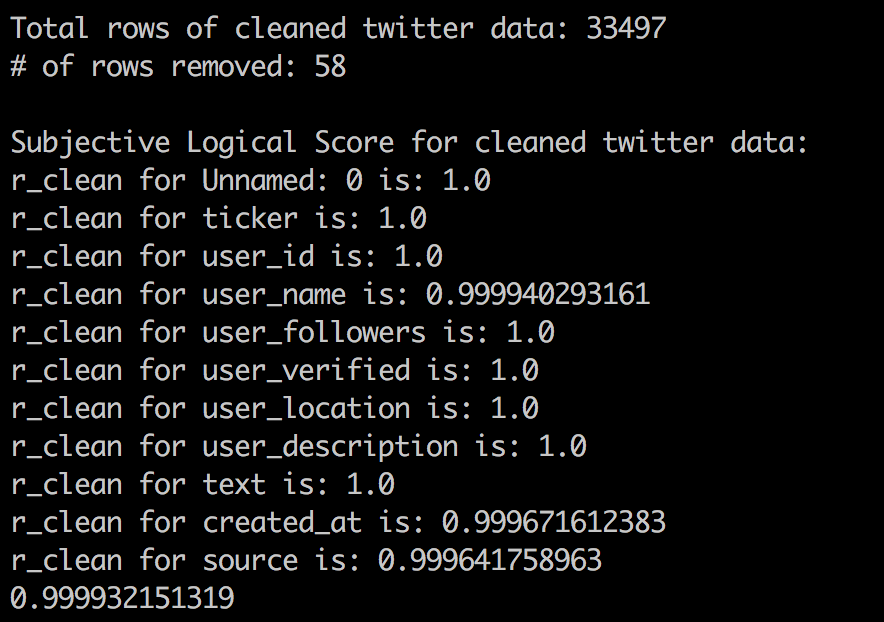
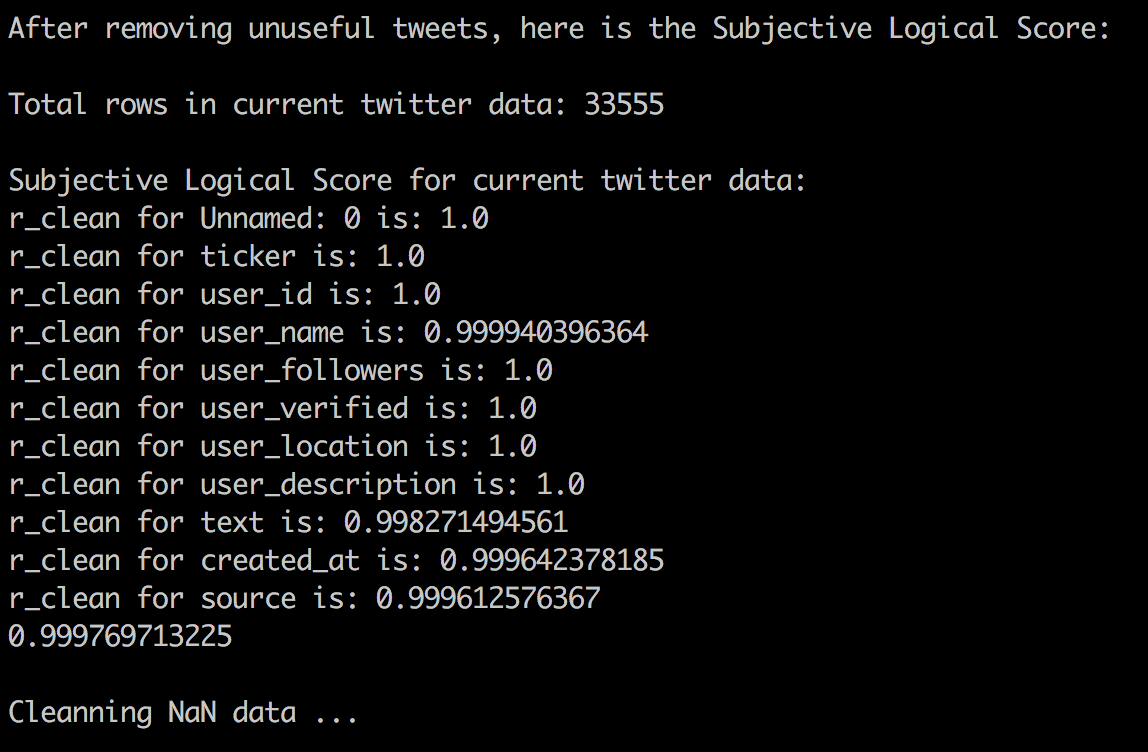


Fig 4: Subjective Logic Score after removing all N/A values for twitter data

Fig 5: Subjective Logic Score after removing all noisy values for twitter data

Citation

[1] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." *Journal of computational science* 2.1 (2011): 1-8.