Stock Market Prediction Using Sentiment Analysis: Testing The Method's Accuracy and Efficiency

Raj Patel

Department of CSE, Nirma University, Ahmedabad, India

Abstract:- **Nowadays, twitter has become one of the most prominent platform for information sharing and information exchange. It is beneficial for the people using Social Media to get relevant information about what is happening around the world. Sometimes, the information obtained is faster than the information obtained through news channels and newspapers. However, the amount of information stored inside the database can create a situation of quandary and often it becomes a mammoth task for users simply reading the tweets to extract the exact sentiment behind the tweet. If it was possible to classify the exact sentiment behind the tweet, various decisions regarding prediction and classification can be taken based on the analysis. In this paper, we will analyse the recent tweets based on the Bombay Stock Exchange (BSE) and will co-relate the results of the analysis with the actual stock and the Index prices.**

Keywords**: - Stock Market Prediction, Sentiment analysis of tweets, Sentiment analysis using R**

I. INTRODUCTION

It is the basic nature of human beings to believe what they see. Sometimes, it becomes impossible to determine that what exactly a tweet or a person who posted the tweet want to convey. Especially with a large data set like twitter in which approximately 400 million tweets are posted daily basis. Now, if it was possible to analyse a few thousands of latest tweets related to a particular topic or to a particular buzzword, and performing various operations in order to classify the tweets based on the sentiments, myriad of different decisions could be made which would lead to better management in the fields of business and big data analytics.

Sentiment analysis is a part of text mining which itself is one of the technique of data mining. It means that extracting the emotions and the sentiments of the person who posted a tweet for instance without him explicitly stating or mentioning his sentiments in the tweet.It helps us to identify whether a user has given a thumbs up or thumbs down to a particular issue concerned.For example, a user tweets concerning a review of a particular movie and his tweet goes like 'The movie was extremely awful and totally boring', the user didn't explicitly stated the word bad in the tweet but words with a negative connotation like 'awful' and 'boring' were used, looking on which we can say that the user was certainly not happy with the movie. But if we imagine the how difficult it would be to determine the sentiments of a user as the number of tweets posted everyday on twitter are in millions. After performing the sentiment analysis on the data, the data is usually classified into different categories.

Usually, classification of sentiment analysis is categorized into two types:

I. Full Classification

II. Categorical Classification

On the basis of the analysis obtained at the end of the experiment, we will classify the sentiment of the tweets into two major categories:

I. Positive

II. Negative

In the following experiment, we used the social networking platform Twitter in order to perform sentiment analysis based on the tweets that were posted on the website and consisted information regarding a particular topic or a particular subject. For example, information regarding a review of a movie or regarding what he/she is feeling regarding the stock market etc.

II. FETCHING THE TWEETS

We used R Programming Language in this project, as it is one of the most advanced and user-friendly language when it comes to statistical analysis of the data. In order to do sentiment analysis of the tweets, the first step that was required to do was fetching the tweets that are available on the twitter database.

For fetching the tweets, we need to first have a twitter account and by that username and password need to log-in by making a developer account on apps.twitter.com. After making a proper account we will get a set of different keys which we will need for getting the tweets that have been posted on the website by myriad of users.

For getting the latest updates on a particular topic, we used the '#' feature of twitter which helps in finding a particular subject in a relatively easy manner. We will require the 'twitteR' library in R in order to use the methods available by which the tweets would be fetched. In order to get an accurate idea we fetched then the recent 1000 tweets from the twitter that contained the word '#sensex' in order to get all the

recent tweets that were concerning about the Indian Stock Market.

Note: In order to get the tweets, we will first need to authenticate to our developer's account using the keys that were auto-generated after making the account on the website.

III. APPLYING DATA TRANSFORMATION AND CLEANING

Just getting the raw tweets doesn't suffice our job, as we will also need to apply different types of data transformation, data cleaning and data pre-processing techniques on the raw data in order to get a proper format of the tweets obtained.

As, we fetched the tweets and saved them in an array format in their raw form, we needed to convert them into a proper format as applying classification in their raw format would not be appropriate in order to get accurate and substantial results.

While applying transformation, we first converted this array of raw data into proper string format and then after getting the proper string format of the data, we applied the transformation of this string into word vectors so that we could search for each word in the tweets individually.

After getting the tweets into proper word vector format, we will applied data cleaning which included removing Punctuations as punctuations are completely trivial in order to classify the sentiments of the user. For that we used library defined functions of R which was included in the library 'stringr'.

We also removed the numbers from the tweets as like the Punctuations, numbers didn't help in the classification of data. Some other contents that were removed were white spaces and the word 'sensex' as once from the hashtag it is known that the tweet is concerning about the stock market, there is no need of included the word 'sensex' again in the tweets.

Though at first the task of data cleaning looks like a trivial and a time-consuming task, but for accurate and efficient results in terms of time and space complexity, it is necessary to clean the data.

IV. COMPARING AND CLASSIFYING

The next step would be comparing the words from the tweets from some word list/s by which the classification process would be carried.

For this step we used, a set of two word lists which consists of a positive word list and another consisted of a negative word list. Most of these data sets are available on the Internet. We can use any of the data sets as most of the data sets are derived from the Natural Language Toolkit (NLTK) library usually made for python.

We compared the words obtained from the tweets by the word data sets one by one. This kind of string matching algorithm takes more time but gives accurate results. The calculation that we used for classifying the tweets was as follows:

Score =((-1)*(negative_word) + (+1)*(positive_word))/(positive+negative)……(1)

In the above equation, negative_word represents if the word from the tweet matches with the word in the negative word data-set while positive_word represents if a particular word of the tweet matches with the word in the positive words data set.

For example, if a particular tweet goes like this, "The announcement of the new scheme in the Union budget seems completely useless and glib. Although it has the potential to strengthen the Indian Currency at some level. #sensex".

In the above example, if we apply our equation then we can see that there are two negative words 'useless' and 'glib' in the tweet compared to only one positive word 'strengthen', which means the score would be calculated as follows:

score = (-1)*(1+1) + (+1)*(1)/(3) = (-1)/3 = -0.3333

Like given in the above described example, we calculated the scores of rest of the tweets.

V. RESULTS AND INFERENCES

After getting the results, we calculated mean and median to have an idea about the central tendency of the data and to determine whether the overall sentiments going on the social media regarding stock market were positive or negative.

The overall scores were calculated on the scale of (-1) to (+1).

Below is the histogram calculated on the basis of the classification of tweets:

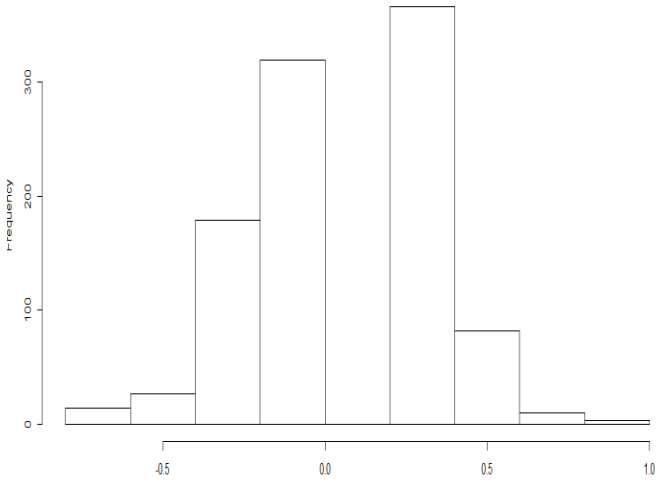

Fig 1: Histogram based on the classification of sentiment analysis

This calculations were done on 29th February, 2016 after 3.15 p.m. IST as that is the closing time for the Indian Stock market. The median score for that day was (0.018) while the mean calculated was (0.000).

Similar kinds of calculations were performed on 21st February, 2016 and 13th March, 2016. All of the above calculations were performed after the closing time of the stock market in order to get a proper idea of what will happen the next day. As through the sentiments of the people, we can get an overall idea of how will the market perform as the buying and selling of the stocks majorly depend on the minds of the people that are dealing in the stock market.

We also found of the co-relation between the mean and median of the sentiment analysis performed with the actual closing price of the Bombay Stock Exchange (BSE) also known as Sensex. The closing price was calculated on the same day on which the sentiment analysis was carried out.

The table for the co-relation between the actual price and the analysed score on the basis of the sentiment analysis is given below:

Date	Price	Mean	Median
21st-Feb-2016	23,800	0.075	0.120
29th-Feb-2016	23,000	0.000	0.018
13th-March-2016	24,800	0.16925	0.250

Table 1: Co-relation between the actual price versus the mean and median calculated at the same day.

As we can see from the above analysis that the actual prices of the Sensex varied positively with the change in the sentiments. As from the 23th to 29th February, the decrease in the sentiments that means the increase in the negativity among the people on the social media platform had some effect on the actual stock market.

Accordingly, between 29th February, 2016 and 13th March, 2016 the increase in the positivity in the social media regarding the Stock Market also affected on the actual stock prices.

Moreover, we also performed the similar analysis on other individual stocks, most of which gave a positive co-relation with the actual prices.

VI. CONCLUSION

By the sentiment analysis performed above, we can say that the fluctuations of the stock market are dependent on the sentiments of the people going around on the social media. But a precise co-relation and conclusion cannot be given on the basis of the above calculated results.

As we know, the prices of the stock market are dependent on various other factors, for instance the changes in government policies, results, budgets etc. So, only on the basis of the sentiments on social media, we cannot identify the exact future of the stock market.

Moreover, for a detailed idea about the sentiments, many other factors such the type of the user, user's history, the actual relation of the user who posted the tweet has with the stock market has to be known and taken into consideration. All of which is extremely difficult to get and therefore, prediction of the stock market cannot be done solely on the basis of a single factor.

REFERENCES

[1]. Data Mining Concepts and Techniques By Jiawie Han and Micheline Kimber
[2]. Lallindra De Silva, Ellen Riloff. User Type Classification of Tweets with Implications of Event Recognition
[3]. Ellen Riloff. User Learned Extraction Pattern for text Classification

.