T-TRADE: Twitter Thoughts Recognization and Analysis to Determine stock Exchange Value

ASHWINI MAILE[1], MOHIT SABLE[2], ROHAN SONAWANE[3], MEGHANA SONCHHATRA[4], Prof. S. C. SURYAWANSHI[5]

[1,2,3,4,5] *Department of Computer Engineering, Sinhgad College of Engineering*

Abstract—*Stock prices are highly fluctuating. It is one area where change in sentiment of people immediately affect the stock market. Twitter is one of the social media platform where people share their views. The tweets posted on twitter helps us to predict stock price movement by doing sentiment analysis of twitter data. The system uses R language to fetch and process the tweets. It uses pre-processing techniques like removing URL, stop words to reduce noise in data. The prediction is performed for the next day stock price by using twitter data and stock price gathered from yahoo finance using linear regression. Sentiments gathered from twitter are combined with historical stock price to predict future stock price index.*

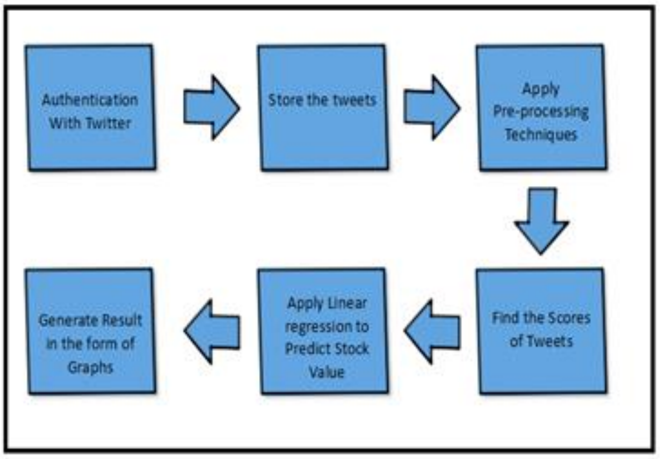
1. INTRODUCTION

Stock market prediction has been an active area of research for a long time. Stock market prices are largely fluctuating. It is one area where change in sentiment of people immediately affect the stock market . People are more likely to express their views on the social media. Comments, reviews and opinion of the people about the company on social media play an important role to determine whether a given population is satisfied with the product, services of the company or not. Abundant data on such sites provide great opportunity to understand human behaviour and thinking. Twitter is one such social networking and micro blogging site with more than 500 million tweets per day. The tweets posted on twitter can be used to predict the stock market movement .So by finding the sentiment analysis of the tweet ,stock market index can be forecasted. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing (NLP), text analysis and computational linguistics to identify and extract subjective information from the source materials. Sentiment analysis is the process of automatically detecting whether a text segment contains emotional or opinionated content, and it can furthermore determine the text's polarity. Sentiment analysis of tweets helps to understand the emotions of people. The raw Twitter dataset cannot be used directly in sentiment analysis since it contains unwanted information such as punctuations, emoticons, and misspellings. The data must also be grouped and transformed to extract relevant sentiment information.Pre-processing is required before working on such tweets . Existing system do not use pre-processing techniques. This project is based on the system which uses pre-processing techniques such as removing url's, stop-words to reduce the noise in data. The prediction is performed for the next day stock price by using twitter data and stock price gathered from yahoo finance using linear regression. Sentiments gathered from twitter are combined with historical stock price to predict future stock price index

2. RELATED WORK

Zhao Jianqiang, Gui Xiaolin proposed a Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. This paper discussed the effects of text pre-processing method on sentiment classification performance in two types of classification tasks, and summed up the classification performances of six pre-processing methods using two feature models and four classifiers on five Twitter datasets. Also evaluates effect of various pre-processing methods such as removing URL, negation repeated words and stop words. Results shows that removing stop words, numbers, URL is appropriate to reduce noise using Expanding acronyms improves accuracy. Proposed methods limits that it cannot process the punctuation marks and sarcastic sentences.[1]Sunny Kumar, Paramjeet Singh, Shaveta Rani presents Sentimental Analysis of social media using R language and Hadoop. In this paper ,Rhadoop is used for storing and analysis of big-data. Rhadoop provides huge set of packages to analyze the text or comments . Sentiments functions of R-language shows poor performance with the big data. Result shows that use of R-Hadoop increase efficiency , where the limitations of the algorithm are it is less efficient for small data and only support for batch processing.[2] Jintak Park, Henry Leung and King Ma proposed Granger causality analysis to find correlation between public sentiment and stock market movement. Twitter data is combined with historical stock price information to predict future stock price index. The minimum number of followers and daily tweets was also analyzed. Prediction is shown to have improved performance as compared to conventional methods. Analysis shows that authorative user's sentiment affects the other users and also the stock price.[3]

Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda proposed Sentiment Analysis of Twitter Data for Predicting Stock Market Movements in which they provide the way to predict stock market values. In the proposed system we are using efficient market hypothesis(EMH)states that financial market movements depend on news, tweets about company will have significant impact on a company's stock value. This system applied sentiment analysis and supervised machine learning principles to the tweets extracted from twitter and analyze the correlation between stock market movement of a company and sentiments in tweets. In a elaborate way, positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase.[4] Carlos Simoes,Rui Neves,Nuno Hortal proposed a system that predicts stock market index using tweets collected from twitter.To predict market, sentiment model is built that detects the sentiment present on tweets from different companies .Using that sentiment a trading rule was build that was optimized by a genetic algorithm to achieve maximum profit .To gurantee that no human sentiment is left behind authors adopted emotions from circumplex model of affect and their synonyms and used them as search term on the twitter API.[5]


Authentication With Twitter
Store the tweets
Apply Pre-processing Techniques
Find the Scores of Tweets
Apply Linear regression to Predict Stock Value
Generate Result in the form of Graphs

3. METHODOLOGY

A. Twitter Authentication:

The proposed system connects to the Twitter Searches API using the developer‘s username and password of Twitter Developers account. As soon as the username and password is authenticated and handshake is done with twitter API, it provides some methods through which we download the data from twitter server. The resultant stream of tweets is stored into some files that can be used for analysis purpose.

B. Collection of Tweets and stock prices.:

The tweets were collected using Twitter API and filtered using keywords related to company.Retrieved tweets data was JSON formatted. For calculating the sentiment analysis purpose, we only collect tweet id, posting time, user who posted the tweet and tweet. Stock price data is collected from Yahoo Finance using the ticker of the required company. Information being collected were the open stock price and close stock price of the companies for each day. This data were then being combined with the result of sentiment analysis to create prediction model.

C. Pre-processing techniques:

Pre-processing is required before working on collected tweets.Most researchers consider that URLs do not carry much information regarding the sentiment of the tweet. The URL matching the tokens are removed from tweets to refine the tweet content. Words that contain repeated letters are reverted to their original English form.Numbers are of no use when measuring sentiment and are removed from tweets to refine the tweet content.Removing stop words. Stop words usually refer to the most common words in a language, such as ”the”, ”is”, and ”at”. Most researchers consider that stop words play a negative role in the task of sentiment classification, and they are removed before feature selection by researchers. The classic method of removing stop words is the method based on pre-compiled lists. Removal of additional white spaces: There may be consists of extra white space in the data and it needs to be removed. By removing white spaces the analysis to be done more efficiently.

Removal of Hashtag: A hashtag is a prefixed with the hash symbol . Hashtag are used for naming subjects or phrases that are currently in trend. For example, #google, #twit-ter. During pre-processing we remove unwanted spaces, html links, punctuation marks, hashtags, numbers, etc.

D. Sentimental Analysis of tweets:

The next step is to check the sentiments of different tweets of every user. Sentiment analysis is a process to classify the polarity of given data. Sentiment analysis of tweets helps to understand the emotions of people about that company .For doing the sentiment analysis we use sentiment R package. This package attempts to take into account valence shifters while maintaining speed. Using this package number of positive, negative, neutral tweets are calculated.

E. Apply linear regression to predict stock price.:

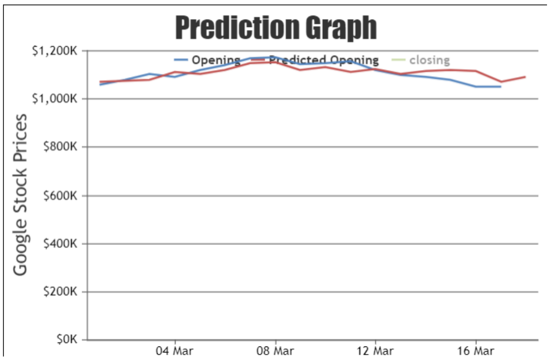
Price prediction used linear regression, since the value that was being predicted was not categorized.Linear regression is one of regression method to be used for classifying numerical class.It creates linear function by calculating weight values for each feature. The function can be seen as follow:

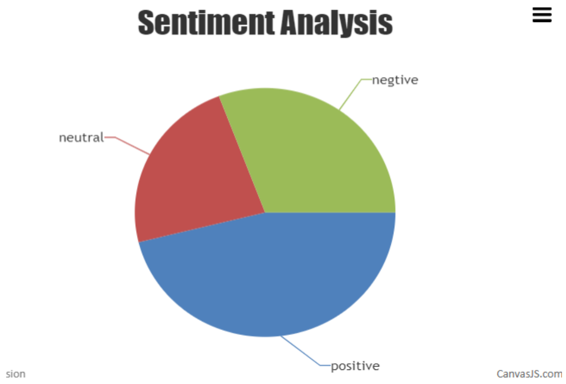
$$Y = a + bX$$

where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept. The purpose of linear regression is to predict the stock price of one company of future day using information from previous 47 days as the features. The sentiment analysis data of each day is combined with stock price and this historical data is given to linear regression for training data. The next day stock price are predicted using sentiment analysis data of tweets of previous day. Maximum fifteen days historical data is used for training.

4. RESULT AND DISCUSSIONS

The accuracy of this model is upto 60 %.Highest accuracy is achieved when tweets collected are relevant to company.It also shows that the accuracy of the model tend to decrease with increasing of number of previous day.





5. CONCLUSION AND FUTURE SCOPE

The purpose of this system is to predict the stock market index values on the basis of the sentimental analysis. The user can get through idea about the future market conditions and he can take decisions accordingly. The system takes input as tweets from the twitter data, store it into R database. The real stock price values are collected from Yahoo finance. Linear regression is used to predict the future stock price using sentiment analysis data of twitter and historical stock price collected from Yahoo finance . The result is generated in the form of graphs , tables and pie-charts. Above all, we hope to provide a comfortable user experience.

As the size of data exceeds the size of physical memory of R environment then, R gives poor results . So in this case, we can use R-Hadoop instead of R.Apache flumes can also introduced along with R-Hadoop.

REFERENCES:

[1] 'Comparison Research on Text Pre-processing Methods on Twitter Sen-timent Analysis ',IEEE Journal,2016 by Zhao Jianqiang , Gui Xiaolin.

[2] 'Sentimental Analysis of Social Media Using R Language and Hadoop: Rhadoop ', IEEE 2016 by Sunny Kumar, Paramjeet Singh, Shaveta Rani.

[3] Information Fusion of Stock Prices and Sentiment in Social Media using Granger Causality,IEEE 2017 by Jintak Park, Henry Leung and King Ma.

[4] 'Sentiment Analysis of Twitter Data for Predicting Stock Market Move- ments',IEEE 2016 By Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda.

[5] Using Sentiment from Twitter optimized by Genetic Algorithms to Predict the Stock Market ,IEEE 2017 by Carlos Sim˜oes,Rui Neves,Nuno Horta.

[6] Stock Price Prediction using Linear Regression based on Sentiment Analysis, IEEE 2015 by Yahya Eru Cakra and Bayu Distiawan Trisedya.

[7] Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm, IEEE 2016 by Huma Parveen and Prof. Shikha Pandey