ENGG5108 Big Data Analytics

Project Proposal

Group: ωετhεbεsτ

WONG Tsz Fung (1155077547)
XIA Xin (1155116404)
HUANG Ruiyang (1155130026)
WU Mengyang (1155119011)

1. Abstract

It has been demonstrated that the stock market is profoundly affected by the public mood states, showing the significance of sentiment analytics for prediction work. Twitter, as a popular social media, is then attracted world-wide researchers. Recent works on sentiment analytics are limited with single index and the region is somehow limited in Europe. Here, we plan to develop a system applied on different index or company's stock in the US to predict the stock market by sentiment analytics. Public opinions related to the trade-war between the US and China will be collected from twitter, where the distributed computing environment Hadoop and Map-Reduce algorithm will be applied for data retrieval, dataset cleanup and format transformations. Text processing tools of WNAffect, word2vec and NLTK tool kits will be employed to investigate the relation between the stock market and the public motion, and machine learning, including basic regression and neural networks, will be utilized for the prediction.

2. Motivation & Background

2.1 Background

The stock market plays a variety of roles in the economic system in the modern world, such as financing, investment, risk diversification, resource allocation, and regulation, reflecting the global economic level. In the past. Stock market movements are considered to be random and unpredictable. However, with the development of data analysis technology in recent years, many scholars have shown that investor sentiment has a significant impact on the stock market and the price of a single stock as well as the income it brings, which means investors' sentiment can be used to predict stock market returns.

Today, the methods of stock trend forecasting are increasingly being updated. And behavior has become a new hot spot for forecasting the stock market. At the end of the 20th century, the financial market found a large number of phenomena that traditional financial theory was difficult to explain, and various asset pricing models were unsustainable. As a result, behavioral finance was introduced into the financial research framework to analyze investor behavior, which improves the interpretation of the stock market. Today, it is possible to effectively analyze investor sentiment and prices of a single stock as well as its profit because of the powerful technology of internet and Big Data Analytics.

In addition, the investigator also plays an important part in the stock market. Investors are influenced by the stock market and also affect the stock market. Behavioral finance believes that investment behavior is also a psychological behavior, which will be related to investor sentiment.

Then, emotions can influence the decision of the agent significantly. investors are considered to be bounded rational, and the decisions of irrational investors are not completely random. The

emotions can influence the decision of investment. For example, if the sentiment in the society of investment, and investors tend to be optimistic.

Finally, the social media contains a great amount of investors. It means there exists a huge social network which consists of lots of investors in social media. The investment information is published and disseminated in that network, which has formed a huge amount of data, which can effectively analyze the influence of investor sentiment on the stock returns and prices.

2.2 Motivation

Investors are affected by their emotions, and sentiment of investors is affected by the stock market, then investor sentiment also affects the stock market. The interaction between investors and the stock market is a dynamic process. We can find the global stock market crashed after the financial crisis in 2008, and a large number of financial theories did not predict this result. It can be seen that the existing theory is flawed in explaining stock market volatility. As an important part of the financial market, investors play a significant role of financial market theory. The assumptions of traditional financial theory for investors have been difficult to meet the reality. So, studying the impact of investor sentiment on the stock market has obvious theoretical significance. First, it can provide a theoretical explanation for whether investor sentiment affects stock returns as well as prices, and how it affects stock returns and prices. Secondly, there are many phenomena in the stock market which can be difficult to explain by the existing financial theory. If it can be based on investor sentiment analytic, it is obviously beneficial to the supervision and management of the stock market and understand the stock market. It can also be helpful to know the cause of abnormal fluctuations and avoid risk factors.

3. Relation to course topics

Distributed file system:

Our data source is directly from twitter, produced every second and in huge volume, which meets the 4V characteristic of big data, especially in velocity and volume. Such jobs require the construction of big data analysis platform. In the data retrieval and preprocessing stage, we will make use of the distributed computing environment Hadoop. Map reduce algorithm will be applied in dataset cleanup and format transformations.

Text processing and NLP:

To analyse the relation between market and public emotion from twitter, probability or machine learning model will be built for predictions. But before the analysis or learning process, we still need an effective approach to transfer the semantic information inside each tweets to a quantifiable formation. Hence, NLP and test preprocessing tools and algorithm will be applied in our implementation, such as WNAffect, word2vec and nltk tool kits.

Machine learning:

With the vectorised twitter data and stock data from past few years, many mathematical models are in our considerations. In this stage, we will make use of different classification models learned from this course including basic regression and neural networks. Comparison will also be made based on their accuracy and efficiency.

4. Expected deliverables

Our idea is to construct a stock market prediction system by analyzing public emotion on Twitter. There are many papers showing that public mood can be an effective factor on forecasting the stock market and they can be extracted using sentiment analysis on tweets.

We would like to use Machine Learning techniques to do the prediction. This essentially means that our system has to be separated in two phases, training and predicting. In either two phases, we have to retrieve tweets from Twitter, which indicates that our system has to include a tool to retrieve tweets according to user input of specific tags and time range, as shown in **Figure 1**. We will describe the procedure in detail in session 5.

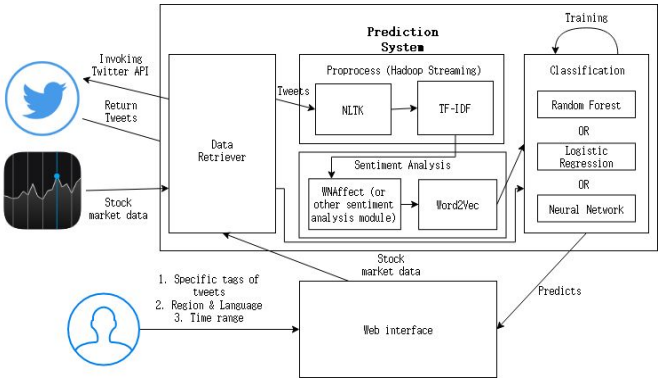

Figure 1: Proposed structure of prediction system

Then, our system is necessary to preprocess and do sentiment analysis on tweets. We planned to use tools like Hadoop streaming and NLTK for preprocessing and WNAffect and word2vec for sentiment analysis, as shown in **Figure 1**. By converting words to vectors, we can easily input them into many machine learning models, such as Random forest, Logistic Regression and Neural Network.

To simplify the problem, we will model the stock market prediction to a simple classification problem. Essentially, the input will be the word vectors, while the output will be a 0 or 1. 0 representing the stock value of predicting day will be lower than the previous day, and else for 1. We will further elaborate in session 6. After finishing the training, the system will continue to retrieve tweets and do the prediction for future days.

Since Twitter is mainly used by American people, we will conduct the prediction on US stock market. We planned to predict the Dow Jones Industrial Average (DJIA), but our system should not be limited by a single index, since it should be to apply on different index or company's stock. Afterall, the prediction result will be shown graphically for better user-friendliness, and the details will be provided in session 7.

5. Dataset & Data collection

Twitter: (Based on Twitter API)

As twitter provides useful tools for developers retrieve data automatically [1][2], we will take advantage of the official APIs to collect the tweets in corresponding days we are going to analyse. The text data will be organised by the dates and their hashtag, for example, "#Trade war", "#Tariff", "#Agriculture" or "#farmer". Because the data retrieval could be real time and automatic, we could use current data as an additional part of the data stream to boost the training.

Stock: (Based on Kaggle competitions)

There are many commercial data set about historical stock market, but most of them are expensive to access. However, Kaggle provides some open source data for their big data analysis competitions [3][4], which are free to use and meet our requirements. The data we plan to apply in our training is the historical daily price and volume data for US-based stocks. We will select some of them as examples. The data is in 'csv' format with column such as date, open, high, low, close, and volume.

6. Techniques & Algorithms

6.1 Preprocessing

After retrieving the related tweets on Twitter, we will put them into Hadoop for preprocessing.

Basically, we will make use of Hadoop streaming and NLTK package provided in Python.

By using NLTK, we can tokenize and lemmatize tweets to countable words for further processing. After that, we may use techniques like TF-IDF to remove non-important words, such as “I”, “We are”, “They” from the tweets to increase accuracy.

6.2 Sentiment analysis

According to Bollen and Mao’s paper, they proved that the “calm” mood has fulfilled the Granger causality requirement, which indicates that “calm” can certainly predicts the stock market (DIJA)[5]. Therefore, they extract the words with “calm” mood from tweets, using their frequency and construct a “calm” time series. We would like to use word vector instead of time series, because word vector will be a better representation for data to be inputted into different machine learning model.

Nevertheless, we will learn from Bollen and Mao’s paper by extracting word with specific mood as well[5]. NLTK provided SentiWordNet, which gives us a score with range [-1, 1] for each word, but it only tells us if a word is meaning positive or negative. Which is not enough for our semantic analysis, as shown in the Bollen and Mao’s paper, the predicted results using OpinionFinder are unsatisfied, such that OpinionFinder is an NLP tool which only tells positive or negative of a word. To recognize the mood represented by the words, we need to make use of packages other than NLTK.

There are other packages in Python would allow us to categorize words with specific emotions. For example, WNAffect can categorized the emotion represented by WordNet resources. It provides an emotion hierarchy for developers to recognize the mood of a word, from general to specific. For instance, the first layer of the hierarchy would be “Neutral emotion”, “Ambiguous emotion”, “Positive emotion” and “Negative emotion”. Then, for the second layer, there are some emotions like “calmness”, “joy”, “self-pride” under “Positive emotion”.

By using these resources, we can construct a sum of vectors $V_{t,m}$, such that t indicates that it is the vector of day and m is the specific mood. $V_{t,m}$ will then be normalized and each dimension is a feature of the input. .

6.3 Modeling & Training

With the input vector $V_{t,m}$, our desired output is a prediction of stock price (up / down) after the period of t days. To model this mapping function, we can formulate it as a classification problem, $f(v) -> [0, 1]$. In which, 0 indicates a price down prediction and vice versa. According to Bollen and Mao’s paper, the vector $V_{t,m}$ we used should be lagged three days, so it should predict stock market value S_{t+3} [5]

To achieve this, random forest, logistic regression and neural network could be applied as potential solutions. During our training, the stock price dataset will provide the labels to supervise the training. As the data stream could be updated with the passing time, online learning or reinforcement learning could also be taken into our consideration. We will choose the model with the best performance in our final demonstration.

6.4 Prediction

In the prediction stage, we will regard a previous date as the pseudo "current day", and inference the results with our best trained model. The prediction should include the trend of stock price (up / down) in the following market day. The result will be demonstrated with our GUI for visualisation and an average rate of successful

prediction will be calculated with a selected set of testing days.

7. Demonstration methods

To make the result of analytic easily understood, we will use a website to show the trend of stock we predict by analyzing the data of sentiment of users from twitter.

Amcharts is a gorgeous tool which is written in JavaScript and HTML5[8]. Amcharts can draw the beautiful chart for the input data. Then the file formats Amcharts support can be .xml and .csv. It can also use the dynamic data. So, we choose Amchart to make our prediction visible and add some extra contents that fit our demands.

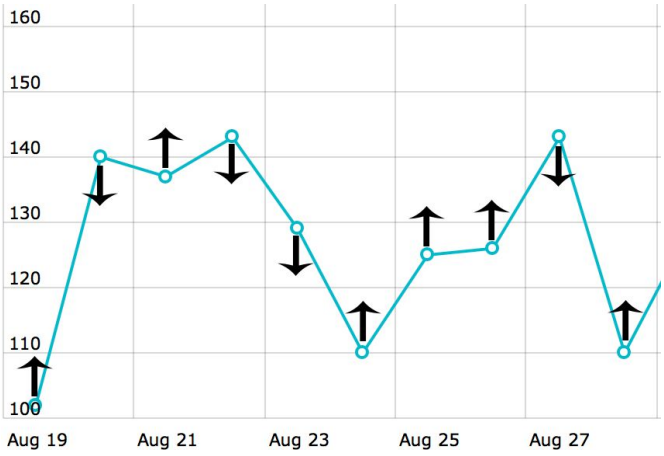

Figure 2: Anticipative demonstration of result

We add an arrow on each point in the chart, as shown in **Figure 2**, showing the trend of the stock to express the result we predict by analyzing the data from users of twitter. The up arrow means in the next time, the value of the stock will be more, or it will fall in the next time.

As a result, we can see the accuracy of our forecasting by analyzing the data from users of twitter, when the arrow is in a point which is in the past.

8. Related works

Twitter, as one of the most popularly-used social online social media, provides an open platform for people to share their emotions about topics they concern, which may impact individual behavior and the stock market as well. Therefore, sentiment analysis through Twitter is an attractive research field for prediction of stock market movements.

To investigate whether public mood states collected from large scale Twitter feeds are correlated to the Dow Jones Industrial Average (DJIA) value or even predictive to it, Bollen and Mao[5] used a Self-Organizing Fuzzy Neural Network to demonstrate their hypothesis, using the tools of OpinionFinder and Google-Profile of Mode States (GPOMS) over time. They measured a six-dimensional daily time series of mood states (Calm, Alert, Sure, Viral, Kind and Happy), demonstrating the ability to predict changes in DJIA value from public mood collected from tweets, while the results showed that only the mood dimensions of Calm and Happiness measured by GPOMS show the potential of predictive effect, compared with their counterparts measured by OpinionFinder tool.

Based on the demonstration of the correlation between the public mood collected from twitter and the DJIA value, Pagolu[6] conducted sentiment analysis to observe how well the correlation between the changes in stock prices of a company and the public mood states in tweets, with the textual representation methods of Word2vec and N-gram employed. A sentiment analyzer was developed to judge the type of collected sentiment on the basis of machine learning, finding that the stock price is strongly correlated with the public opinions from twitter and an accuracy up to 71.82% was shown.

Nisar[7] applied a short-window event study to evaluate the relationship between politics-related sentiment collected from twitter, and FTSE 100 movements in UK. Their results demonstrated the potential to forecast market movements by politics-related sentiment analytics on twitter in a short term. By regression analyses, a stronger causation between mood states and closing price of the stock market was found as well.

Our work will focus on the prediction of US stock market, by analyzing the sentiment related to the trade-war between the US and China, collected from twitter. Unlike former works, our system will not be limited by a single index, since it should be to apply on different index or company's stock.

9. Timeline

Time	Milestone
Week 3	Determine the topic and dataset Literature Review
Week 4	Determination of the Techniques and Algorithms Write the proposal
Week 5-7	Collect and format both datasets, build Hadoop preprocess pip line.
Week 8-9	Implement machine learning model.
Week 10-11	Implement the demonstration part and finalized the project report.

10. Reference

[1] Socialsensor. (n.d.). socialsensor/twitter-dataset-collector. Retrieved from https://github.com/socialsensor/twitter-dataset-collector.

[2] About Twitters APIs. (n.d.). Retrieved from https://help.twitter.com/en/rules-and-policies/twitter-api.

[3] Your Home for Data Science. (n.d.). Retrieved from https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/download.

[4] Two Sigma: Using News to Predict Stock Movements. (n.d.). Retrieved from https://www.kaggle.com/c/two-sigma-financial-news/data.

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

[6] Pagolu, V. S. , Challa, K. N. R. , Panda, G. , & Majhi, B. . (2016). Sentiment analysis of twitter data for predicting stock market movements. *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Paralakhemundi, 2016, pp. 1345-1350.

[7] Nisar, T. M. , & Yeung, M. . (2018). Twitter as a tool for forecasting stock market movements: a short-window event study. *The Journal of Finance and Data Science,* S2405918817300247.

[8] amcharts. (n.d.). amcharts/amcharts4. Retrieved from https://github.com/amcharts/amcharts4.