CSCI5408 - Data Mgmt, Warhsng Analytics

Stock Market Prediction

Project Proposal

Yunfei Guo, B00777458

Zhenbang Wang, B00769297

a. Role of each team member:

1. Yunfei Guo, B00777458: (Data Engineer) data gathering and processing, data cleaning

2. Zhenbang Wang, B00769297: (Data Analyst) model development, model testing

b. Project objective:

In this project, we are trying to predict the future price of a company stock by analyzing users' tweets sentiment. Stock market is full of uncertainty and is affected by many factors [1]. Stock market prediction is very important for finance and business investors. We aim to help traders make decisions on stock market investments. Traditionally, there are two main methodologies to predict stock market: one is using fundamental analysis, another is technical analysis [2]. In this project, we will mainly focus on technical analysis.

We can get information of the stock from some financial channel. As Twitter has been a platform for users to express opinions about stocks, for example, the "$AAPL" tag refers to the Apple stocks. Therefore, we can retrieve financial tweets for stocks of interest. For example, we will retrieve many real time financial tweets of one company, and by analyzing the sentiment of the tweets, we synthesize the scores of the tweets, then make the prediction of the stock of this company. At the very end of the project, we will use the visualization tool to find a parameter to describe the relationship between the probability and the score.

c. Programming language, tools, and resources required for the project:

For this project, we are using Python as programming language. Python can provide multiple library which can help us in coding, so it would be the best choice. Twitter Search API will be used for this project. We also use Apache Spark for tweets streaming and building the classifier model. Because the Apache Spark can perform a very fast speed in getting the related tweets and provide multiple choice of classifier such like LR and MVC. We plan to use AWS EC2 to deploy our project and running on the cloud.

Also, we will use the ‘tableau’ to visualize the relationship of the price of the stock and the sentiments we get. By observation, we will try to find the most accurate parameter of the probability of rising and falling with the sentiment score.

d. Work breakdown Structure:

For this project, we conduct Agile methodology and our work are distributed as following: (Estimation is based per hour work)

ID	Backlog/Story	Estimation	Priority
1	Setting up Github and Git flow	1	1
2	Setting up Apache Spark environment	2	2
3	Get related tweets of one company	3	8
4	Clean the tweets and remove noise	5	5
5	Text processing for sentiment analysis	7	5
6	Train the classifiers	8	6
7	Build model using machine learning	10	8
8	Train model using training dataset	6	7
9	Test model and predict stock market	8	6
10	Visualize results using tableau	3	2

e. Value proposition:

Stock market prediction is difficult, as stock price is affected by many different variables. Many researches have focused on the fundamental analysis, which analysis the company’s past performance and its credibility of accounts. Our project looks at this problem at a different angle. With the help of big data technology, we propose a method to predict the future price of a stock based on the people’s opinions, emotions and feelings about a company.

We are not conducting sentiment analysis on general tweets, but on tweets which are related to stock market, which will make our prediction more precise.

f. Milestones/Sprints:

Milestones	Content	Achieve goal
1	Data gathering and cleaning	Gather the tweets that are related to stock market and get rid of noise.
2	Setting up ETL job	Set up our project pipeline: create Apache Spark project to process python stream; load trained model to Spark; label each tweets and export results.
3	Model testing	Use our model to predict real time stock market, and modify our model based on results
4	Analyzing through visualization	Find relationship between stock price and tweets sentiment score
5	Project presentation and submission	Present final project results; submit all codes and report.

References:

1 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7373006

2 https://en.wikipedia.org/wiki/Stock_market_prediction