China Concepts Stock Volatility Analysis

by

# Yikai Feng

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business Honors Program

NYU Shanghai

May 2018

Professor Marti G. Subrahmanyam Professor Yuxin Chen

Professor Yiqing Lu

Faculty Advisers Thesis Adviser

Table of Contents

[**Abstract** 3](#_Toc514349743)

[Introduction 4](#_Toc514349744)

[Background 4](#_Toc514349745)

[China Concepts Stock 4](#_Toc514349746)

[Online Stock Forum 5](#_Toc514349747)

[Regression with Multiple Features 5](#_Toc514349748)

[Research Purpose 7](#_Toc514349749)

[Approach & Model 8](#_Toc514349750)

[Multiple features regression 8](#_Toc514349751)

[Market Volatility Context Influence 8](#_Toc514349752)

[Model Establishment 9](#_Toc514349753)

[Results & Analysis 12](#_Toc514349754)

[Difficulties & Future Work 17](#_Toc514349755)

[References 18](#_Toc514349756)

[Tables 19](#_Toc514349757)

**Abstract**

In this research, we planned to target on the volatility analysis of China Concepts Stocks, especially in technology industry. We started from a hypothesis that the volatility of certain China Concept Stock was more related to the financial hotspots and the public’s sentiment view in China than the overall volatility in US stock market environment, and by analyzing word power, building regression for prediction, the test on this hypothesis could be addressed. We proposed a multi-feature regression model to analyze the significant words in posts from the online financial forum in China. By generating the word dictionary and computing the weight on each of the featured words, we could draw the correlation between each featured word and the volatility of the stock. We trained the model to converge and obtained the regression equation of predicted stock price. In the overall environment context of US market volatility, we observed a positive correlation between NASDAQ Internet Index and certain China Concepts Stocks and used this as a baseline to test the performance of regression estimation. We drew a conclusion that online posts power was strong and could serve as a factor to determine the changes in stock prices.

Keywords*: China Concepts Stocks, Online words, Multi-feature regression*

Introduction

In this paper, we looked at the China Concepts Stock volatility. We made two hypotheses: 1) By analyzing online post, we could conclude that words online influenced the stock price. 2) The volatility of certain China Concept Stock was more related to the financial hotspots and online related posts in China than the overall volatility in US stock market environment.

# Background

The main concepts introduced in this paper were China Concepts Stock, Online Stock Forums, Regression with multiple features.

## China Concepts Stock

China Concepts Stock is a series of stocks that trade overseas and are listed on global stock markets including the NYSE, NASDAQ, and NYSE MKT. Companies operate in China having significant assets while listing their stocks in oversea stock markets are seeking to gain foreign investor capital. In recent decades, China Concepts Stock of technology companies are rising in a rapid speed. From year 2005 when Baidu listed its stock BIDU on NASDAQ, to year 2017 when we saw the dramatic increase of the value of Alibaba’ stock BABA, technology China Concepts Stock has thrived in the US market. Since these stocks are in a newly developing area and they accept a hybrid influence from both China and United States, not many researches are done regarding to the promising future of them. It is crucial to have a volatility analysis in interpretation of these specific stocks to benefit future investors and the company managers themselves.

## Online Stock Forum

As stock market is usually capricious and unpredictable, financial securities portal websites collect financial data through stock markets and companies’ reports and form a platform for people who have interest on financial market to discuss and predict the trend of stock. These forums usually contain many subjective information as people post their ideas about the stock before events actually influence the stock price happen. However, financial hotspots and domestic news will spark a flurry in these online forums and the discussions will influence user behaviors that will eventually cause effects on the stock changes. This kind of influence can be especially observed in technology companies (i.e. Tesla), where the public valuation of the company largely comes from the expectation and predictions of the company’s future performance.[1]

## Regression with Multiple Features

To interpret a large dataset, one usual and effective way is through regression. Simple linear regression takes single feature to predict the response variable :

However, in a multiple features context, the calculation is more complex.

Gradient Decent We consider using Machine Learning (ML) to deal with the training of multiple-feature regression with the method of gradient decent (GD). The entire setting is a supervised learning where label is given according to every group of . The goal of this supervised learning is to minimize a loss function , where the loss is defined to be the difference between predicted value and the ground truth value that uses a parameterized function to average on samples. [3]

The process of finding the local minima is to update the weight that is assigned to each feature in each iteration,

where is the learning rate which discouts the value that is used for updating. After several epoches of training process, the loss will be sufficiently small to have the equation reach a convergence.

Stochastic Gradient Decent The stochastic gradient descent (SGD) algorithm is a computational simplification of GD, where in each update iteration, it only takes a sample from the dataset to estimate the gradient on it.

It can efficiently deal with large datasets and usually achieve a fast convergence under sufficient regularity conditions where the learning rate is .[4]

From a finite traning dataset, the SGD picks random samples at each iteration. The number of iterations is usually set to be larger than the size of the training set, so that the average traverse of the data can cover most of the data points in the training set while going through some of them multiple times to directly optimize the expected risk.

# Research Purpose

Many researchers think that China Concepts Stock will have a promising future based on the overall optimistic market value they are presenting nowadays. However, since the performance of these stocks have unclear basis, we can hardly tell in which context the trend of its volatility is following. Therefore, an analysis on the volatility of China Concepts Stock is highly applauded.

        This research is aiming to find a way to identify primary factors that influence the volatility of China Concepts Stock and, in a subsequent step, analyze the factors and compare the influence power across factors.

The way that this research approaches the problem is to combine the quantitative analysis and qualitative analysis. The correlation between numbers can tell the influence, so can sentiment context that leads the trend of a change on stock price. By comparing the two effects on the stock price, the factors that influence China Concepts Stock can be interpreted. Therefore, it is crucial to analyze on both sides to integrate information from real-time stock data and from the public’s sentiment view.

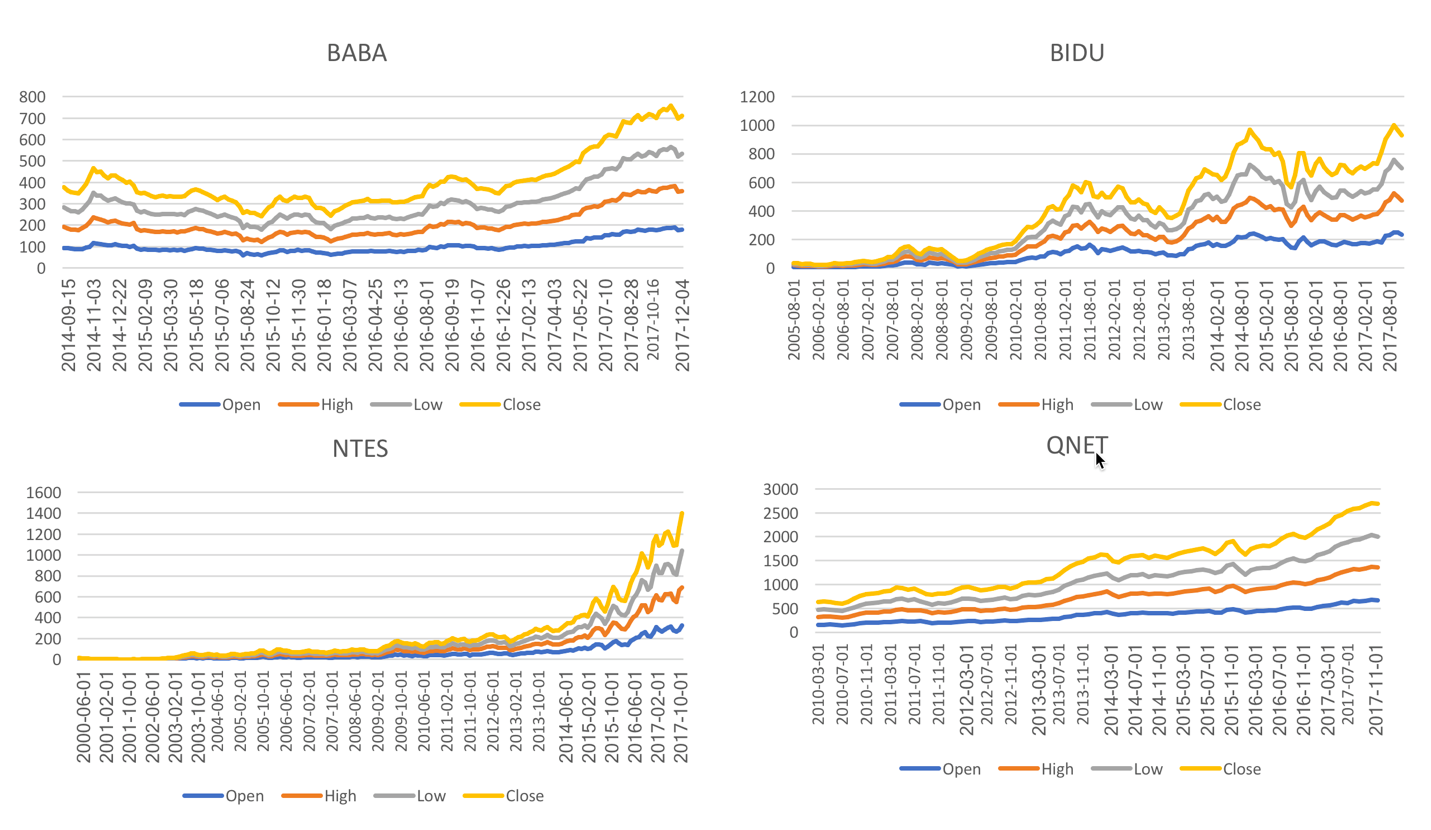
# Approach & Model

To achieve the goal, this research focused on two aspects: conducting multiple features regression for online posts and stock price changes and identifying correlation from the US market volatility context.

## Multiple features regression

To build a multiple feature regression, we used the SGD algorithm. The regression returned the predicted stock price from the post words. We formed a multiple feature matrix () containing the significant words in the posts and represented each post in this form. Then to predict the stock price, we assigned a weight (parameter) matrix to the feature matrix and computed the weighted sum over each post representation. By comparing to the ground truth stock price, we could see the error of this regression and took a next step to improve it.

## Market Volatility Context Influence

****

We used the NASDAQ Internet index (QNET) to represent the overall market influence. With the comparison above, we could roughly see the trend of stock price is of positive relation with QNET. Therefore, we used QNET chart as the baseline for our estimation model.

## Model Establishment

To apply a multiple features regression analysis on the posts data corresponding to the certain stock, we established a neural network model based on SGD method.

Dataset

For this research, we obtained datasets of Alibaba Group Holding Limited (BABA), NetEase Inc. (NTES), and Baidu, Inc. (BIDU). We focused on these three internet companies to draw the correlation between the hybrid influence. Also, NASDAQ Internet Index, published on March 1st, 2010 and corresponded to the internet industry, was my main focus on serving as a context influence on volatility.

        We obtained posts collections related to these three companies using “Bazhuayu”, a web crawler from Guba, which is an open online discussion forum established by East Money Information Co.,ltd. for users to react to instant events happened on certain stocks*.* [2]

For Alibaba, we obtained 6114 posts ranged from June 27th, 2014 to May 7th, 2018 as the dataset. Then we obtained the close price of BABA stock from NASDAQ within the according date range as the label set. We did the same for the NetEase (464 posts ranged from May 21st, 2014 to May 8th, 2018) and Baidu (1090 posts ranged from May 20th, 2014 to May 8th, 2018).

Since we observed that the datasets for NetEase and Baidu were relatively small, we decided to apply the model on Alibaba dataset for the purpose of eliminating bias.

Data Process

Since we got 6114 posts from Alibaba forum, we were considering using them as the primary dataset to apply our model. To process the data, we first tokenized the posts into Chinese phrases and words using “Jieba”, a classic sentence segmentation tool especially for Chinese. Then we obtained a dictionary of words that appeared 15 and more times in all posts. This dictionary served as the feature matrix to represent each post. We normalized the input dataset of posts to a binary representation according to the feature matrix. That is, if the word in the post appeared in the feature matrix, then the value of that word would be set to 1, otherwise, the value of the word would be set to 0. After normalization, we got a dataset with binary representation of posts as input (i.e. 6114 vectors for Alibaba).

For the label set, we matched the close price of stock with the post by the date. That is, posts published on the same day would have the same close price of that day. Since Alibaba listed its stock on September 19th, 2014, the posts before were all labeled with the close price on that day.

Model training

The goal of the training process was to approximate a multiple variable regression equation that could to some degree interpret the stock price changes. Therefore, the loss function we defined here was the Mean Square Error between the predicted stock price and the ground truth stock price, which we sent into the network as labels.

At first, the model initialized a random matrix of parameters as the weights for each feature in the data, (i.e. a random vector of size 494 for Alibaba data).

Then during each training epoch, the model performed two actions: *forward* and *backward*. In the *forward* action, the model computed the weighted sum over the input data and calculated the MSE as the loss between the predicted value and ground truth. In the *backward* action, the model back-propagated the calculated loss at a discount and updated the weight matrix with the gradient of the loss.

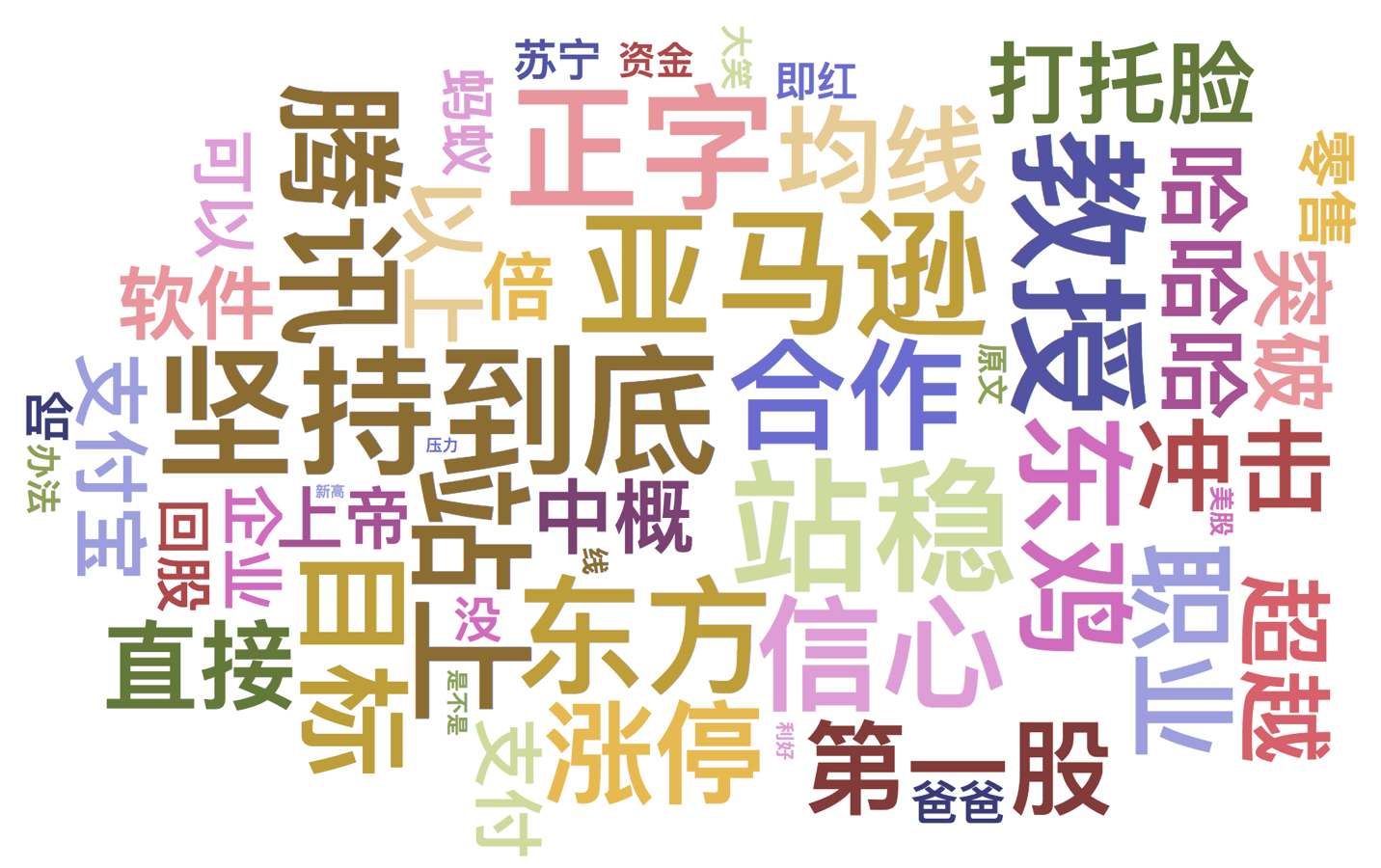
At last, through this process, the model found the local minima on the loss as the weight matrix approached convergence. We could tell the weight (or score) associated with each feature by simply looking at the matrix that already converged.

# Results & Analysis

**Related Words Extraction**

In the original dictionary dataset that we generated from Alibaba forum dataset (figure 1 below), we saw a lot of unrelated phrases, such as tone particles: 呵呵(hehe), 嘛(ma). Although these words are usually considered useless in context understanding and representation, the frequencies of these words are relatively high since people are accustomed to adding them to the end of the sentences. Instead of considering these words as stop words and tokening them out, we were interested in if they had implicit relation with the stock price. Therefore, when running parameter training, we took these words into consideration and saw if their weights were automatically lowered during the process so that we would not have the primary part of weight wrongly assigned to these words.

*Figure 1. Dictionary contains words frequency15 in all Alibaba related posts*



*Figure 2. Words weighted top 50 in Alibaba related posts*

After training the SGD algorithm, we got the words with top 50 weighted scores and were most related to the changes in stock price. The previous problem of tone particles was also addressed here as the top 50 words contained almost none of the unrelated phrases.

As we could see from the word cloud visualization, the phrases like “支付宝” (Alipay), “蚂蚁” (Ant Financial), “中概” (China Concepts Stock), “教授” (the informal name people usually call Jack Ma) directly defined the characteristics of Alibaba’s stock BABA. Besides, it contained a lot of positive words like “突破” (break through), “合作” (corporation), “站稳” (stand firm), “超越” (exceed), “涨停” (rise by the daily limit of 10%), which represented the positive attitude the public held towards the stock. Other words like “亚马逊” (Amazon), “腾讯” (Tencent) and “苏宁” (Suning) were the competitor companies that had some similar business sectors to Alibaba. This was rather interesting since people’s comparison between the similar companies’ stocks with Alibaba’s was also a factor influencing Alibaba’s stock.

We ran the same model for NetEase and Baidu, and got the results as follows:

*Figure 3. Words weighted top 50 in NetEase related posts*

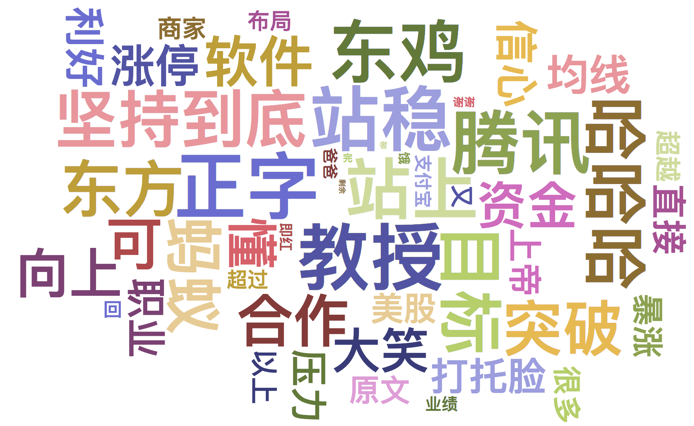
*Figure 4. Words weighted top 50 in Baidu related posts*

In the NetEase word cloud, it contained words about its released games. NetEase is opening a big market on game apps recent years, we could see that people’s attention on the game sector of NetEase’s business was influencing its stock price.

In the Baidu word cloud, we saw some words related to technology as this is a hotspot of Baidu in recent years to attract investors. Therefore, the attention on technology could serve as a factor that influenced stock volatility.

**Decreasing and Increasing Trend Divisions**

To see the differences of related words between decreasing trend and increasing trend, we divided the dataset of Alibaba into two parts. The decreasing part only contained the data where the close price fell compared to the price on the previous day, and vice versa. The result was as follows:

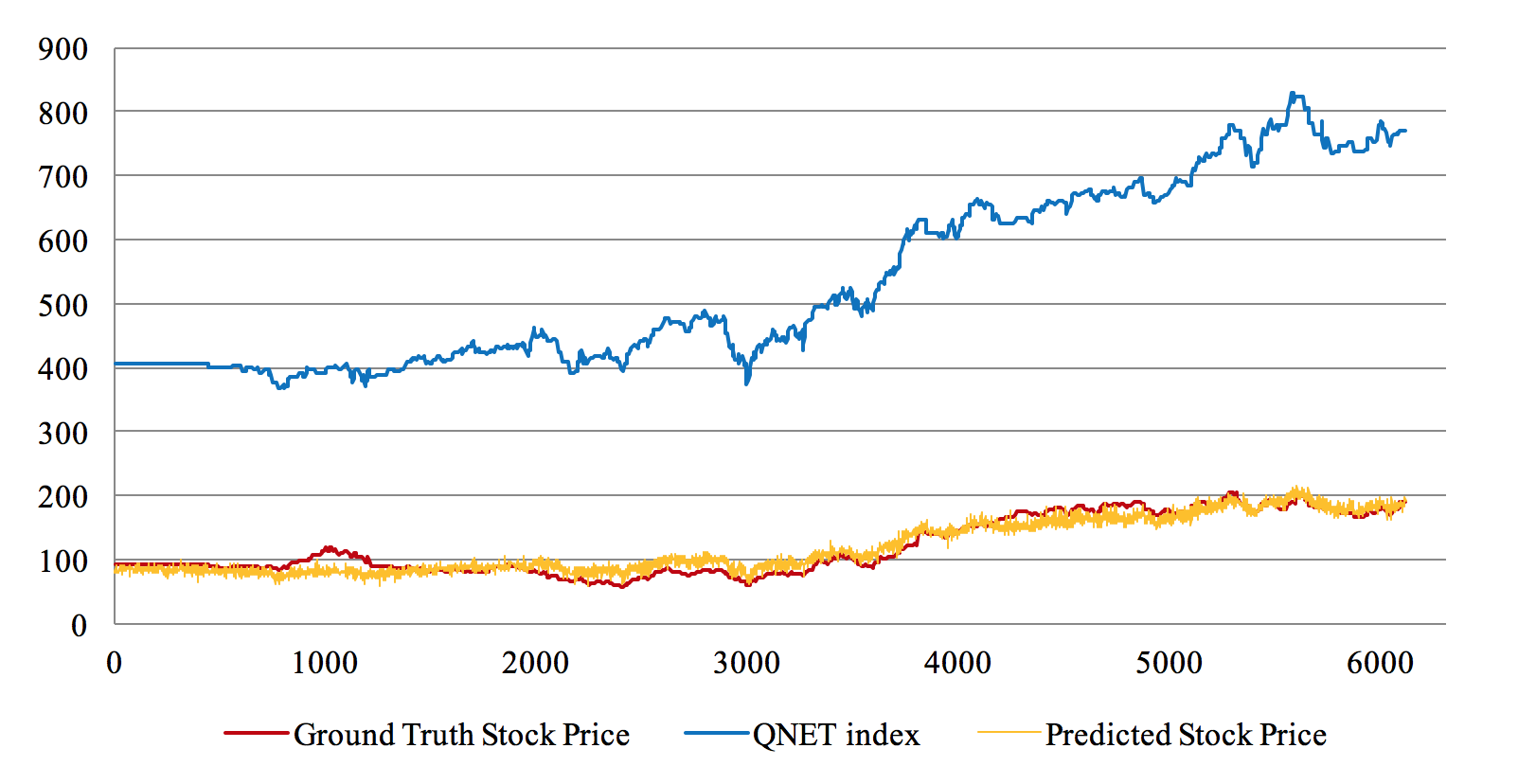
 

*Figure 5. Words weighted top 50 in decreasing dataset Figure 4. Words weighted top 50 in increasing dataset*

Though some of the words remained in the similar position, the weight of them varied (refer to the Table1 & 2 in the appendix). Overall, the increasing trend was related to more positive words. However, there was not many differences that could be told from the division. The reasons might be: 1) the amount of negative information dataset was not sufficient for producing a change 2) because of information delay, the influence of posts would be mixed together and could not be told apart from the divisions.

**Comparison between “Words Power” and Context Influence**

After collecting the weights on significant words, we formed an equation of predicted stock price.



We could see from the chart that the predicted stock price approached the ground truth better than the NASDAQ Internet Index. Thus, we concluded that the word power from the online posts was relatively stronger than the influence of the context power.

# Difficulties & Future Work

The difficulties we faced during this research was that the influence of real-time events on stock volatility might have delayed effects. We used the time lag of one day in this research. However, we usually could not tell how long it took for a post to produce influence. Therefore, we need to take the time gap as a factor and try other ways to determine this factor.

Regarding the future work, since there is a lot more to do in improving the performance of a function approximator, we can explore more algorithms and methods of regression. One research direction is to incorporate Recurrent Neural Network (RNN) as a tool to interpret the relation, since in RNN, the information will be stored in hidden states every time when passing to a new iteration. Therefore, in our case, we may also take the relationship between posts into account. This is still open to further exploration.

References

[1] Bollen, Johan, and Huina Mao. “Twitter Mood as a Stock Market Predictor.” *Computer*, vol. 44, no. 10, 2011, pp. 91–94., doi:10.1109/mc.2011.323.

[2] Liu, Yifan, et al. “Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis.” *Advances in Artificial Intelligence: From Theory to Practice Lecture Notes in Computer Science*, 2017, pp. 192–201., doi:10.1007/978-3-319-60042-0\_22

[3] Bottou, L´eon. “Large-Scale Machine Learning with Stochastic Gradient Descent.” *Proceedings of COMPSTAT'2010*, 2010, pp. 177–186., doi:10.1007/978-3-7908-2604-3\_16.

[4] Murata, Noboru. “A Statistical Study of On-Line Learning.” *On-Line Learning in Neural Networks*, 1998, pp. 63–92., doi:10.1017/cbo9780511569920.005.

Tables

Table 1

Words and weight related to the decreases in stock price:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Words | 教授 | 正字 | 站稳 | 腾讯 | 东鸡 | 站上 | 目标 | 哈哈哈 | 坚持到底 | 东方 |
| Weights | 42.41 | 38.38 | 34.37 | 33.06 | 31.78 | 31.00 | 28.07 | 26.42 | 25.94 | 25.46 |

From the table above, we filtered out words by their sentiment meaning that are more related: 教授 represents Jack Ma, 正字represents positive number, 站稳 represents stand firm, 腾讯 represents Tencent, 目标 represents goals, 坚持到底 represents hold on straight to the end.

Table 2

Words and scores related to the increases in stock price:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Words | 站稳 | 坚持到底 | 亚马逊 | 东方 | 冲击 | 教授 | 第一股 | 信心 | 全力以赴 | 站上 |
| Weights | 41.13 | 36.42 | 35.46 | 33.68 | 31.93 | 30.42 | 27.95 | 27.87 | 26.97 | 25.70 |

From the table above, we filtered out words by their sentiment meaning that are more related: 站稳 represents stand firm, 坚持到底 represents hold on straight to the end, 亚马逊 represents Amazon, 冲击 represents impacts, 教授 represents Jack Ma, 第一股 represents the best stock, 信心 represents confidence, 全力以赴 represents do one’s best, 站上 represents stand upon

Note: The predicted value of stock price is given by the weighted sum so that the higher the weight is assigned to the word, the more related the words is to the stock price change.