Stock Market Forecasting based on Sentiment Analysis

Final Project of LING 131

Yuehuan Zhang, Ran Dou

Xuanbo Mao, Ilsa Wu

December 2019

0. Introduction

In this article, we are initiating research on sentiment analysis and investment forecasting based on current Chinese stock market. And we are mainly going to discuss four aspects, including research background and content, construction of sentiment-analytical models, construction of time-series predicting models and summary.

1. Research Background

As we all know, the stock market is a barometer of the nation's economy as well as a significant component of the country's market economy. It is especially susceptible to the influences of policies, news and public's opinions and therefore fluctuates violently. It is quite necessary for us to pay attention to Chinese stock market under huge changes occurred in Chinese domestic and international economic markets. Moreover, with the development of media, people are inclined to publish, disseminate and exchange information through the Internet. Those online stock comments are enriched with valuable financial information and investors' sentiment "messages". Online stock reviews possess the characteristics of fast dissemination, a wide coverage, huge

influence and accurate real-time response. Not only ordinary investors but also financial professionals select online stock community to express their opinions about Chinese stock market. Therefore, at a certain point, online stock reviews are capable of reflecting the whole society's opinion on the stock market in China. In terms of sentiment analysis to quantify stock investors' sentiment, the majority of studies concentrates on employing indirect indicators to reflect investors' sentiment toward the stock market, such as closed-end fund prices, market resale rates, etc. There are very few of them based on direct investor sentiment indicators. This phenomenon brings up an interesting point and inspires us to conduct this project, that is, to perform sentiment analysis and extract investors' sentiment index through online stock comments and then apply those sentiment indexes as an exogenous variable to build the ARMAX model for the sake of studying the relationship between investors' sentiment and time series in the stock market.

2. Research Contents

Our research on Chinese stock market will start from mining and analyzing those investors' reviews toward the stock market. In addition, according to Behavioral Finance, the factors that affect stocks' prices and performance include not only their intrinsic value but also investors' psychology and behaviors. This creates a theoretical basis for our data mining on investors' stock comments. And the rise of the techniques, including Text Mining, Machine Learning and Time Sequence models, has made our mining and analysis on stock reviews to be possible. We are aiming to apply those techniques into our research. In this case, our research content is to implement sentiment analysis based on investors' stock reviews and conduct sentiment indicators corresponding to those reviews. And then we are going to establish a time sequence model in combination with stock prices to predict trends of Chinese stock market. This study consists of two significant parts.The first one is to establish a sentiment-analytical classification model to achieve the rapid recognition and judgment of investors' sentiment tendency from the reviews. The other mission is to construct a time-series forecasting model to predict stock market performance so that providing a reference for investors' selection on their investment strategy in

Chinese stock market. Before moving to those two tasks, we have to first collect investors' reviews from Chinese online stock forums.

3. Data Preparation and Data Handling

Regarding the data acquisition, we select the section "SSE Composite Index Bar (zssh000001)" from Guba Stock Comments, a Chinese online stock community for real-time market comments and stock exchange, to extract stock review data. For this purpose, we applied Python to design a crawler program and eventually grabbed 1,300,000 reviews in total, which were posted by investors between 08/06/2018 and 12/17/2019. Furthermore, we manually marked 20,000 pieces of total stock comments with the labels "-1", "1" and "0", which each respectively represents investors' bearish (negative) comments, bullish (positive) comments and neutral attitude toward the future direction of Chinese stock market. (An example for the labeled stock review data is in Appendix I). We also concluded a list of sentiment indicators to help us judge whether the messages delivered by investors is positive, negative or neutral. The table (Table 1) below is part of indicators. Then, we divided the entire data set into 2 files, the file whose comments with all positive tags and the one with all negative tags and applied 'Jieba', Python Chinese word segmentation module, to most accurately split sentences of comments for text analysis and quickly scan out all words that can be formed in the sentence. By removing all punctuation and keeping words with part of speeches, 'n', 'v' and 'a', we eventually got the output of words that differ in frequency in Bull and Bear, and also obtained the frequency proportion for each word in these two categories. For instance, the frequency of '跌破 (fall below)' occurred in the categories of Bull and Bear is 0.09% and 1.29%. It represents that '跌破 (fall below)' is a word that delivers investor's bearish mind toward stock market trends (Table 2 presents some words' frequency proportion in Bullish and Bearish category).

Table 1 - Sentiment Indicators

Bullish (+1)	反弹 Rebound	收红 Accept Red	利好 Good/ Benefits	抄底 Bottom-Fishing	加仓 Buy in more shares	高走 Go high	春天来了 Spring is coming
Bearish (-1)	阴跌 Fall down	利空 Bear news	下套 Bait-and-switch	跳水 High diving	清仓 Clearance	垃圾 Trash	韭菜 Chinese Chive
Neu (0)	是否 If...	看情况 It depends on...	为什么 Why	小心 Take care	看不出来 Cannot tell	就怕 Afraid of	? (All posts with questions)

Table 2 - Words Categorization

Words	**Words in Eng**	**Bullish (+1)**	**Bearish (-1)**	**Assigned to**
收复	Reoccupy	0.0018	0.0001	Bullish
萎缩	Contract	0.0002	0.0018	Bearish
拉到	Pull to	0.0011	0.0001	Bullish
卖光	Sold out	0.0002	0.0031	Bearish
黑	Black	0.0005	0.0035	Bearish
观望	Wait-and-see	0.0006	0.005	Bearish
火箭	Rocket	0.0021	0.0001	Bullish
开门红	Good start	0.0014	0.0001	Bullish
老套路	Stereotype	0.0002	0.0012	Bearish
蓝筹股	Blue chip	0.0016	0.0001	Bullish
杀入	Break into	0.0032	0.0001	Bullish

4. Classify Process

We tried three methods of vectorization: word count, Tf-idf score and word count binary. Based on these three methods, we used three different algorithms to train our model: Linear Regression,

Logistic Regression and Multinomial Naive Bayes. So now we have got nine different combinations in total. In order to evaluate our models, we must keep some labeled data for the test set. If the test set is too small, our evaluation may be inaccurate. However, a large test set usually means that the training set is small. If the amount of labeled data is limited, this setting will have a significant impact on performance. Our solution to this problem is to perform multiple evaluations on different test sets and then combine the scores of these evaluations, which is the Cross-validation method. In particular, we subdivide the original corpus into N subsets that called folds. For each of these folds, we trained the model using all the data except the data in this fold, and then tested the model on this fold. Even though individual folds may be too small to give an accurate evaluation score on them, the comprehensive evaluation score is based on a large amount of data and is therefore quite reliable. The advantage of using cross-validation method is that it allows us to study how much performance varies on different training sets. If we get very similar scores from all N training sets, then we can be quite confident that the scores are accurate. In our case, we divided the corpus into 5 folds to initiate model comparison.finally and finally we pick the one with the highest accuracy: Tf-idf with Logistic Regression. Because scores of our models are all above 0.8 and are quite similar to each other, we can confidently consider that the scores are accurate. The table below shows our results.

Table 3 - Model Comparison

Model	**Accuracy**	**Positive Precision**	**Negative Precision**	**Positive Recall**	**Negative Recall**
Binary LinearSVC	0.818	0.829	0.809	0.783	0.851
Binary LogisticReg	0.830	0.850	0.815	0.787	0.872
Binary MultiNB	0.831	0.850	0.816	0.787	0.872
Count LinearSVC	0.817	0.827	0.808	0.782	0.849
Count LogisticReg	0.830	0.848	0.815	0.786	0.87
Count MultiNB	0.831	0.851	0.815	0.785	0.873
Tfidf LinearSVC	0.824	0.829	0.819	0.798	0.848

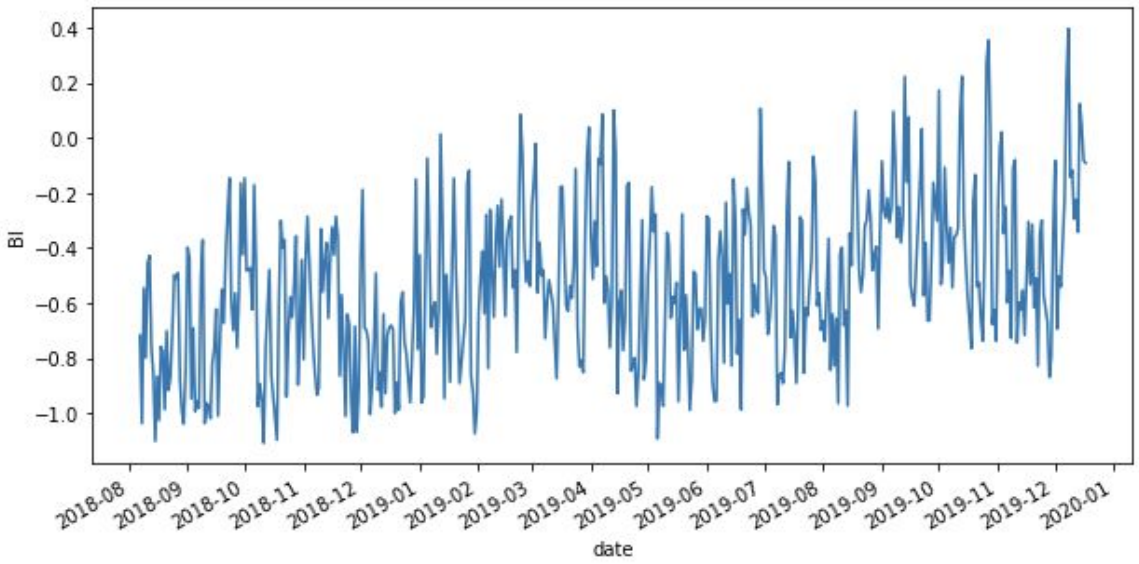
Tfidf LogisticReg	**0.835**	**0.852**	**0.822**	**0.796**	**0.872**
Tfidf MultiNB	0.832	0.847	0.819	0.793	0.867

5. Build Investor Sentiment Index

For building our investor sentiment index, we selected the bullish index as our sentiment indicator to analyze the relationship between investors' bullish sentiment and Chinese stock market trends. Consequently, we applied the BI (Business Intelligence) formula (presented below) to calculate time series data for sentiment index in daily units.In the formula, M^{bull} and M^{bear} represent the number of bullish and bearish stock comments. In previous steps, we have already traversed all the words in our data set, counted their proportions in Positive, Neutral and Negative categories and eventually obtained the words with different frequencies in those three classes. Now our program has already determined which words have bullish attributes and which words have bearish attributes. And therefore, we are already able to evaluate whether a piece of review delivers bearish sentiment or bullish sentiment. Then we applied our best classification model, Tf-idf scores with Logistic Regression, into the whole data set to calculate the number of stock comments with investors' bullish and bearish opinion. With the help of this process, we would be able to realize how many bullish and bearish stock comments (M^{bull} and M^{bear}) generated every day. In this case, we can calculate daily BI index, a sentiment ratio, to realize investors' opinion toward stock market trends. Appendix II is an example that presents BI index for each date between 2018 and 2019. Positive and negative values indicate bullish and bearish sentiment respectively (the higher the value, the stronger the sentiment). The chart below is our line graph that plots daily BI index from 08/2018 to 12/2019.

$$BI = ln\left(\frac{1+M^{bull}}{1+M^{bear}}\right) \qquad (1.1)$$

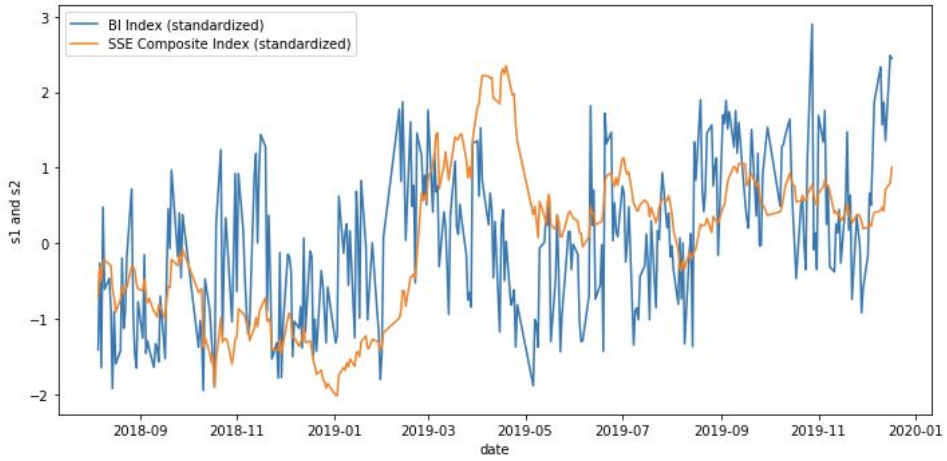
Plot 1 - Computed BI Index for 16 months



6. Comparison between Sentiment Index and SSE Composite Index

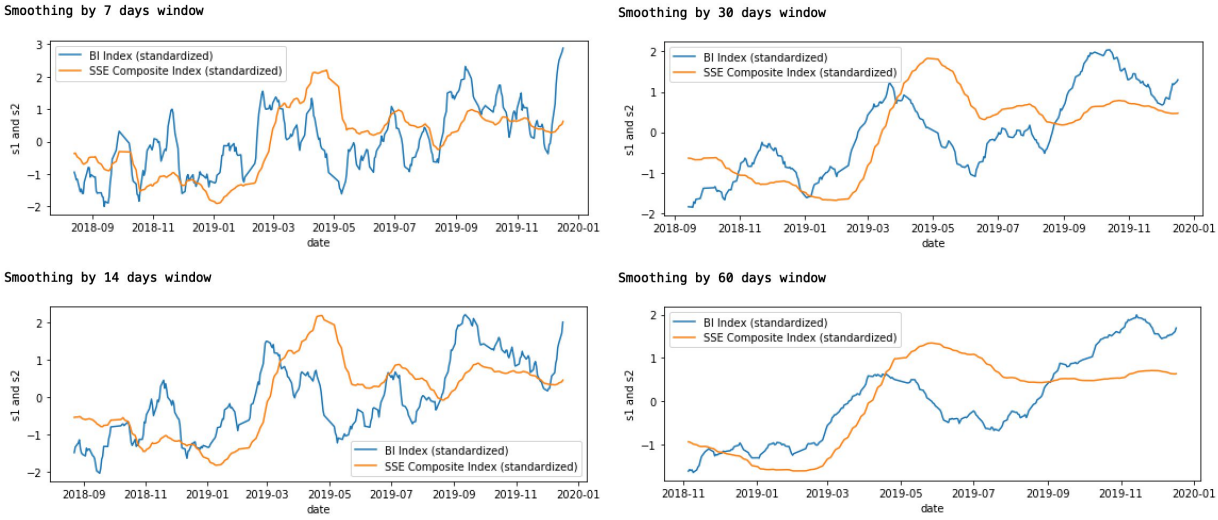
After finalizing the method of generating BI index, we computed the index between 2018-08-06 and 2019-12-17. The plot above shows that, although the attitude of shareholders varies everyday, there are distinct increasing/decreasing trend during the specific periods. To investigate how well can our BI index reflect the fluctuation in the Chinese stock market, we compared the standardized BI index with the standardized SSE Composite Index, which is an authentic stock market index of all stocks that are traded at the Shanghai Stock Exchange.

Plot 2 - BI Index vs. SSE Composite Index for 16 months



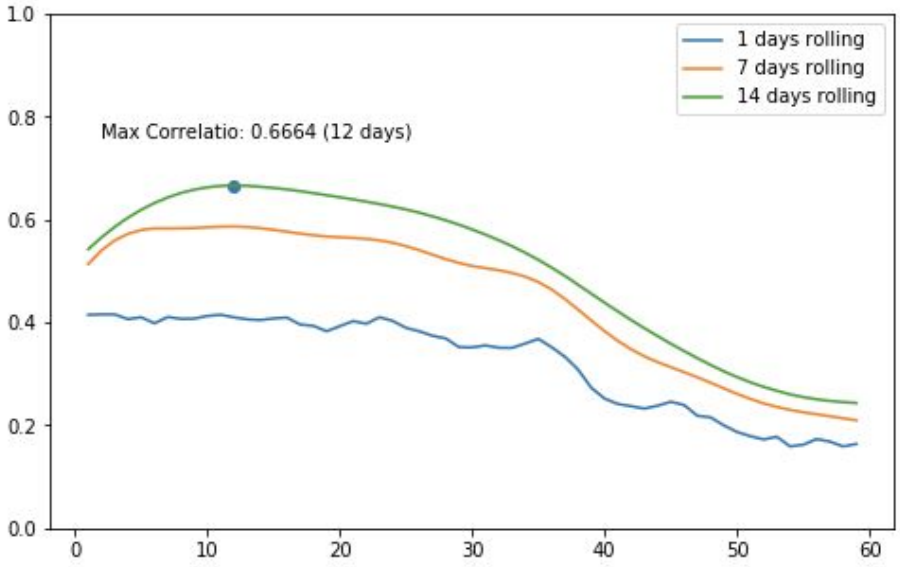
Since the Stock Exchange closes on weekends, we excluded the BI indices for each weekend during the target period to maintain the consistency of analysis. Moreover, as the sentiment index varies a lot in every single day, it's difficult for us to observe the start and the end points of each significant increase/decrease. Therefore, we smoothed the data by 7, 14, 30 and 60 days window. By observing the four subplots below, we think there is a high correlation between BI index and SSE index, and we made a preliminary conclusion that stockholders' attitude may influence the future fluctuation of the stock market.

Plot 3 - Rolling Average Comparison



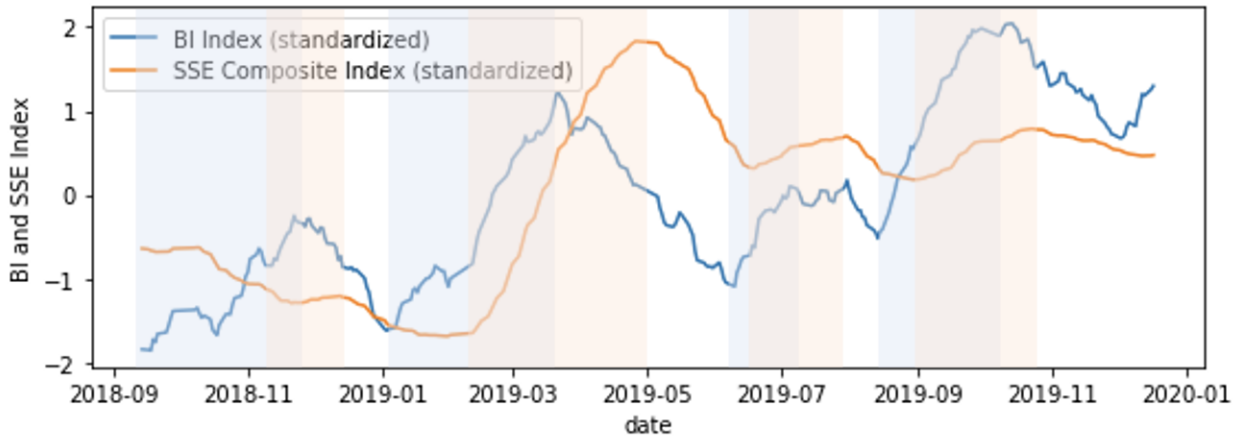
To have a better understanding of the relationship between the two indices from the numerical perspective, we computed the Pearson correlation score between them. In the plot below, the x axis is the assumed days of lag-effect of BI index on the real stock market, and the y axis is the computed Pearson correlation score (the three lines represent the score for 1, 7 and 14 days of rolling windows). As marked on the plot, the highest correlation 0.6664 occurs with the 14-day rolling window and 12-day lag-effect. It means that the sentiment of stockholders, no matter positive or negative attitudes, will significantly influence the SSE Index after 12 business days, that is approximately two weeks.

Plot 4 - Pearson Correlation Analysis



After having an overview of lagging effect, we also conducted detailed analysis on Chinese stock market regarding the three distinct increase shown on the plot (30-days rolling window). The four increase of BI index are marked by blue shaded rectangles, and the corresponding growth of SSE Composite Index are marked by yellow shaded rectangles on the plot.

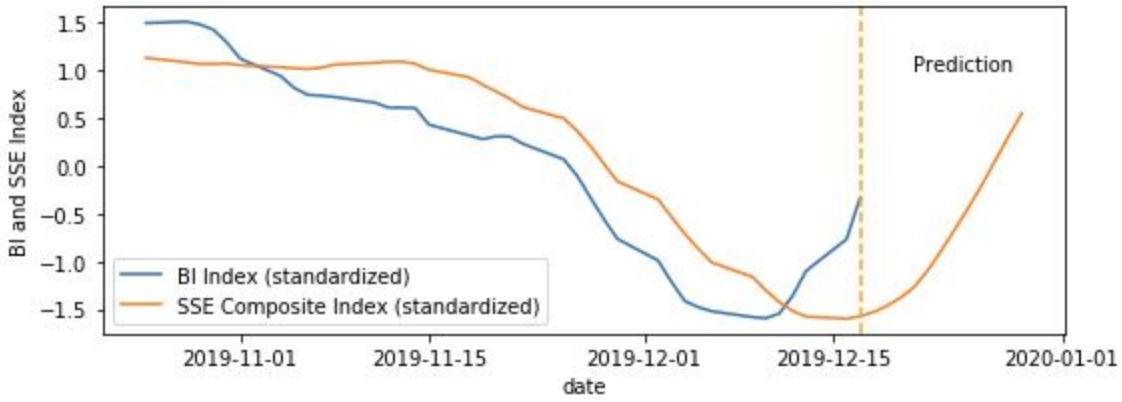
Plot 5 - Stock Market Analysis



The first significant increase of BI Index occured between September 2018 and December 2018, which reflected on the increase of SSE Index at the end of 2018. After a long-term economic depression, the market was finally influenced by the stockholders and bounced back from the lowest point. In January 2019, People's Bank of China announced that, to support the development of the real economy and lower the capital cost, it decided to bump down the deposit reserve ratio by 1 percent. The director said that the decrease of the deposit reserve ratio will release about 150 million Yuan. The new policy has brought a bright prospect of A-share market to Chinese's stockholders and, therefore, raised positive attitudes among them. Later in February 2019, the SSE Composite Index started to climb up under the positive sentiment and continually increased for two months. Moreover, in June 2019, technology stocks officially join the market. The technology board brings higher development velocity and better market tolerance to the overall A-share market. With the registration of new technology stocks, the stockholders saw great opportunities in the market and hold very positive attitudes toward the market from June to July. The sentiment of the stockholders, positively influenced the growth of the stock market, that is, the SSE Composite index continually increased from mid-June to August.

With the assumption that the sentiment of stockholders will significantly influence the SSE Index after 12 business days, we simply predicted the SSE Index for the 12 continuous business days after '2019-12-17'. The predicted values are shown on the Plot 6 below.

Plot 6 - Simple Prediction From 2019-12-17 to 2019-12-29



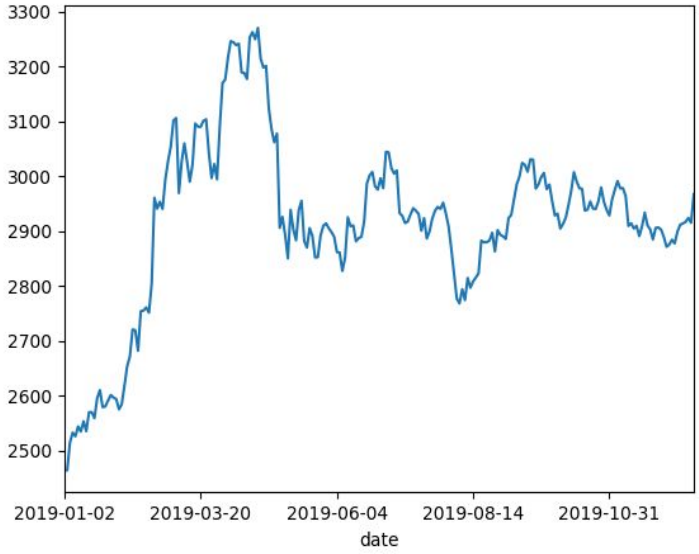
7. Build Time Series Forecasting Model

a. ARIMA (Based on the closing price of SSE Composite Index)

The first step is a stock forecasting model based on stock prices. We want to create time series on SSE Composite Index. This does not consider sentiment indicators first, based solely on stock prices.

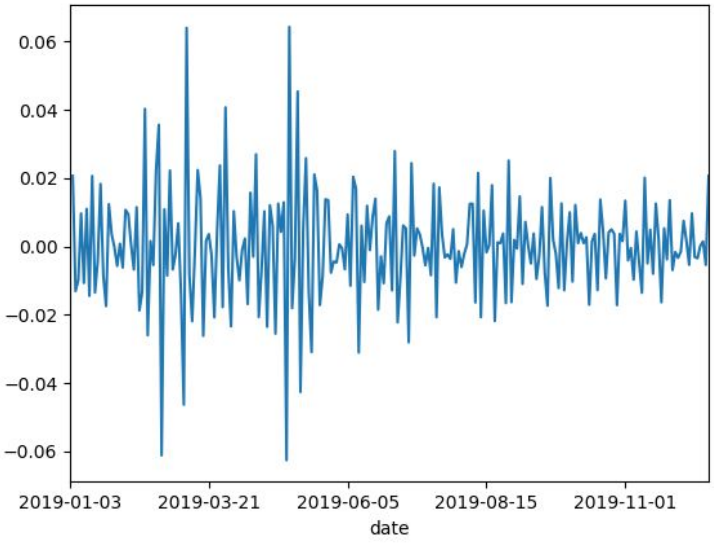
Time series forecasting requires stationary data. We want to see whether SSE Composite Index has trend or seasonality. ADF unit root test is used. The following figure is the original closing price time series.

Plot 7 - SSE Composite Index From 2019-12-17 to 2019-12-29



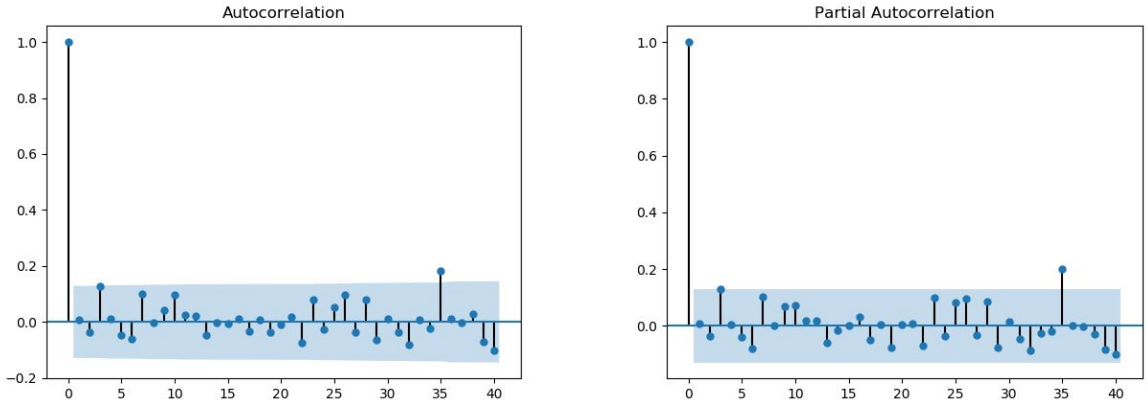
From visual perspective, there is obviously a trend. And also the p-value of the test result is greater than the significance level. So what we do next is differencing to remove trend and seasonality.

Plot 8 - SSE Composite Index after Differencing



This is the last step of data handling, now we are going to train ARIMA model. The key to get a precise and robust ARIMA model is to find optimal p and q. We use acf and pacf plot to estimate p and q value, then we try different pq combinations to our model, and pick the one with highest BIC score.

Plot 9 - Autocorrelations

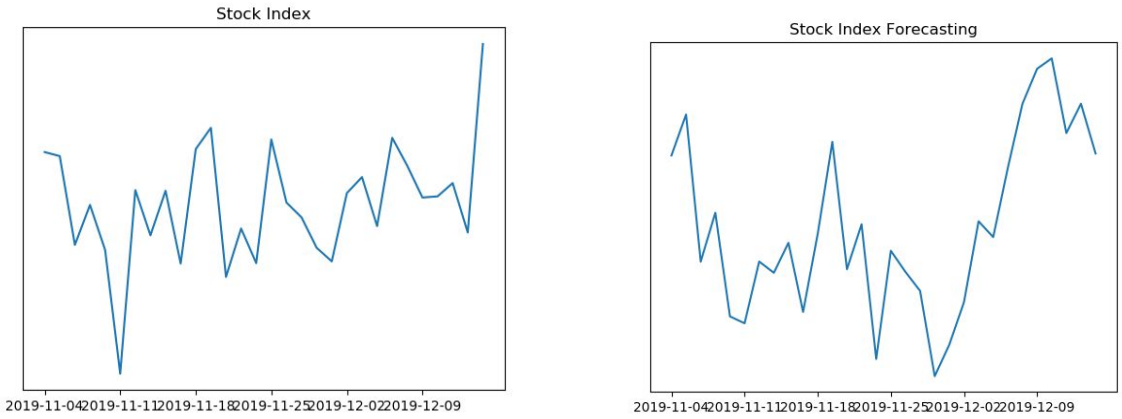


b. ARMAX (ARIMA Model based on an exogenous variable)

Furthermore, we generate an ARIMA model based on an externally generated variable (exogenous variable). For financial time series analysis, the exogenous variables are the variables that would be affected by external factors in the economic mechanism and are the variables that would also affect internal variables. The stock investor sentiment index can be regarded as a comprehensive reflection of many exogenous variables, reflecting external factors such as macroeconomics influence, company fundamentals information, policies, major events, etc.

Consequently, this project constructs an ARIMA model with an endogenous variable, the closing price of SSE composite index, and an exogenous variable, the stock investor sentiment index. As shown in the figure, after adding BI index to our model, the predictive power of our model improved significantly. We did 30 days forecasting on our model, and compare the result to the real stock market index, let's see:

Plot 10 - Stock Index Forecasting



From the figure above, we can see our model predicted the sharp decline of SSE Composite Index in November. Also, the index bounced back in December, not as much as predicted, but the trend of forecast result is generally consistent with real world stock index.

8. Explanation and Conclusion

Accompanied by the advanced development of the Internet, an increasing number of people rely on it to search for information and exchange their perspectives with other users, especially those stock market investors. They are more inclined to obtain financial information and share investment views through online platforms. After two decades of rapid development, the hard power of Chinese stock market has entered into the international advanced level but its soft power remains to be improved. For instance, in China, investors do not pay enough attention or have the capability to get the overall expected trend of the market and development situation of the company. Instead, they rely solely on various gossip, resulting in multifariously illusory news generated in all kinds of channels, which is difficult for investors to distinguish whether the news is true of false. This situation causes negative impact on the stock market. At the same time, investors' comments to the entire stock market or a certain stock not only reflects the market's situation to a large extent but also affect the stock market's up and down. The stock review forum, such as Guba Stock Comments, contains a wealth of financial data and investors'

sentiment information and those relevant information of the forum has become a significant factor affecting investors' psychology and behavior toward the stock market. Therefore, this analysis could be of great significance to study stock market volatility in China through applying investors' sentiment on the stock review forum.

For this analysis, we first obtained text data through crawling investors' reviews from Guba Stock Comments forum and then selected historical reviews over 2019 as our sample data to manually label investors' sentiment category. With the application of Text Mining, we further obtained words within each sentiment category, Bull and Bear, so that we could judge investors' Bullish and Bearish sentiment in our further processes based on the specific text within the stock comments from 2018 to 2019. Furthermore, we selected our best classification model in virtue of cross-validation technique and then employed it to judge sentiment behind each review. And then, we calculated the number of Bullish and Bearish comments in total for assisting us to come up with our BI investor sentiment index, the percentage of bullish sentiment. In order to investigate how well can our BI index reflect the fluctuation in the Chinese stock market, we compared the standardized BI index with the standardized SSE Composite Index. Since the highest Pearson correlation score is above 0.5, we would be able to conclude that shareholders' Bullish and Bearish sentiment can significantly impact the SSE Index's up and down, resulting in fluctuations in the stock market. Finally, we constructed the Time Series Forecasting Model based on the stock's closing price for providing prediction ability to the stock trend. The first step is to conduct an ARIMA model with the closing price of SSE Composite Index as our endogenous variable. So we initiated a Stationary test to check if there is any trend or seasonality behind the historic closing price of SSE Composite Index over 2019. And then we selected the optimal p-q combination with the highest BIC score to come up with our best ARIMA model, so we would be able to use this model to predict the development trend of the stock market. Subsequently, we performed a correlation analysis between time series of BI index and closing prices' rises or falls and eventually concluded that BI index reflects the trend of stock price changes to some extent. Secondly, we established the ARMAX model with stock's closing price as endogenous variable and BI sentiment index as an exogenous variable to carry out future stock price prediction. TO achieve this, we first applied ARIMA model that we previously

constructed as a basis and then added BI sentiment index as our exogenous variable to come up with our predicted results in a short term. We used rolling prediction method to forecast the future data by adding the latest data and eventually found that all the trends are predicted correctly thought there is a slight deviation in the value. Through plotting our predicted results in comparison with historical data of stock's closing price, our prediction looks more consistent as a whole. Therefore, due to the strong correlation between the BI sentiment index and changes of stock's closing price, we would be able to roughly realize the overall trend of the stock market. More importantly, we can also use the ARIMA model with BI as our exogenous variable to predict the stock market's development by combining the historical price sequence and BI sentiment index.

In the future, we can use Ensemble Learning method to improve the accuracy of sentiment analysis. Moreover, the promotion of research objects is not limited to the SSE Composite index but also the analysis toward a single stock. Meanwhile, we can further extend our research indicators that no only include the stock's closing price and BI sentiment index but also cover the stock's opening price, turnover rate or newly established sentiment indicators. Instead of a short-term forecasting model, we can also explore models and methodology for long-term forecasting for Chinese stock market.

Appendix I - Comments Crawling and Tagging Example

text	author	time	num1	num2	label
看见大盘涨成这样，我以为大牛市来了，再看看自己的股票才NM的涨1%不到。	﹀ㄣ那个叫声	12-13 14:43	1	9	0
今天真是个涨指数的好日子	2016Optimus	12-13 09:56	1	9	1
今天没赚钱的举手！！！	435066371	12-13 16:25	5	4	0
周一会不会补回来2926缺口！	7673.211005	12-14 01:22	1	3	0
哎，今天这个缺口如果回补，不知道要埋多少小散！	7673.211005	12-13 15:02	2	7	0
机构疯炒科技股	a12381238778	12-10 23:37	7	0	0
年底大盘上3100，明年大盘上6124点！	A1云风	12-13 20:25	4	2	1
补完上方缺口补下方缺口，总之补缺口是一定要的。	A666亮	12-13 12:17	3	1	0
机构最新数据显示继续减仓，三日内必有大跌，拭目以待	A9A9007	12-12 10:53	2	4	-1
假牛市来了，快买啊！	A9A9007	12-12 14:26	1	9	0
沪市的主力真的是弱爆了那种[想一下]	A9A9007	12-06 13:33	2	2	0
最后一次拉高出货！	AAAAA15	12-13 10:46	1	7	0
你们随便拉！不玩了！	AAA双曲线 AAA	12-06 14:46	2	4	0
本人看来，消费电子板块股见顶了，大家赶快逃顶了	AAA好运来2019	12-12 09:51	2	2	-1
感觉，跳水	AA平常心AA	12-13 14:05	2	1	-1
这就是稳。一点一点割肉。所以别想多了稳的意思。	ABC123555	12-09 10:25	2	9	0
放量突破60日线向3000点进发[大笑][大笑][大笑]	abc670038	12-13 15:38	3	5	1
高开低走，该滚的滚	abccqw	12-12 09:33	2	2	-1
不明白，凭啥就突然直接跳空高开？涨要有涨的理由啊。	abc阿阿	12-13 13:56	2	0	0
尾盘要炸的意思？？？	Action5	12-06 13:54	2	3	0
九连阳	Action5	12-12 14:25	2	6	1
今天拉指数，感觉周一不妙啊	ag完蛋曹	12-13 14:44	1	7	-1
你看看，又是高低走	ag完蛋曹	12-13 09:36	1	7	-1
高开低走[大笑]	ag完蛋曹	12-13 09:36	1	8	-1

Appendix II - BI Index and SSE Composite Index Example

Date	BI	Close
8/6/18	-0.9837	2705.156982
8/7/18	-0.718	2779.374023
8/8/18	-1.0381	2744.070068
8/9/18	-0.546	2794.38208
8/10/18	-0.7988	2795.310059
8/13/18	-0.7642	2785.87207
8/14/18	-0.8597	2780.965088
8/15/18	-1.1022	2723.258057
8/16/18	-0.867	2705.191895
8/17/18	-1.0268	2668.966064
8/20/18	-0.9873	2698.466064
8/21/18	-0.7019	2733.825928
8/22/18	-0.9175	2714.60791
8/23/18	-0.867	2724.624023
8/24/18	-0.6974	2729.430908
8/27/18	-0.4907	2780.898926
8/28/18	-0.879	2777.980957
8/29/18	-0.9947	2769.294922
8/30/18	-1.0392	2737.737061
8/31/18	-0.8359	2725.25
9/3/18	-0.9479	2720.733887
9/4/18	-0.6919	2750.580078
9/5/18	-0.994	2704.336914