Sentimental Analysis of Chinese New Social Media for stock market information

Guanhang Chen
Xi'an Jiaotong-Liverpool University
Suzhou, China
(86) 13777822859
guanhang.chen18@student.xjtlu.edu.cn

Lilin He
Xi'an Jiaotong-Liverpool University
Suzhou, China
(86)15606205254
lilin.he18@student.xjtlu.edu.cn

Konstantinos Papangelis
Xi'an Jiaotong-Liverpool University
Suzhou, China
(0512)81883272
K.papangelis@xjtlu.edu.cn

ABSTRACT
The popularity of social media provides a new platform to collect big social data. With the development of social sentiment analysis, high business value extracted from social data are applied to various fields. Asset price prediction, as an emerging topic based on the behavioral economics, is closely linked to social data analysis. This research aims to explore the effort of sentiment analysis data in the prediction of China composite index. Data from Sina Weibo and financial community is processed to get the useful sentiment information. A linear regression model and a multilayer neural network algorithm are used to prove the relationship between social data and price market prediction. The experiments show a strong relationship between the numbers of negative sentiment and a multilayer perceptron model is effectively built to predict the composite index.

CCS Concepts
•Computing methodologies→Model verification and validation

Keywords
Sentiment analysis; stock market

1. INTRODUCTION
Stock price prediction is a vital and challenging problem in investment activities. In traditional economic theories, assuming that asset prices fully reflect all available information and all people are rational, stock price is considered to be unpredictable due to the unpredictability of news [1]. Nevertheless, numerous studies found that these hypotheses do not always hold in real word. On the one hand, information is no longer unpredictable when social media offers the possibility to track early useful information. On the other hand, behavioral economics theory discusses that investors are not always rational and their decision-making could be affected by psychological phenomena [2]. Behavioral economics provide theory support for stock price prediction by analyzing psychological effect on the decision. How to analyze general sentiment and quantify its effects is an important direction of research on stock market [4]. As the explosive growth of social network users, social media platform becomes one of the major digital data sources nowadays. Big social data is derived to satisfy the needs of capturing the opinions of general public [5]. Based on the huge volume of big social data, the studies of opinion mining and social sentiment analysis make significant progress. In China, Sina weibo had more than 431 million active users in June 2018 [7]. Analysis of China big social data should enable people to extracting remarkable business value in financial market prediction. This research will seek to answer the following questions:

Can we find the relationship between stock price movement and sentiment measured by big social data analysis?

Can we use big social data to develop an accurate stock prediction model?

What is the difference between the accuracy of prediction based on different social media platforms?

PRAI '19, August 26–28, 2019, Wenzhou, China

 ACM ISBN 978-1-4503-7231-2/19/08…$15.00

DOI:https://doi.org/10.1145/3357777.3357778

2. RELATED WORK
The value of social network data is widely discussed and studied with the technological progress of sentiment analysis. Sentiment analysis was identified as a branch of Natural Language Processing and it was very tough because of its informal and special network language [10]. In the research of Hassan, Yulan, and Harith [11], a new method was proposed to forecast sentiment for Twitter by measuring the relationship of the semantic concept of each extracted entity and sentiment. The result shows that the F harmonic accuracy score for recognizing sentiment rise by an average of 6.5% and 4.8% over the baselines of unigrams and part-of-speech features. Deep learning algorithms also play an important role in the sentiment analysis area, Dos Santos and Gatti introduced a deep convolutional neural network based on the Stanford Sentiment Treebank (SSTb) with sentences of movie reviews and the Stanford Twitter Sentiment corpus (STS) with Twitter messages, achieving a high accuracy of over 85% in sentence sentiment prediction for both corpus [12].

The emerging work related to predicting future events based on social network data is applied in various fields including politics, finance, entertainment, marketing [8]. In Vasu's study [6], more than 60GB raw data of tweets are classified as Positive, Negative, Neutral and Irrelevant to analyze people's preferences about movies. Although the classification algorithm and sentiment analysis did not perform well in the experiment, the result still shows a satisfied accuracy in predicting 5 movies' box office.

Many studies about the relationship between stock market and public emotions (or opinions) have pointed out that people's emotions can be a factor which has effects on stock prices. To prove the effectiveness of sentiment analysis in stock price prediction, Nguyen et al. compared the average of accuracy of model using sentiment analysis and their model only use historical prices [9]. The outstanding performance of model using sentiment proved that sentiment analysis is conducive to forecast stock price. Different approaches may lead to different results in this area. Some researchers used twitter data and vector auto-regression to build a stock price prediction model, reporting the better performance of non-parametric topic-based sentiment time series approach [14]. In 2011, Bollen proposed a model to predict stock movements based on daily Twitter posters data [3], this model achieves 87.6% accuracy in forecasting the daily changes in the closing values of Dow Jones Industrial. It proves the feasibility of the application of sentiment analysis in stock price prediction. Wong C. et al analyzed daily newspaper in South Korea through sentiment analysis which shows a way to predict the stock price [21]. Besides, Wang Z. also stated that irrational behavior of people is an important factor affecting the stock market [19]. Currently, with the growing popularity of social media network, people prefer to express their views and opinions on internet. Unlike previous studies which focus on the traditional media data, our study focused on the new social media data which can provide more effective and up-to-date data than the traditional media, and the new social media media is also more convenient to retrieve.

3. RESEARCH METHODS

3.1. Data Collection

3.1.1 Sina Weibo

Sina Weibo, the biggest micro-blog platform in China, was used as the main data resources of our project. Sina Weibo API provided by Weibo open platform was used to continuously retrieve data from Sina Weibo from Oct. 22 - Nov. 4 on AliCloud [10], all Weibo containing key words “Shanghai Composite Index” were collected.

Totally, 286007 weibo were retrieved in real time from October 22, 2018 to November 4, 2018 on Alicloud.

3.1.2 Financial Community

A web crawler built with free HTML parser library JSOUP was sent to retrieve data from a famous Shanghai Composite Index community EastMoney. All posters posted from October, 8 2018 to November 4 2018 in this Shanghai Composite index community were retrieved.

Totally, 92821 posters and 174599 corresponding replies were retrieved from October 8, 2018 to November 4, 2018.

3.1.3 News

News containing key word “上 证 指 数” (Shanghai Composite Index) were manually retrieved from Baidu news from October 8 to November 4. Data was stored with the following information: content (content in the news).

3.1.4 Composite Index

Shanghai Composite indexes were manually retrieved from the stock market website from October 8, 2018 to November 5, 2018.

3.1.5 Discussions

Our data set includes Sina Weibo, posters and replies from financial community, news and composite indexes. Sina Weibo were retrieved in two weeks (Oct. 22 to Nov. 4) while other data were retrieved in four weeks (Oct. 8 to Nov. 4). Therefore, in the section 3.5 machine learning section, we will combine two different data sets with two emotion lexicons to predict the stock market and make a comparison: posters and replies from financial community and news to predict the Shanghai composite index from Oct. 22 to Nov. 4 with NRC emotion lexicon and DUTIR emotion lexicon and Sina Weibo , posters and replies from financial community and news to predict the Shanghai composite index from Oct. 8 to Nov. 4 with NRC emotion lexicon and DUTIR emotion lexicon.

3.2 Data Pre-processing

3.2.1 Data Pre-processng Methods

All punctuations, numbers and English words in the titles, contents and formats were deleted. Besides, all words with the certain speeches were also deleted: preposition, conjunction and auxiliary.

3.2.2 Data Simulations

Word segmentation was conducted to simulate the retrieved data. We randomly reassigned the words to make different sentences and used our new data set as our simulated data. Simulated data were stored for empty titles in Weibo data.

3.3 Word Segmentation

The widely adopted Chinese word segmentation tool: the Natural Language Processing Information Retrieval (NLPIR) was used to do the word segmentation [8, 18, 20, 23, 24]. After the word segmentation, each sentence was divided into words with their speeches. Then, the unweighted term frequency and the weighted term frequency were calculated.

For the Sina Weibo and posters from financial community, the weighted term frequency was calculated based on the times of read(how many times of other people read this poster), times of reply(how many times of other people reply to this poster), times of comment(how many times of other people make comments of this Weibo), times of share(how many times of other people share this Weibo) and the unweighted term frequency(unweighted term frequency in this post or Weibo).

For the replies from the financial community and news, unweighted term frequency was only calculated based on the frequency of words.

After calculating the weighted term frequency of posters and Sina Weibo and the unweighted term frequency of replies and news, every day's weighted term frequency was calculated based on the weighted term frequency of posters and Sina Weibo and the unweighted term frequency of replies and news.

After the word segmentation, four data sets were merged into two data sets: weighted term frequency of posters, replies and news which were retrieved from October 8 to November 4, weighted term frequency of Sina Weibo, posters, replies and news which were retrieved from October 22 to November 4.

3.4 Emotion Lexicon

Two different emotion lexicons were used to conduct a word-level sentiment analysis: the Chinese version of NRC emotion lexicon and DUTIR emotion lexicon from information retrieval lab in Dalian University of Technology [16].

3.4.1 NRC Emotion Lexicon

The Chinese version of the widely adopted NRC emotion lexicon was used to conduct a word-level sentiment analysis [16]. There are nine different emotion labels in the Chinese version of the NRC emotion lexicon: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

While we were using the Chinese version of NRC emotion lexicon to conduct the sentiment analysis, every day's emotion scores were calculated based on the weighted term frequency and the assigned value in the NRC emotion lexicon. Every day's emotion scores contain 9 emotion scores of nine emotion labels.

After the NRC emotion lexicon was used to do the sentiment analysis, in order to predict the Shanghai composite index which will be discussed in the machine learning section, we classified all emotion labels into two emotion labels: positive and negative.

Positive: positive, anticipation, joy, trust

Negative: negative, anger, disgust, fear, sadness, surprise

3.4.2 DUTIR Emotion Lexicon

The DUTIR Emotion lexicon from Dalian University of Technology was used to do a word-level sentiment analysis [16]. There are 21 different emotions labels in the DUTIR emotion lexicon: happy, peace, respect, praise, trust, fondness, hope, angry, sadness, disappointment, guilt, miss, panic, fear, shy, boring, hate, blame, jealous, doubt and surprise.

While we were using the DUTIR emotion lexicon to conduct the sentiment analysis, every day's emotion scores were calculated based on the weighted term frequency and the assigned value of major/minor emotion in the DUTIR emotion lexicon, including 21 emotion scores corresponding to each emotion labels.

After the DUTIR emotion lexicon was used to do the sentiment analysis, in order to predict the Shanghai composite index which will be discussed in the machine learning section, we classified all emotion labels into two emotion labels: positive and negative.

Positive: happy, peace, respect, praise, trust, fondness,hope.

Negative:angry, sadness, disappointment, guilt, miss, panic, fear, shy, boring, hate, blame, jealous, doubt, surprise.

3.4.3 Results and Discussions

Figure 1 shows the result of sentiment analysis of posters, replies and news based on NRC emotion lexicon on October 8. Figure 2 illustrates the result of sentiment analysis of posters, replies and news based on DUTIR emotion lexicon on October 8. Figure 3 shows the relationship between sentiments and stock index.

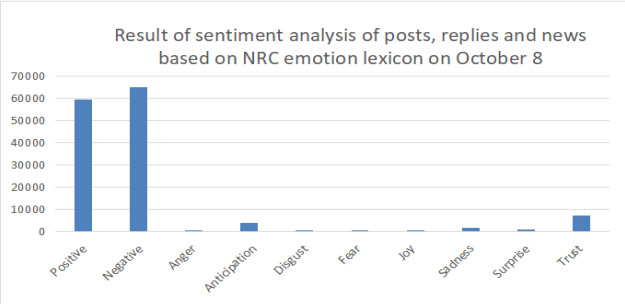

Figure 1: Result of sentiment analysis of posters, replies and news based on NRC emotion lexicon on October 8

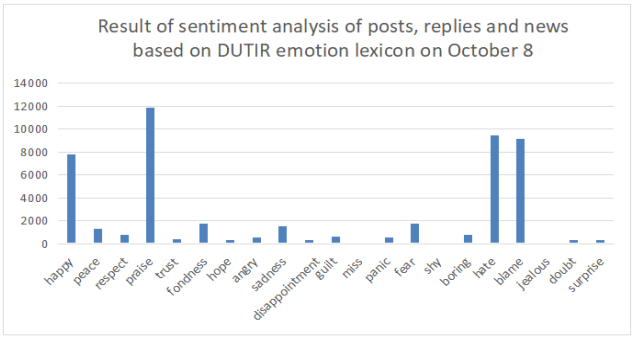

Figure 2: Result of sentiment analysis of posters, replies and news based on DUTIR emotion lexicon on October 8

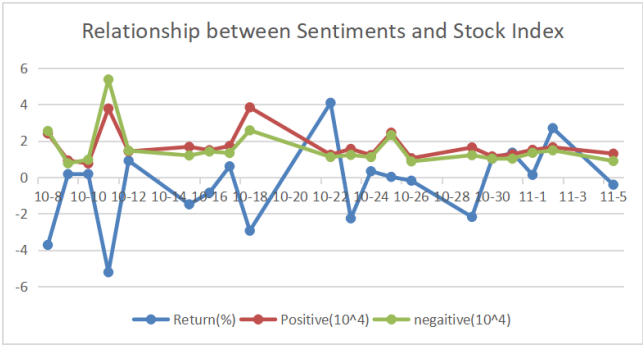

Figure 3: Relationship between sentiments and stock index

We found that for the NRC emotion lexicon, most words are positive or negative compared to other emotions, but for the DUTIR emotion lexicon, such condition didn't occur. Besides, the result based on NRC emotion lexicon also shows that anticipation and trust are the two most frequent appeared emotions besides positive and negative, which can represent the positive emotion is stronger than the negative emotion.

In the results of DUTIR emotion lexicon, the top four most frequent appeared emotions are happy, praise, hate and blame. Happy and praise belong to positive emotions, while hate and blame belong to negative emotions. Since their emotion scores are close to each other, it can be speculated that in the results based on DUTIR emotion lexicon, positive and negative emotions scores are in a balance.

Besides, on October 11, the Shanghai composite index lost 5.22% which was a large crash to the stock market, the negative emotion increased rapidly on that day accordingly. However, it is found that on October 8 and October 18, although the Shanghai composite index lost 3.72% and 2.94% which were also a large crashes to the stock market, the emotion scores of sadness didn't increase rapidly as it did on October 11.

3.5 Machine Learning

The data sets which are used for machine learning of this study are split into four subsets based on the two emotion lexicons and various data sources: Chinese-Combine (DUTIR Emotion Lexicon for data from Sina Weibo, Financial Community and news during October 22 to November 4), Chinese-Community(DUTIR Emotion Lexicon for data from Financial Community and news during October 8 to November 4) , NRC-Combine(NRC Emotion Lexicon for data from Financial Community and news during October 22 to November 4) and NRC-Community (NRC Emotion Lexicon for data from Financial

Community and news during October 8 to November 4). Each subset is analyzed separately and compared. The emotion cores of positive labels and negative labels measured by each day's data were used to predict tomorrow's composite index. All variables were normalized before to reduce amplitude variation, and the ratio of these two variables was also taken account of in the analysis.

Various statistics methods such like CHI-Square and correlation implemented by SPSS were applied to study the relationship between the dependent variable and multiple independent variables, while only linear regression didn't reject (with a 95% confidence interval). In multiple linear regression model, it is hypothesized that there is a significant linear relationship between the independent variables x (the emotion scores of positive labels and negative labels, and the ratio of emotion scores) and dependent variable y (composite index). The model could be represented as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where ε denotes the residual terms. Table 1 shows the comparison of results of the four different data sets. Adjusted R Square represents a higher degree of correlation in consideration of the number of terms. Sig. is the statistical significance referred to p-value of test. The coefficient of each independent variable is also recorded in this table.

Table 1: Comparison of results of 4 data sets

	Chinese-Combine	**Chinese-Community**	**NRC-Combine**	**NRC-Community**
Adjusted R Square	0.661	-0.164	0.574	-0.021
Sig.	0.023	Reject	0.031	Reject
Coefficient of Positive Label	0.446	Reject	0.328	Reject
Coefficient of Negative Label	0.749	Reject	0.703	Reject
Coefficient of Ratio	Reject	Reject	Reject	Reject

According to the table 1, the Sig. values in the 'Chinese-Community' column and 'NRC-Community' column are greater than 0.5 and the hypotheses are rejected. It could be inferred that data from financial community and news is not representative and related to the composite index. Since the adjusted R square of Chinese-Combine data set is more than that of NRC-Combine data set, the sentiment analysis based on DUTIR emotion lexicon is more effective for stock market movement prediction. The sig. values of variable 'positive to negative ratio' are all more than 0.1, implying that it should not be considered in this analysis. As shown in table 1, the coefficient value of positive label and negative label in 'Chinese-Combine' column are 0.446 and 0.749. Therefore, there are both positive correlations between the positive label, negative label and the response variable composite index. Negative emotion plays a more important role in the prediction.

From the results above, the 'combine' data sets will be applied to develop a multilayer neural network model by python. The dependent variable is classified into 3 categories: less than -0.1%, -0.1%-0.1%, more than 0.1%. The 1 hidden layer model and 2 hidden layers model were chosen in this study subjected to the data size. After comparing the performance of models with different hidden layers, nodes and activation functions, the results of superior models is shown below.

Table 2: The comparison of accuracy of combine data

	Chinese-Combine	**NRC-Combine**
2 hidden layers with 10 nodes	0.749283	0.677293

The accuracy shows that the model based on DUTIR emotion lexicon outperform the model based on NRC in the forecasting which is similar to their performance in linear regression model. The experiments reveals that data from Sina Weibo, financial community and news has a high relationship with the composite index, and DUTIR emotion lexicon performs better than NRC emotion lexicon when conducting the sentiment analysis in the stock price prediction.

4. DISCUSSIONS

The Shanghai Composite Index is influenced by various factors including political factors, economic factors and public emotion factors. The Shanghai Composite Index has lost a lot since the beginning of 2018 because of the stock war between China and America. Therefore, it's reasonable that the public emotion has changed dramatically from October 8 to November 4. In addition, we found that only very few people post their posters or reply to other people's posters on weekends in the financial community. Perhaps it is because that the Chinese stock market stopped on weekends.

Unlike previous works which focus on traditional social media data, our study used data from the new social media which provided more data to predict the stock market price. Therefore, our study might could have a higher accuracy.

In our study, we found that the data set which contains the Sina Weibo, posters, replies and news performs better than the data set which contains the posters, replies and news. Probably because people prefer to express their own opinions instead of following experts' opinions in Sina Weibo. Nevertheless, in the financial community, there are more experts than Sina Weibo, leading people to prefer to repeat experts' opinions. In a lot of conditions, experts' opinions can't represent public emotion so that experts' opinions have a weaker relationship with the stock market price.

Besides, we also found that DUTIR emotion lexicon performs better than the Chinese version of NRC emotion lexicon in the

prediction of stock market price. Previous work about sentiment analysis also shows a similar result. Tiffany claimed that it might because that the Chinese version of NRC emotion lexicon is a translation version and there is a narrow between different languages [16].

5. LIMITATIONS AND FUTURE WORKS

In the present study, due to the time limitation, only 28 days' data is retrieved, which is not long enough for predicting stock market correctly. Hence, in the future, we will retrieve more data to improve the accuracy of our prediction.

Besides, our present study only considered the word-level sentiment analysis. The correlations between various words were ignored. Therefore, some machine-learning approaches such like Support Vector Machine(SVM) will be considered to improve the present study.

6. CONCLUSIONS

Nowadays, with the development of Internet, social data has grown explosively. A large number of users choose to post their opinions on the new social media such as Twitter and Facebook. Therefore, although the stock market price prediction was regarded as an impossible task in the last century, the large number of user data has made it possible to predict the stock market price. In this project, we used the result of sentiment analysis of new social media data to predict the stock market price. Specifically, we conducted a comparative research on the sentiment analysis of two emotion lexicons and two data sets. With the combination of the sentiment analysis and machine learning, we found the DUTIR emotion lexicon perform better than the Chinese version of NRC emotion lexicon in the stock prediction field. In addition, we also find that the prediction result of data set of Sina Weibo, posters, replies and news is more effective than the result of data set of posters, replies and news. In conclusion, our study shows a strong relationship between stock market and new social media data.

7. ACKNOWLEDGMENTS

Thanks go to Mr. Irwyn Sadien for his valuable suggestions to the whole project, thanks also go to Mr. Peeranut Apinyayanyong and Mr. Yunfan Li for their helps in the related work part.

8. REFERENCES

[1] Baker M. and Wurgler J. "Investor sentiment in the stock market," *Investments*, vol. 21,no. 2, pp. 129-152, 2007.

[2] Basu, S. "Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis," *Investments*, vol. 32, no. 3, pp. 663-682, 1977.

[3] Bollen J., Mao H. and Zeng X. "Twitter mood predicts the stock market," *Investments*, vol. 2, no. 1, pp. 1-8, 2011.

[4] Cambria E., Rajagopal D., Olsher D. and Das D. "Big social data analysis," *Investments*, vol. 13, pp. 401-414, 2013.

[5] CIW Team,"Weibo monthly active users (MAU) grew to 431 million in Q2 2018" Retrieved from *https://www.chinainternetwatch.com/26225/weibo-q2-2018/*. 2018.

[6] Jain, V., 2013. Prediction of movie success using sentiment analysis of tweets. *The International Journal of Soft Computing and Software* , 3(3), pp.308-313.

[7] Kahneman D. "Maps of bounded rationality: Psychology for behavioral economics," *Investments*, vol. 93, no. 5, pp. 1449-1475, 2003.

[8] Li X. and Zhang C. "Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method," *2013 IEEE 4th International Conference on Software Engineering and Service Science*, IEEE Press(2013), 267-270.

[9] Nguyen T.H., Shirai K. and Velcin J. "Sentiment analysis on social media for stock movement prediction," *Investments*, vol. 42, no. 24, pp. 9603-9611, 2015.

[10] Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

[11] Saif, H., He, Y. and Alani, H. Semantic sentiment analysis of twitter. In *International semantic web conference*. Springer Press(2012). 128-132.

[12] Santos, C. and Gatti, M. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Springer Press(2014), 69-78.

[13] Schoen H., Gayo-Avello D., Takis P., Mustafaraj E., Strohmaier M., and Gloor P. "The power of prediction with social media," *Investments*, vol. 23, no. 5, pp.528-543, 2013.

[14] Si J., Mukherjee A., Liu B., Li Q., H. Li, and X. Deng, "Exploiting topic based twitter sentiment for stock prediction," *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, vol. 2, pp. 24-29.

[15] Straur N., Vliegenthart R. and Verhoeven P. "Lagging behind? Emotions in newspaper articles and stock market prices in the Netherlands," *Public Relat*. Rev. 2016.

[16] Tiffany Y. T., Pinata W., Guan A. and Chen G. "The Foreign Language Effect" and Movie Recommendation: A Comparative Study of Sentiment Analysis of Movie Reviews in Chinese and English. *In proceedings of the 2018 10th International Conference on Machine Learning and Computing*.ACM Press (2018), 79-84.

[17] Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics (2012), 115-120.

[18] Wang X., Chen X., Yao C., Wu W. and Ding Y. "The application of automatic summarization technology in document management," *2013 IEEE 4th International Conference on Software Engineering and Service Science*, Beijing, 2013, pp. 919-921.

[19] Wang Z.. A research and design of stock market risk prediction model based on sentiment analysis. *Beijing University of Posts and Telecommunications*. 2018.

[20] Wen F., Yun L., Yan X., Jing C. and Xiao Y. "Research on Semantic Retrieval for Communication Ontology," *2015 8th International Conference on Intelligent Computation*

Technology and Automation (ICICTA), IEEE Press(2015), 756-760.

[21] Wong C., and Ko I. Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices. *in 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 2016.

[22] Xu Y., Wang Y.P. and Qin W.X. "Social Network analysis on Sina Weibo based on K-means algorithm", *in Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, IEEE Press(2016), 127-132.

[23] Zhang H.P., Liu Q., Cheng X.Q., Zhang H. and Yu H.K. "Chinese lexical analysis using hierarchical hidden Markov model," *in Proceedings of the second SIGHAN workshop on Chinese language processing*, IEEE Press (2003), pp. 63-70.

[24] Zhang H.P., Yu H.K., Xiong D.Y., and Liu Q. "HHMM-based Chinese lexical analyzer ICTCLAS," *in Proceedings of the second SIGHAN workshop on Chinese language processing*, IEEE Press(2017), 184-187