Summer 6-26-2019

Mining the Impact of Investor Sentiment on Stock Market from WeChat

Haiyuan Zhao
School of Government, Beijing Normal University, China

Dannuo Wang
School of Government, Beijing Normal University, China

Mingyan Wang
School of Government, Beijing Normal University, China

Xinrui He
School of Government, Beijing Normal University, China

Jian Jin
School of Government, Beijing Normal University, China, jinjian.jay@bnu.edu.cn

Mining the Impact of Investor Sentiment on Stock Market from WeChat

Haiyuan Zhao, Dannuo Wang, Mingyan Wang, Xinrui He, Jian Jin[*]
School of Government, Beijing Normal University, China

Abstract: In this study, the CSI 300 Index in China mainland and original articles from authoritative stock WeChat public accounts are investigated regarding their relations. First, a sentence-level sentiment classification approach for analyzing investor sentiment polarities in text corpus is proposed by expanding synonyms. Then, the Granger causality test is utilized to examine the impact of sentiment index on the stock price and volume-values. It shows that the influence of overall investor sentiment on volume-values is more rapid than that on stock price and the impact of positive sentiment is found to be more lasting than the negative in both stock price and volume-values. Furthermore, it is worth noting that there is a dual-stage phenomenon in the impact of positive sentiment on volume-values, which indicates that some investors react to positive information immediately while others may choose to wait and follow the trend.

Keywords: investor sentiment, sentiment analysis, stock market, investor reaction, WeChat

1. INTRODUCTION

Classical financial theory argues that the asset transaction price is the reflection of all information, and investors can evaluate the value of assets rationally. However, due to subjective factors such as knowledge and individual preferences, investors may not perform rationally when making complex decisions involving uncertain factors. Recent studies show that investor sentiment may affect investors' decision-making to a certain extent and further affect stock price trend [1]. Some studies investigating the impact of investor sentiment based on Internet information on the stock market also supported this result [2-5].

However, few research concerns were paid on analyzing original articles in WeChat, which becomes one of the most popular information platforms in recent years in China mainland. Compared with other platforms, WeChat information spreads faster with a huge number of users and user stickiness is stronger. Thus, mining investor sentiment from WeChat becomes a new choice. Accordingly, in this study, relations between original articles from the authoritative stock WeChat public accounts and the CSI 300 Index are investigated. Based on some previous studies, two hypotheses are proposed.

Hypothesis 1: The influence of investor sentiment on stock volume is more rapid than on stock price.

Some study shows that stock price changes are affected by many factors [6]. In contrast to stock prices, stock volume-values are directly affected by investor decisions. Some studies investigated the correlation of trading volume and trading prices. They argued that price changes are more complex than those of volume changes [7] since price changes can be triggered by multiple elements while the volume is affected mainly by subjective elements. Thus, the influence of investor sentiment on stock volume is more rapid than on stock price.

Hypothesis 2: The influence of positive sentiment is more lasting than negative sentiment on stock market.

Some study shows that, given positive and negative information, people may react differently [8]. Some study also shows that positive emotions have a more significant impact on the stock market than negative emotions in the long term [9]. Meanwhile, investors are often overconfident in their estimates of returns, so they are more willing to believe in good news [10]. Thus, the influence of positive sentiment is more lasting than

[*] Corresponding author. Email: jinjian.jay@bnu.edu.cn

negative sentiment on stock market.

In this study, two hypotheses will be evaluated by analyzing online investment information in WeChat public account articles. Categories of experiments will be conducted, and analytical discussions will be made.

The rest of this paper is organized as follows. Section 2 elaborates recent studies about the correlation between investor sentiment and stock market. Section 3 introduces the daily CSI300 index and WeChat dataset. Also, the approach of sentiment analysis and Granger causality test are explained; Section 4 presents the results of Granger causality test and some interesting finding are presented. Section 5 summarizes this study.

2. RELATED WORK

2.1 Research on the Construction of Investor Sentiment Index

Generally, studies about investor sentiment indicators can be divided into two categories. One is to use financial market indicators, economic index indicators as proxy indicators. For example, Zheng and Lin [11] selected the discount rate of closed-end funds. Lu and Leng [12] adopted the consumer confidence index. The other is to analyze emotions of financial news, reviews and articles to obtain the time series of investor sentiment. For instance, Tanya et al. directly utilize the company-specific news sentiment data provided by Thomson Reuters News Analytics [13]. Yan et al. [14] constructed a comprehensive stock emotion dictionary for incomplete emotional dictionary and one-sided sentence analysis in stock text emotional analysis, and then analyzed stock text emotions from sentence tendentiousness, degree and relevance.

2.2 Research on the Relevance between Investor Sentiment Index and Stock Market

Studies on the correlation between investor sentiment index and stock market can be divided into three categories.

The first are to employ financial models to analyze the sequence of investor sentiment and stock market index. Some researchers get the time series of investor sentiment and market trading indicators and then a regression analysis is made to study the predictive ability of investor sentiment on stock market. For example, Yu et al. [3] selected 300 listed companies, extracted the financial news data of each company from the China Securities Network, and constructed a time series of weekly news emotional indicators by word segmentation. Then, different portfolios were divided to make CAPM, FF3, FF4, FF5 regression for each company's news sentiment index sequence and stock return sequence. It is found that positive news sentiment can effectively predict the rise of stock returns in the next week. Smales [15] constructed emotional indicators by using news sentiment data by Thomson Reuters News Analytics. Through Fama-French factor asset pricing model regression, it was found that there is a significant correlation between news sentiment and stock returns and the correlation is closely related to the industry. Oliveira et al. [16] analyzed Twitter's user messages, and further explored the diversity of traditional emotional indicators using four regression methods and statistical tests.

The second are to use time series analysis methods and establish an ARMA-GARCH family model and a VAR model for investor sentiment sequence and transaction index sequence, and the impulse response function analysis is carried out to study their interactions. Meng et al. [2] collected reviews of six high-profile stocks in *Eastern Fortune Internet Stock Bar* and got individual stock sentiment index by an emotional dictionary and the naive Bayesian sentiment classification. Then, an ARMA-GARCH family model was built to analyze emotional index and individual stock return. Network sentiment was found to present a certain predictive effect on stock return in the short term, and in most cases, the impact of stock return on network sentiment has a long-time lag.

The third are conducted from the perspective of causality, which mainly apply the correlation analysis and different types of Granger causality test to explore the interaction between investor sentiment and the stock market. According to the Chinese financial news related to listed companies on Taiwan Stock Exchange, Wei et al [17] constructed a comprehensive news sentiment index. Through the Granger causality test, correlation between the index and earnings, transaction value, turnover rate and volatility index was analyzed. It shows that the weekly and monthly comprehensive news sentiment index has a certain predictive effect on market returns and provide theoretical support for portfolio decision-making. Bu [4] extracted the East Fortune Internet stock bar posts and used the Naive Bayesian approach to establish the index of investor sentiment. Through Granger causality test and instantaneous Granger causality test, it is found that the formation of investor sentiment depends on the pre-market returns. Investor sentiment has no predictive ability on stock market returns, trading volume and volatility, but has a current impact on stock returns and trading volume. Xu et al. [6] extracted the time series of network emotion from Sina Weibo texts. Using the mean Granger causality test and the quantile Granger causality test, whether there is a causal relation between network emotion volatility and stock market returns was explored. In some specific quantile intervals, network emotion volatility has a significant causal relation with stock market returns, and it provides evidence for the predictability of stock market returns under specific conditions.

3. DATA AND APPROACH

3.1 Data description

In this study, an article dataset and a stock dataset will be explored.

In the article dataset, WeChat public accounts' articles were extracted from Qing-Bo big data platform, which is a WeChat data website. In order to demonstrate the overall sentimental tendency of the WeChat public media, top 50 public accounts in the stock section leaderboard were selected according to the activity index ranking. Totally, these public accounts have more than 13 million hits per month. Thus, articles published by them have a huge impact on the sentiment of most investors. 21,353 articles from November 2016 to November 2018 were finally gained with filtering out irrelevant ones. For each article, timestamp, author, title and abstract are recoded.

In the stock dataset, considering that current major financial markets are concentrated in Shanghai and Shenzhen in China, stock data was extracted from daily CSI300 Index. It mirrors the fluctuations of Chinese stock market more comprehensively. The stock dataset includes closing-values and volume-values of CSI300 from November 1, 2017 to November 19, 2018 (involving 502 trading days).

3.2 Sentiment analysis

It is difficult for investors to distinguish the real information from noises. In this case, when an individual investor makes decision, he/she is assumed to refer to a clear-cut statement or imitates the behavior of other popular groups and his/her investment sentiment is more likely to be influenced by opinions from WeChat public accounts' articles. Hence, the sentimental tendency of articles is used to evaluate investor sentiment in this study.

However, the challenge is that, compared with manual judgment, many public tools for sentiment analysis usually return a deviated result. Table 1 shows that some examples of fallacious results given by TencentNLP and SnowNLP, which are two famous tools for analyzing the sentiment polarity of Chinese corpus. As presented, some sentences, which are easy to judge by humans, are assigned to incorrect results by these two tools.

Table 1. Examples of fallacious result

Examples	Tools	Returned result	Manual Judgment (ground truth)
连续两日大涨，市场人气明显回升 (The market has risen sharply for two consecutive days.)	TencentNLP	negative	positive
受消息刺激，继续上涨 (The stock price is stimulated by the news and continues to rise)	SnowNLP	negative	positive
全球股市暴跌！周期的宿命！ (Global stock market plunged! The fate of the cycle!)	SnowNLP	positive	negative

To improve the performance of sentiment analysis, similar with the approach in Meng et al. [1], a dictionary-based method is proposed in this study. First, some sentiment words with obvious bullish and bearish sentiment were selected by term frequency and expert screening. All seed words are listed in Table 2.

Table 2. Seed sentiment words

Bullish sentiment words	Bearish sentiment words
暴涨(skyrocketing), 牛市(bullish), 反弹(rebound), 利好(good), 利多(profitable)	暴跌(plunge)，熊市(bearish)，探底(bottom)，利空(loss)，回落(fall back)

On the basis of seed words, a Chinese Synonym tools named *Synonyms* was used to expand our sentiment dictionary. *Synonyms* returned 10 synonyms and their cosine similarity distances with each seed word. Afterwards, the weight of each seed word was assigned to 1, and weights of their synonyms were assigned to the cosine similarity with seed sentiment words. Due to that a sentiment word can be the synonyms of many seed words, the highest value was taken as its weight. Finally, 44 sentiment words were selected after irrelevant synonyms are filtered out. In Table 3, some exemplary sentiment words as well as corresponding weights are listed.

Table 3. Some exemplary sentiment words and their weights

Bullish sentiment words		Bearish sentiment words	
暴涨(skyrocketing)	1.0	暴跌(plunge)	1.0
涨停(raising limit)	0.8381	跌停(limit down)	0.8202
飙升(soaring)	0.7042	下跌(fall)	0.7761
激增(surge)	0.5714	震荡(wild swings)	0.5559
强力(strong)	0.4694	停滞(stagnant)	0.4464

For the sentiment polarity of one article, all words in it are checked through the sentiment dictionary. The bullish sentiment index SI^{bull} is then defined as the sum of all bullish sentiment words' weights $\mathcal{W}_i^{bull}$, and, similarly, the bearish sentiment index SI^{bear} is defined as the sum of all bullish sentiment words' weights $\mathcal{W}_j^{bear}$. If SI^{bull} is greater than SI^{bear}, then this article is assumed to be bullish, vice versa.

$$SI^{bull} = \sum_i \mathcal{W}_i^{bull} \tag{1}$$

$$SI^{bear} = \sum_j \mathcal{W}_j^{bear} \tag{2}$$

To further illustrate the applicability of this approach, an evaluation was conducted through random sampling. The other two current tools, SnowNLP and TencentNLP, were used as competitors of this proposed method.

Due to that the task of sentiment analysis was generally considered as a classification task, some metrics, such as Jaccard similarity, recall, precision and F1 score, are used to evaluate the overall performance. 3 random sampling were conducted. For each sampling, 500 samples were taken as test sets for each validation and these samples are manually labeled. In Table 4, the average values of all these metrics regarding 3 validations are presented. It shows that the proposed approach outperforms both the SnowNLP and the TencentNLP by a significant margin except recall. One major potential reason is that the proposed approach can identify explicit emotional semantics precisely but fail to judge implicit semantics. The superiority of the proposed approach in precision helps to improve the performance of sentiment analysis in this study.

Table 4. Test performance of each approach

	Jaccard similarity	Precision	Recall	F1 score
SnowNLP	0.5710	0.6274	**0.7327**	0.6754
TencentNLP	0.6211	0.7456	0.5798	0.6506
Our approach	**0.6858**	**0.9007**	0.5450	**0.6772**

The proposed approach is then used to analyze the title and abstract of each article in the article dataset. Given that there will be multiple articles published on the same day, the positive sentiment $S_t^{positive}$ in the t-th day is the sum of the bullish sentiment index SI_i^{bull} of all the articles and the negative sentiment $S_t^{negative}$ in the t-th day is the sum of the negative sentiment index SI_j^{bear} of all the articles.

$$S_t^{positive} = \sum_i SI_i^{bull} \tag{3}$$

$$S_t^{negative} = \sum_j SI_j^{bear} \tag{4}$$

To mirror the sentiment of all investors, the overall sentiment index $SENT_t$ was defined for a single day. Both positive and negative sentiment in the t-th day was integrated by $SENT_t$. To reduce the absolute value, the logarithmic transformation is used in the ratio. If $SENT_t$>0, then the overall investor sentiment is bullish. If $SENT_t$<0, the overall investor sentiment is bearish.

$$SENT_t = \ln\left(\frac{1+S_t^{positive}}{1+S_t^{positive}}\right) \tag{5}$$

3.3 Granger causality test

The Granger causality test was applied to identify the causality relevance of investor sentiment tendency and stock market variation, particularly CSI300 index closing-values and volume-values. The test rests on the assumption that, if a variable X causes Y, then changes in X will systematically occur before changes in Y. The result shows that lagged values of X will exhibit a statistically significant correlation with Y. However, correlation cannot derive causality. The Granger causality analysis is not to test actual causation but whether the time series has predictive information about the other or not.

To analyze the variation of stock market, the change rate of closing-values and volume-values was calculated at first. On this basis, both the sentiment time series and the stock time series are normalized. For the time series, the normalization is affected by the time span significantly. Accordingly, the normalization method developed by Bollen et al. [9] was taken in this study, in which a sliding window of k days before and after a given day is established. Specifically, the z-score of time series X_t, denoted $\mathbb{z}_{X_t}$, is defined as,

$$\mathbb{z}_{X_t} = \frac{X_t - \bar{x}(X_{t\pm k})}{\sigma(X_{t\pm k})} \tag{6}$$

$\bar{x}(X_{t\pm k})$ and $\sigma(X_{t\pm k})$ represent the mean and standard deviation within the period [t-k, t+k]. This normalization induces time series to be expressed on a scale around a zero mean and one standard deviation. In this study, the parameter k was set to be 5.

In addition, the Ganger causality test is greatly affected by the number of lag days and, that is, the news before different days have different effects on the stock market. Thus, the results of different lag days were considered in present study.

4. EXPERIMENT RESULT

4.1 Results

In this section, the article dataset and the stock dataset are integrated. Both stock time series and investor sentiment time series were calculated by the proposed approach in Section 3. Specifically, the stock time series involves the change rate of closing-values and volume-values. The Granger causality test was conducted on the time series of stock and investor sentiment.

4.1.1 Impact on the change rate of closing-values

The change rate of closing-values was examined with overall sentiment index, positive sentiment index and negative index and the p-value is utilized to analyze the Granger causal relationship. P-values of different lag days are listed in Table 5 and different level causality confidence markers are marked.

As shown in Table 5, the overall sentiment index was significant between 2 and 5 lagging days. This indicated that the impact of investor sentiment on change rate of closing-values is not immediately. It shows that there is a certain lag for stock price in the react to investor sentiment. Also, the positive sentiment index is found to be significant between 3 and 6 lagged days. It implies that the impact of positive sentiment, though lagging, affect the stock price in the long run. Comparatively, the negative sentiment index is found to be significant between 2 and 3 lagged days, which presents that the negative has a faster and more transient impact on the stock price.

Table 5. P-values (change of closing-values & news sentiment)

Lag	Sentiment	Positive	Negative
1 day	0.5178	0.8509	0.2988
2 days	**0.0009*****	0.3419	**0.0011****
3 days	**0.0021****	**0.0991***	**0.0016****
4 days	**0.0058****	**0.0624***	0.3268
5 days	**0.0007*****	**0.0004*****	0.1326
6 days	0.1939	**0.0114****	0.2593
7days	0.6529	0.3033	0.3525

(p-value < 0.001: ***, p-value < 0.05: **, p-value < 0.1: *)

Overall, the positive sentiment index is found to have more significant causal explanation for the change rate of closing-values. In Figure 1, the positive sentiment index is compared with the change rate of closing-values. It shows that trends of the stock market price and the positive sentiment series are similar, and the investor sentiment is found to present a strong causing relation with the stock price time series.

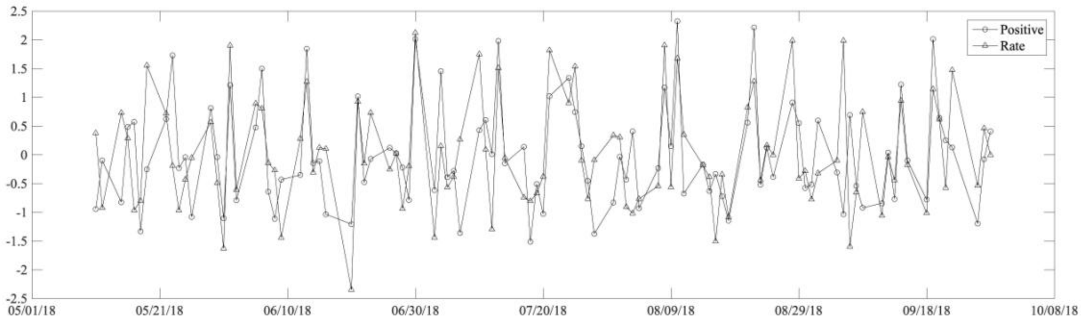

Figure 1. comparison of positive indicator and change rate of closing-value

4.1.2 Impact on the change rate of volume-values

Similarly, the Granger test relation between the change rate of volume-values and the sentiment index was studied. In Table 6, p-values of different lag days in 3 types of sentiment index are listed.

Table 6. P-values (change of volume-values & news sentiment)

Lag	Sentiment	Positive	Negative
1 day	**0.0050****	**0.0170****	0.1398
2 days	0.1988	0.1344	0.8624
3 days	0.5335	0.6076	0.3835
4 days	0.3476	0.1572	0.5829
5 days	**0.0600***	**0.0005*****	0.9192
6 days	**0.0183****	**0.0050****	0.2335
7days	0.9628	0.6890	0.2594

(p-value < 0.001: ***, p-value < 0.05: **, p-value < 0.1: *)

As shown in Table 6, both overall sentiment index and positive sentiment index are significant in 2 different periods. It indicates that the significant influence of positive sentiment on the public has two stages. The first influence may be caused those who is sensitive to positive messages and they respond to these messages quickly, so the positive sentiment index was significant in lagging 1 day. The other influence may be caused by some who choose to wait a period of time, and then follow the trend to buy stocks, affecting the stock market. It is the potential reason that positive sentiment index is found to be significant between 5 and 6 lagging days. The negative sentiment is shown not to be significant in the whole 7 lagging days. In contrast to positive sentiment index, the negative sentiment does not present to affect the volume-values.

To visualize this phenomenon, p-values are illustrated in Figure 2. It shows that the Granger test relation in positive sentiment index is presented by an "inverted U-shaped" distribution and there is one peak within a dual significant period.

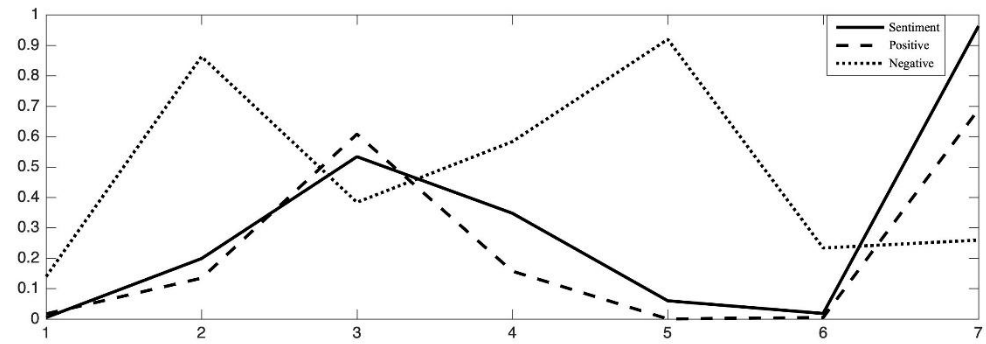

Figure 2. p-values (change rate of volume-values)

4.2 Discussion

According to the above results, some interesting findings are as listed.

- **The influence of investor sentiment on stock price and volume-values is different.** According to the Granger test, the impact of investor sentiment on volume-values is more rapid than that on stock price. One potential reason is that the volume-values and investors' decision are directly related. Investor sentiment affects volume-values by influencing decision making.
- **The influence of positive information is more lasting.** The positive sentiment has a more significant Granger test relation with both stock price and volume-values and it has a longer impact than negative. It shows that when people receive positive information, they are more likely to react and enter the market.
- **In the react to positive information, there are two kinds of people in the market, one is to directly react to the news, and the other is to wait.** According to the previous analysis, it shows that the Granger test relationship is significantly "inverted U-shaped" distributed with a single peak within a dual significant

period. It indicates that there is a dual-stage impact on investors' decision when positive information is received. The first influence may be caused by someone who is sensitive to news sentiment and they respond to the news sentiment quickly. The other influence may be caused by some people who choose to wait a period and then follow the trend to buy stocks, which affect the stock market later.

5. CONCLUSION

This study indicates that the influence of investor sentiment on stock price and volume-values is different, and the positive information has a long-term impact on the stock market, corresponding with some preliminary results obtained in other studies. Nevertheless, in this study, some results which are inconsistent with previous studies are also presented. For instance, the significance of the Granger causality test presents the "inverted U-shaped" distribution within different number of lagging days, which shows that some stock investors are immediately active followers while others may be the wait-and-see followers.

This study complements many studies on investor sentiment and stock market in China mainland from the perspective of WeChat public account articles, which demonstrates to be an effective verification of behavioral finance in the era of big data. This study also shows that WeChat public platform can be used as an effective research object for analyzing the stock market in China mainland. Managers can gain public opinions of the stock market by analyzing WeChat public account, so as to take corresponding measures and guide investors to invest rationally.

ACKNOWLEDGEMENT

The work described in this paper was supported by a Grant from the National Nature Science Foundation of China (project no. NSFC 71701019/G0114).

REFERENCES

[1] De Long, J Bradford，Andrei Shleifer，Lawrence H. Summers. (1990). Noise trader risk in financial markets. Journal of Political Economy. 98(4):703-738.

[2] Meng Zhiqing, Zheng Guojie, Zhao Yunwen. (2018). Research on the relationship between online investor sentiment and stock market price: based on text mining technology analysis. Price theory and practice (8):127-130(in Chinese).

[3] Yu Qin, Zhang Bing, Yu Wenwei. (2017). Is news sentiment behind stock returns? Financial Economics Research (6):95-103(in Chinese).

[4] Bu Hui, Xie Zheng, Li Jiahong, Wu Junjie. (2018). The Impact of Investor Emotion on Stock Market Based on Stock Evaluation. Journal of Management Science, v.21; No.166 (4): 91-106(in Chinese).

[5] Xu Qifa, Bo Zhongpu, Jiang Cuixia. (2017). Research on the relationship between network sentiment based on Quantile Granger causality and stock market returns. Management Science (3):147-160(in Chinese).

[6] Zhang Weiping, Zhuang Xintian (2019). The stability of Chinese stock network and its mechanism. Physica A: Statistical Mechanics and its Applications.515: 748-761

[7] Stosic Dusan, Stosic Darko,Stosic Tatijana, H.Eugene Stanley.(2015).Multifractal properties of price change and volume change of stock market indices. Physica A: Statistical Mechanics and its Applications.428:46-51.

[8] Li, P., Baker, T. E., Warren, C., & Li, H. (2016). Oscillatory profiles of positive. negative and neutral feedback stimuli during adaptive decision making. International Journal of Psychophysiology.107:37-43.

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

[10] Zheng Zhenlong, Lin Jung. (2015). Pricing bias and investor sentiment of Shanghai and Shenzhen 300 stock index futures. Mathematical statistics and management. 34 (6):1129-1140(in Chinese).

[11] Dowie G , Willows G.(2016). An investigation of investors' estimates of returns earned and the effect of anchoring on these estimations. South African Journal of Accounting Research.30(1):29-40.

[12] Lu ping, Leng Jun. (2017). Investor sentiment and earnings management: an empirical study based on accrued earnings management and real earnings management. Financial and economic issues. (2): 88-96(in Chinese).

[13] Arajo Tanya, Eleut Rio Samuel, Louang Francisco. (2018). Do sentiments influence market dynamics? A reconstruction of the Brazilian stock market and its mood. Physica A: Statistical Mechanics and its Applications, 505: 1139-1149.

[14] Yanfeng, Du Tengfei, Mao Jianhua, Liu Xuefeng. (2017). Emotional analysis of stock market text based on emotional dictionary and LDA model. Electronic measurement technology. (12):82-87(in Chinese).

[15] Smales, L. A. (2015). Time-variation in the impact of news sentiment. International Review of Financial Analysis.37:40-50.

[16] Oliveira, N., Cortez, P., and Areal, N. (2016). Stock market sentiment Lexicon Acquisition using microblogging data and statistical measures. Decision Support Systems. 85: 62-73.

[17] Wei, Y. C., Lu, Y. C., Chen, J. N. Hsu, Y. J. (2016). Informativeness of the market news sentiment in the Taiwan stock market. The North American Journal of Economics and Finance.39: 158-181.