

# ***Research on How Social Media Sentiment Affects the Stock Market***

**Yixuan Cao<sup>1,a,\*</sup>**

<sup>1</sup>*School of Management Science and Engineering, Shandong University of Finance and Economics,  
Jinan, Shandong, 250014, China*

*a. 202206240125@mail.sdufe.edu.cn*

*\*corresponding author*

**Abstract:** With the rise of social media, investors are increasingly focused on the expression of emotion on social media. Studies have shown that social media sentiment can significantly affect all aspects of the stock market, such as stock yields, trading volume, market volatility, etc. However, there are still some shortcomings, such as a deep understanding of emotional transmission mechanisms, limitations of emotion measure methods, and less research on long-term effects. This paper covers the state of the research and current trends while examining the influence of sentiment on the social media platforms. This review collected 30 related journal articles through a systematic search in the Web of Science and CNKI databases. This paper also summarized the research status and trend of the impact of social media sentiment on the equity market. These studies are significant for investment decision-making, market volatility prediction, and risk management. But it also calls for future research to further explore the emotional communication mechanisms and long-term effects to more fully understand The effect of mood on social media on the equity market.

**Keywords:** Social media, sentiment analysis, stock market

## **1. Introduction**

The impact of sentiment expressed on social media on the equity market has gradually become one of the hot spots in the current economics market research. In recent years, with the rise of social media platforms such as Twitter, Weibo, and Xiaohongshu, many users share their views on the equity market on these platforms. A large amount of text data in social media has become a new source of dynamic information that can reflect public sentiment and market trends.

With the popularity of social media and the rapid dissemination of information, investors are increasingly paying attention to emotional expression on social media platforms, believing that these emotions reflect the emotions and views of market participants. Previous research has yielded the following findings: incorporating the particular dimension of public sentiment can notably enhance the accuracy of forecasting the Dow Jones Industrial Average; investor sentiment forecasts the value premium; social media post sentiment can anticipate stock returns for future trading days; pre-market investor sentiment predicts the opening price; investor sentiment exerts a significant positive influence on stock returns, trading volume, and order imbalance in bulk trading; and the implicit sentiment expressed in news and forums impacts stock prices. The examination of social

media emotion's impact on the stock market holds significant research implications across various domains such as market forecasting and investment decision-making, market volatility and risk management, trading strategy formulation, algorithmic trading, market behavior analysis, and market efficiency assessment. Research methods encompass emotional index construction, analysis of influence mechanisms, development of trading strategies, and analysis of market behavior, among others. With the ongoing advancement of technology, an increasing number of researchers are leveraging big data from social media platforms to conduct sentiment analysis. Their aim is to delve into the correlation between social media sentiment and stock market trends, yielding some notable achievements. However, there are still some research gaps, such as the current understanding of the emotional transmission mechanism is not deep enough. There are some limitations in the understanding of polys, implicit emotions, and specific context in coping text. As social media data changes from time to time, generating a huge amount of data, most of the existing research focuses on the short-term impact of social media sentiment on the stock market, with little research on the long-term impact. Some studies have considered other factors (search index, macroeconomic indicators, etc.), and have not deeply explored them. There may be a complex interaction relationship between social media emotions (SME) and other factors, which needs to be further explored.

To address the influence of social media mood on the equity market in recent years, this paper examined thirty pertinent, authoritative journal publications and methodically compiled pertinent research. This paper aims to systematically summarize recent studies on the influence of social media sentiment on the stock market, point out the shortcomings of existing research, and propose directions for future improvement.

## 2. Primary Source of the Literature

CNKI, developed by CNKI, is one of the largest academic literature databases in China, covering a wealth of Chinese journals, doctoral and master papers, conference papers, and other resources. It has a wide application and influence in Chinese academic circles, especially in Chinese literature retrieval. Operated by Clarivate Analytics, Web of Science is one of the most authoritative academic literature retrieval platforms in the world, including academic journals, conference papers, and patents from all over the world. Its retrieval system is more complex and sophisticated, supporting a variety of advanced retrieval functions and citation analysis, and is widely recognized as one of the important tools of academic research evaluation and scientific research index evaluation. This work gathers 30 journal articles by searching for keywords like "investor sentiment," "stock market," "social media," "stock price," and "sentiment indicators" using the Web of Science and CNKI database as data sources. The study on whether social media sentiment affects the stock market started early. With technological progress, refined machine learning techniques, advancements in natural language processing, and sophisticated models, the precision of forecasting stock market fluctuations through sentiment data has been consistently improving. Due to the lockdown caused by COVID-19, the influence of social media sentiment on the equity market has been increasingly valued by researchers, so the largest number of papers in 2019 was 6, with a slightly downward trend since 2020 (Figure 1).

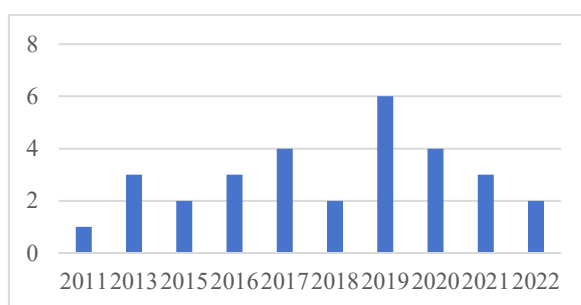


Figure 1: Trends of related publications (Photo/Picture credit: Original).

### 3. Research on the Influence of Social Media Sentiment on the Equity Market

There have been many recent studies on social media sentiment analysis, which usually employ natural language processing techniques to analyze and quantify emotions in social media texts. This paper divides 30 journal articles into three time stages: 2011-2016, 2017-2019, and 2020-2022. It finds that with the progress of technology, the influence of social media mood on the equity market is more and more widespread, and the forecast rate is also steadily improving.

#### 3.1. Selected Data Sources and Model Methods Adopted Between 2011 and 2016

Table 1: Related research on the stock market from 2011-2016

Literature	Data sources	Model algorithm	Superiority	Limitations
[1]	Twitter	Autoregressive conditional heteroscedasticity model	Taking into account the impact of social media sentiment; the dynamic relationship between mood index and equity market index; the heteroscedasticity and autocorrelation of time series data	Depends on the accuracy of emotional analysis; ignores other influencing factors; may not be able to fully capture the complex dynamic features in the data
[2]	Bloomberg Database	Principal component method	Obvious effect of dimensionality reduction; elimination of multicollinearity; retention of data information	Information loss; hypothetical linear relationship; principal component; difficulty in interpretation; sensitivity
[3]	Sina Weibo	Granger causality Test and Impulse response function method	Strong statistical validity, the principle is simple, the causal direction between variables can be determined, and the results are intuitive and easy to explain.	High data requirements; inability to determine the mechanism of causality; existence of lag effect; possible influence of exogenous variables
[4]	Eastern wealth net stock	K-nearest neighbor algorithm	Simple and easy to implement; non-parameterized; suitable for multi-classification problems; strong adaptability	High computational complexity; high storage overhead; slow prediction speed; sensitive to outliers
[5]	Sina	Support vector	Efficient processing of	Sensitive to missing data;

Table 1: (continued).

	Weibo	machine model	high-dimensional data; strong generalization ability; effective processing of small sample data sets; strong anti-interference ability; can flexibly adapt to different problems by adjusting hyperparameters	complex tuning of model parameters; sensitive to the selection of kernel functions; not suitable for large-scale multi-category problems
[6]	Eastern wealth net Shanghai index bar Posting content	Vector autoregressive model(VAR)	Suitable for multivariable analysis; does not need to preprocess data; strong interpretability; provides future prediction ability	Vulnerable to external factors; need to meet some assumptions; high data requirements; unable to deal with nonlinear relationships
[7]	Weibo related to Yu'e Bao	Regression prediction model	Strong interpretability; wide applicability; provide probability prediction	Lag effect; the model hypothesis is limited; unable to deal with external interference factors.
[8]	Sina Weibo	Regression prediction model	Strong interpretability; simple and intuitive; high controllability; wide applicability	There is a lag effect; the linear hypothesis is limited; data quality is uncertain
[9]	the gem of Eastern wealth Internet shares Bar	Regression prediction model	Analysis of specific stock market, close to the actual investment environment; timely access to market sentiment; diversified sources of information; high controllability	The quality of information varies; there is the possibility of market manipulation; it is difficult to capture emotional and semantic information; the lag effect affects the results of analysis.

As shown in Table 1, summing up these studies, we can see that different methods are applied to analyze the relationship between social media data and equity market indexes. These methods include a time series model, principal component analysis, causality test, machine learning algorithm, and regression prediction model. Each method has its advantages and limitations. Overall, these studies suggest that social media data can be used as an auxiliary tool to help analyze the rise and fall of stock market indices. However, these methods have some common challenges, such as the accuracy of emotional analysis, the reliability of data quality, the robustness of the model, and the influence of external interference factors. Therefore, when linking social media data to stock market indices, we need to consider a variety of factors and combine different analysis methods to more accurately understand the relationship between them. Meanwhile, the results of each method need to be carefully explained and verified to ensure its reliability and effectiveness in the actual investment decision.

### 3.2. Selected Sources of Data and Model Methods Adopted in 2017-2019

Table 2: Research on social media sentiment on the stock market from 2017 to 2019

Literature	Data sources	Model algorithm	Superiority	Limitations
[10]	Twitter	hypothesis test	Objectivity; scientific; replicability; quantitative analysis; statistical significance	Limitations; data limitations; causality difficult to determine; time delay; model simplification, incomplete consideration
[11]	Twitter	VAR	Consider dynamic relationship among multivariates; no causality; provide impact response analysis; wide applicability	Uncertain data quality; high model complexity; prone to overfitting; inability to handle non-linear relationships
[12]	Sina Weibo, the Shanghai Composite Index stock comments	SVR technology	Suitable for nonlinear relationships; robust to outliers; flexible dimension; and strong generalization ability	The parameter selection is complex; has high calculation complexity; is sensitive to missing data and has a poor interpretation
[13]	Youkuang Finance's stock bar forum	Correlation analysis and VAR model	Multidimensional data analysis; improved accuracy; real-time and diversity; strong interpretability	Difficult to guarantee data quality and authenticity; difficult to determine causality; high model complexity; hysteresis effect and time series characteristics
[14]	Comment post of Shanghai and Shenzhen 300 stocks in Eastern Wealth Internet Stock Bar	The Naive Bayesian model	Simple to understand; high calculation efficiency; good adaptability to sparse data; suitable for text classification	A mutually independent reality may not hold; requires the quality and accuracy of input data; fails to handle complex relationships between concepts; relies on prior probabilities
[15]	WeChat public account and the stock market-related push	VAR model for pairwise time series	Consider dynamic relationships; strong controllability; high data availability; and strong interpretable results	Data is difficult to select and process; difficult to determine causality; high model complexity and external factors
[16]	Stock reviews on Seeking Alpha	Weighted prediction model (WPM)	Comprehensive assessment of multiple sources; considering author influence; high flexibility; interpretability	Subjectivity; data credibility cannot be guaranteed; high model complexity; data update delay

Table 2: (continued).

[17]	Sina Weibo	WPM	Strong real-time; large data volume; diversity; interactivity	Different information quality; noise interference; emotional deviation; and difficult data processing
[18]	Data on the financial world's website	Regression prediction model	High professional authority; high data integrity ; timeliness	Information overload; uneven data quality; emotional bias impact; data limitations
[19]	Oriental Fortune, Sina Finance, and NetEase Finance.	LSTM model	Consider text sequence relationships; adapt to different types of data; consider semantic information; and dynamic update model	Influence of data noise; subjectivity; data sparsity; poor interpretation
[20]	Twitter	CAPM	Provide theoretical framework; widely used and easy to implement; and consider risk factors	Hypothesis limitation; consider only a single factor; data selection may have subjectivity and error; time scale mismatch

As shown in Table 2, the literature from 2017 to 2019 covers many aspects and methods of the stock market forecast model. They use various data sources, including social networking platforms (such as Twitter and Sina Weibo), financial news websites (such as Seeking Alpha and financial websites), stock forums (such as Oriental Fortune network shares), WeChat official accounts, etc. The model methods include hypothesis testing, vector autoregressive model (VAR), support vector regression (SVR), naive Bayes model, weighted prediction model, regression prediction model, and long-and short-term memory network (LSTM) model.

In terms of advantages, these models include objectivity, science, replication, multidimensional data analysis, real-time, diversity, extensive adaptability, considering dynamic relationships, considering text sequence relationships, and so on. In addition, some models also have the advantages of strong interpretability, flexible dimensions, strong generalization ability, high computational efficiency, high data integrity, and strong authority.

However, these models also have some common limitations, such as uncertain data quality, data limitations, high model complexity, hypothesis constraints, difficult-to-determine causality, subjectivity, data update delay, uneven information quality, noise interference, emotional bias, data processing difficulty, poor interpretability, etc. Moreover, some models have specific limitations, such as complex parameter selection, high computational complexity, sensitivity to missing data, sensitivity to data sparsity, and inability to handle non-linear relationships.

In conclusion, the stock market prediction model has made some progress in improving the prediction accuracy, real-time performance, and interpretability, but still faces challenges in the face of data quality, complexity, interpretability, and limitations. Hence, when choosing and implementing a stock market prediction model, it's essential to carefully weigh both the superiority and limitations of the model, and make informed selections and adjustments according to the specific circumstances at hand.

### 3.3. Selected Data Sources and Model Methods Adopted from 2020-2022

Table 3: Research on the stock market from 2020 to 2022

Literature	Data sources	Model	Superiority	Limitations
[21]	Two stock review datasets in English and Persian	Implied Dirichlet Allocation Model (LDA)	Topic modeling; dimension reduction; language relevance	Number of a priori topics; subjectivity and complexity
		SVM model	High accuracy; wide applicability; processing of high-dimensional data	Sensitive to parameters; inefficient in processing large-scale data
[22]	Eastern wealth net shares bar	Improved Bayesian algorithm	good at processing small samples; strong adaptability; improved model accuracy; and improved data processing efficiency	Data sparsity; selection of a priori assumptions; high model complexity
[23]	Internet forum	Point mutual information log (LNPMI)	Consider lexical relevance; reduce the impact of noise; high flexibility	High computational complexity; data sparsity; dependent on the corpus
[24]	Eastern currency in an online stock forum	LSTM model	Consider contextual information; deal with long sequence data; strong memory ability and strong adaptability	Large data demand; super-parameter tuning; over-fitting risk; black-box model is difficult to explain
[25]	Media comments and responses related to HHPIC	Logistic regression (LR)	The short answer is easy to explain; has fast calculation speed; wide applicability; feature selection ability	Linear assumption limitation; sensitive to outliers; unable to handle complex relationships; need to manually select features
[26]	Sina stock	VAR model	Consider multivariate relationships; without preset function form; dynamic; and strong interpretability	High data demand; dimension exponential growth; over-fitting risk; need lag item selection
		BP neural network	Adapt to the nonlinear relationship; flexibility; parallel processing ability; self-learning ability	The black box model is difficult to explain; requires a large amount of data; difficult to tune parameters; and easy to fall into a local optimal solution
[27]	The economic news database	word2vec algorithm	Consider multivariate relationships; without preset function form; dynamic; and strong interpretability	High data demand; dimension exponential growth; over-fitting risk; need lag item selection
[28]	News articles and	LSTM model	Capture long-term dependencies; processing	High data demand; risk of overfitting; difficulty in

Table 3: (continued).

	forum posts on the PTT platform		sequence data; dynamic learning; processing variable-length input sequences	parameter tuning; difficulty in interpreting the black-box model
[29]	Eastern wealth stock bar comment	VAR model	Consider multivariate relationships; without preset function form; dynamic; and strong interpretability	High data demand; dimension exponential growth; over-fitting risk; need lag item selection
[30]	Stock market-related news headlines, Twitter tweets, financial news articles, Facebook comments	Henry	Quickly identify emotional tendency; simple and easy to use and suitable for rapid screening	Vcible to semantic ambiguity; limited processing capacity for long text
		LR	Suitable for binary classification problems, with fast calculation speed and efficient model training and prediction	The weak fit to nonlinear relationships is weak to capture the complex emotions in the text
		Loughran–McDonald	The emotional tendency is identified more accurately, and the lexical features of specific fields are taken into account	Can only identify the emotional tendencies, lack of fine division
		VADER	Emotional analysis of social media data; considering the intensity of emotional words and emotional modifiers	The effect of emotion analysis for specific fields is not accurate enough
		TextBlob	It supports multiple emotion analysis functions; it can be better applied in financial news articles and Facebook reviews	The effect of emotion analysis for specific fields is not accurate enough
		Linear SVC	Suitable for handling high dimensional data and nonlinear classification problems; good performance in text classification tasks, strong model generalization ability	Extensive training data and feature engineering are required
		Stanford	Ability to learn the complex features and semantics of text data; suitable for analyzing various types of text data	Require a lot of computing resources and time, high model complexity; overfitting to small-scale data sets

As shown in Table 3, the literature from 2020-2022 summarizes the various model methods, strengths, and limitations used in the field of sentiment analysis in the stock market. These

approaches encompass conventional statistical models like the Bayesian algorithm, and VAR model, as well as decision-making tools such as SVM, LR, LSTM, among others. Overall, these model approaches have individual strengths and weaknesses in handling sentiment analysis tasks:

Superiority:

Statistical models such as the VAR model and Bayesian algorithm can consider multivariate relationships without a preset functional form and have strong dynamics and interpretability.

Machine learning models such as SVM, LR, LSTM, etc. can deal with nonlinear relationships, are adaptable, and perform well in handling long sequence data and considering contextual information.

Some models such as the word2vec algorithm can consider the lexical correlation, reduce the influence of noise, and have the advantage of high flexibility.

Utilizing a combination of multiple sentiment analysis models enables the amalgamation of diverse model strengths, thus enhancing the accuracy and robustness of sentiment analysis.

Boundedness:

The statistical model may have problems such as high data demand, exponential growth of dimension, risk of overfitting, and selection of lag terms.

Machine learning models require plenty of data to train, and parameter tuning is difficult, sometimes falling into the local optimal solution, and black-box models are difficult to explain.

Some models perform poorly in handling data sparsity, outliers, and complex relationships, requiring manual feature selection or data preprocessing.

Emotional analysis models have some challenges in dealing with the diversity of language context and emotional expression, and sometimes there may be miscalculation or failure to accurately capture emotional information.

In conclusion, the selection of sentiment analysis models suitable for specific tasks requires comprehensively considering factors such as data characteristics, task requirements, model performance, and interpretability, to achieve better analytical results.

## 4. Conclusions

This paper uses a systematic review and finds that social media sentiment data has become an important aspect of financial market analysis, with different studies using diverse data sources, analytical techniques, and models to reveal the complicated relationship between social media mood and the equity market. While facing challenges such as data accuracy and model stability, the accuracy and reliability of stock market prediction utilizing social media sentiment data are anticipated to advance further with technological progress and the evolution of data analysis methodologies. Future studies could further deepen the integration and analysis of emotional data from different social media platforms, and improve the accuracy and stability of prediction models. At the same time, it can also consider combining other financial data and macroeconomic indicators to build a more comprehensive, multi-dimensional, and multidimensional stock market forecast model to provide more reliable decision support for investors. Through ongoing deepening and refinement of pertinent research, there will be a heightened comprehension and application of how social media affection influences the equity market, thus ushering in increased innovation and development opportunities within the investment realm.

## References

- [1] Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1-8.

- [2] Zhang, T., Yu, J., Lu, D. K. (2013). *Emerging market investor sentiment and value premium difference -- is based on a comparative analysis of mainland China, Hong Kong and Taiwan*. *International Finance Research*, (01), 87-95.
- [3] Cheng, W., Yun, L. J. (2013). *Social media investor sentiment and the stock market index*. *Management Science*, (05), 111-119.
- [4] Jin, X. J., Zhu, Y., Yang, X. L. (2013). *The influence of network media on the stock market -- An empirical study of taking the Oriental Fortune network stock bar as an example*. *Journalism and Communication Research*, (12), 36-51 + 120.
- [5] Huang, R. P., Zuo, W. M., Bi, L. Y. (2015). *Stock market forecast based on microblog sentiment information*. *Journal of Management Engineering*, (01), 47-52+215.
- [6] Yi, H. B., Lai, J. J., Dong, D. Y. (2015). *The impact of different investor sentiment on the trading market -- Empirical analysis based on VAR model*. *Financial theory cluster*, (01), 46-54.
- [7] Zhu, M. J., Jiang, H. X., Xu, W. (2016). *Stock price forecast based on the emotion and communication effect of financial microblog*. *Journal of Shandong University (Science edition)*, (11), 13-25.
- [8] Chen, X. H., Peng, W. L., Tian, M. Y. (2016). *Research on stock price and volume forecasts based on investor sentiment*. *Systems Science and Mathematics*, (12), 2294-2306.
- [9] Yang, X. L., Shen, H. B., Zhu, Y. (2016). *Local preferences, investor sentiment, and stock yields: empirical evidence from online forums*. *Financial research*, (12), 143-158.
- [10] Sul, H. K., Dennis, A. R., Yuan, L. (2017). *Trading on twitter: Using social media sentiment to predict stock returns*. *Decision Sciences*, 48(3), 454-488.
- [11] Oliveira, N., Cortez, P., Areal, N. (2017). *The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*. *Expert Systems with Applications*, 73, 125-144.
- [12] Dong, L., Wang, Z. Q., Xiong, D. Y. (2017). *Stock index forecast based on text information*. *Journal of Peking University (Natural Science Edition)*, (02), 273-278.
- [13] Shi, Y., Tang, J., Guo, K. (2017). *Social media investor attention and the impact of investor sentiment on the Chinese stock market*. *Journal of the Central University of Finance and Economics*, (07), 45-53.
- [14] Bu, H., Xie, Z., Li, J. H., Wu, J. J. (2018). *The impact of investor sentiment based on stock reviews on the stock market*. *Journal of Management Science*, (04), 86-101.
- [15] Shi, S. C., Zhu, Y. N., Zhao, Z. G., Kang, K. L., Xiong, X. (2018). *Investor sentiment and stock market performance based on wechat text mining*. *System Engineering Theory and Practice*, (06), 1404-1412.
- [16] Xiao, T., Lin, L., Huang, Y. F. (2019). *A method of stock market trend prediction based on stock sentiment analysis*. *Application of Electronic Technology*, (03), 13-17.
- [17] Zhao, M. Q., Wu, S. Q. (2019). *Research on Stock Market Weighted Forecasting Method Based on Weibo Emotion Analysis*. *Data Analysis and Knowledge Discovery*, (02), 43-51.
- [18] Yin, H. Y., Wu, X. Y. (2019). *The high frequency of investor sentiment on the stock day yield forecast effect*. *China's industrial economy*, (08), 80-98.
- [19] Cen, Y. H., Tan, Z. H., Wu, C. Y. (2019). *Research on the influence of Financial Media Information on the Stock Market: an empirical evidence based on sentiment analysis*. *Data Analysis and Knowledge Discovery*, (09), 98-114.
- [20] Broadstock, D. C., Zhang, D. (2019). *Social-media and intraday stock returns: The pricing power of sentiment*. *Finance Research Letters*, 30, 116-123.
- [21] Derakhshan, A., Beigy, H. (2019). *Sentiment analysis on stock social media for stock price movement prediction*. *Engineering applications of artificial intelligence*, 85, 569-578.
- [22] Liang, S. L., Chen, Y. X., Chen, P. P., Sun, L. M. (2020). *Research on stock market prediction based on social emotion data mining*. *Journal of Northeast Normal University (Natural Science Edition)*, (03), 105-110.
- [23] Huang, C. X., Wen, S. G., Yang, X., Wen, F. H., Yang, X. G. (2020). *Study on the interactive relationship between individual investor sentiment and stock price behavior*. *Chinese Management Science*, (03), 191-200.
- [24] Wang, G., Yu, G., Shen, X. (2020). *The effect of online investor sentiment on stock movements: An lstm approach*. *Complexity*, 2020, 1-11.
- [25] Huang, J. Y., Liu, J. H. (2020). *Using social media mining technology to improve stock price forecast accuracy*. *Journal of Forecasting*, 39(1), 104-116.
- [26] Zhan, B. Q., Qu, B. Y. (2021). *Analysis and forecast of the impact of market sentiment on stock movements*. *Technology and Industry*, (04), 51-57.
- [27] Jiang, F. W., Meng, L. C., Tang, G. H. (2021). *Media text sentiment and stock return forecasts*. *Economics (Quarterly Journal)*, (04), 1323-1344.
- [28] Ko, C. R., Chang, H. T. (2021). *LSTM-based sentiment analysis for stock price forecast*. *PeerJ Computer Science*, 7, e408.

- [29] Lu, W. B., Zhang, M., Zheng, T. Z. (2022). *Can the stock review information predict the downside risks in the stock market?. Statistics and Information Forum*, (08), 53-66.
- [30] Das, N., Sadhukhan, B., Chatterjee, T., Chakrabarti, S. (2022). *Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak. Social network analysis and mining*, 12(1), 92.