An empirical study of the relationship between investor sentiment and stock market performance based on support vector machine modeling

Suiyu Yang[1, *], Jikai Zhang[2]

[1]School of Information Science and Technology (School of Cyber Security), Guangdong University of Foreign Studies, Guangzhou, China, 510006

[2]Department of Digital Economics, Shanghai University of Finance and Economics, Shanghai, China, 200433

*Corresponding author: sueyyang1@163.com

Abstract. This study explores the impact of investor sentiment on stock market performance and empirically analyzes the value of its application in market prediction by constructing a sentiment classification system using Support Vector Machine (SVM) model. By collecting market-related text data, extracting investor sentiment features using text sentiment analysis, and combining them with stock market data, a machine learning-based prediction model is constructed. The experimental results show that investor sentiment has a significant impact on the volatility of the stock market, and the support vector machine model shows high accuracy and applicability in sentiment classification and market prediction. Compared with traditional sentiment analysis methods, the machine learning method used in this study can capture the impact of sentiment fluctuations on market movements more effectively, providing a new technical means for financial market sentiment analysis. The research results can provide investors with a more refined basis for market judgment, and can also serve as a reference for financial regulators to formulate risk warning strategies. Future research can further optimize the sentiment analysis model, improve the real-time data processing, and explore multimodal data integration to enhance the accuracy and stability of market sentiment prediction.

Keywords: Investor Sentiment, Sentiment Analysis, Machine Learning, Financial Markets, SVM.

1. Introduction

With the rapid development of the global economy and the continuous expansion of financial markets, the impact of investor sentiment on stock market performance has gradually received widespread attention. Investor sentiment, which refers to investors' emotions and expectations about market performance, is an irrational psychological factor that often leads to increased volatility in the stock market. Especially in the era of information explosion, news reports, social media and market events spread rapidly, which are very likely to trigger investor sentiment fluctuations. In recent years, with the development of natural language processing, machine learning and other technologies, sentiment analysis has gradually become a hot area of financial research. Studies have shown that investor sentiment has a significant impact on the stock market, and its influence has gradually leveled off with pure market fundamentals and technical analysis factors. In China's stock market, the composition of investors is complex, and the mood swings of small and medium-sized investors are more frequent, and these mood swings may amplify the market's reaction, directly or indirectly affecting the performance of the Shanghai Stock Exchange Index (SSEI). Especially during the peak of bull market and trough of bear market, the extent of investor sentiment on stock price far exceeds the influence of company fundamentals on it, accounting for about 60% [1]. How to accurately quantify the relationship between investor sentiment and market performance has become an urgent problem. Therefore, studying the relationship between investor sentiment and stock market performance not only provides a better understanding of the inner mechanism of market volatility, but also provides an important reference value for policy makers and investors.

Currently, research on the relationship between investor sentiment and stock market performance focuses on two directions: sentiment analysis and market forecasting. In the field of sentiment analysis, the mainstream methods include dictionary-based sentiment analysis and machine learning-based sentiment classification. Dictionary methods obtain sentiment indicators by constructing sentiment vocabularies and counting positive and negative sentiment words in the text, but this method is susceptible to the quality of the dictionaries and the update cycle, and it is often difficult to cope with a large amount of complex text data. In contrast, machine learning methods can be better adapted to different text environments by annotating text data and training models, especially in social media data, machine learning methods show better performance.

This study adopt the `Linear SVC` model of SVM models, aiming to empirically analyze the correlation between investor sentiment data and the performance of the SSE index. `Linear SVC` is a linear support vector machine classifier that can be effectively applied to linearly differentiable data and has high computational efficiency and stability on large-scale datasets. Since stock market fluctuations tend to have strong linear characteristics, the `Linear SVC` model is able to better portray investor sentiment, thus providing an effective quantitative tool for market analysis.

By constructing a quantitative sentiment model based on `Linear SVC`, this paper hopes to achieve the following objectives: first, to quantify the actual impact of sentiment volatility on the stock market; second, to provide an efficient and accurate model of the relationship between sentiment and market performance; and finally, to provide a new research idea for the analysis and prediction of sentiment in the financial market. The significance of this study is that it not only expands the application of sentiment analysis in financial forecasting, but also provides investors and policy makers with theoretical support and practical tools for identifying and managing market sentiment volatility.

2. Literature Review

In recent years, sentiment analysis has emerged as a key tool for improving stock market prediction by extracting investor sentiment from unstructured text data. Traditional methods, such as time-series analysis and regression models, often overlook sentiment factors, while studies demonstrate that investor sentiment significantly influences stock price fluctuations.

Among sentiment analysis techniques, dictionary-based and machine learning approaches each have strengths and limitations, with deep learning becoming a major research direction. For instance, Lin et al. [3] proposed the SCONV model, combining convolutional LSTM with sentiment analysis, achieving stable prediction accuracy even on small datasets. Similarly, Xu et al. [4] enhanced stock index forecasting by integrating financial news sentiment (analyzed via BERT) with trading data in their SA-BERT-LSTM model. Other studies further optimized performance through model innovations: Kumar et al. [5] improved short-text sentiment classification using VSM, while Yang et al. [6] integrated CNN and BiLSTM for better feature extraction. Hui et al. [7] employed a Naïve Bayes model to validate sentiment's predictive power over opening/closing prices and trading volume.

Recent advances focus on hybrid techniques for efficiency gains. Rizinski et al. [8] combined Transformer with SHAP in the XLex model, boosting interpretability and real-time analysis. Peivandizadeh et al. [9] addressed class imbalance in sentiment analysis via TLSTM and PPO algorithms, achieving top-tier prediction metrics.

Despite improved stability through multi-method integration, challenges persist, including high model complexity and data dependency. Future research should optimize architectures to enhance real-time sentiment analysis accuracy, offering more robust support for market decision-making.

3. Model building

This study constructs a support vector machine-based sentiment classification model process, which starts from stock review data collection, goes through data preprocessing, feature extraction, model training and validation, and finally realizes sentiment analysis and index calculation, providing

a systematic technical path to study the relationship between investor sentiment and stock market performance. As shown in Figure 1:

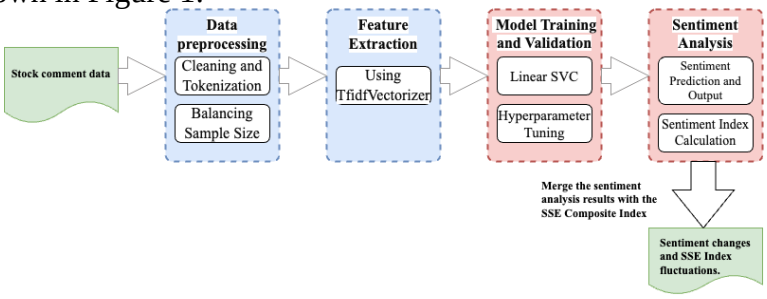

Figure 1. Sentiment Classification Model Flowchart

3.1. Basic concepts of the model

Support Vector Machine (SVM) is a supervised learning algorithm commonly used for classification tasks, and its basic idea is to separate different categories of data by finding the optimal hyperplane for the purpose of classification [11].The advantage of SVM is that it is able to map a low-dimensional nonlinear problem into a high-dimensional space through the introduction of kernel functions, which The advantage of SVM is that it can map a low-dimensional nonlinear problem into a high-dimensional space by introducing a kernel function, which makes the data linearly separable in the high-dimensional space. In addition, Linear SVM (Linear SVC) focuses on dealing with linearly differentiable problems with high computational efficiency and simple model parameters, which is especially suitable for the task of sentiment categorization in large-scale datasets.

In the field of sentiment analysis, after the text data is numerically transformed into feature vectors, the classification of sentiment polarity (e.g., positive, negative) can be achieved by SVM models. The feature vectors are generally generated using the TF-IDF (Word Frequency-Inverse Document Frequency) method.TF-IDF improves the classification performance of the model by measuring the importance of words in a document, effectively reducing the influence of commonly used but low-information words [12]. Combined with SVM classifiers, the text features extracted by TF-IDF can be used to construct sentiment classifiers for accurate portrayal of investor sentiment.

This study choose Linear SVC as the base model, aiming to utilize its fast and efficient classification ability to quantify the investors' emotional state in stock review data as an indicator of emotional polarity, which lays the foundation for the subsequent analysis of the relationship between emotional fluctuations and stock market performance.

3.2. Sentiment Classifier Building Based on Vector Machine Modeling

3.2.1 Data loading and preprocessing

First of all, constructing the sentiment dataset is the foundation of the entire model development. In this study, representative data from the SSE index and user comments from East Money Information Co., Ltd. stock bar during the period from 2020 to 2021 were selected as the research sample. These data were collected through web crawlers and manual retrieval. A portion of the comments were manually labeled with sentiment tags as positive or negative to serve as training samples for sentiment analysis.

The period of 2020–2021 was chosen due to the high market volatility and increased investor participation influenced by global events such as the COVID-19 pandemic. [10]This period provides a rich context for observing the impact of investor sentiment on stock market performance. Moreover, the selected sample is representative of typical investor behavior in a period of intensified market dynamics, making it suitable for validating the effectiveness and robustness of the sentiment-based predictive models.

In order to improve the performance of the model, the loaded data are processed as follows:

Cleaning and Segmentation: The text is cleaned of redundant spaces, symbols, and irrelevant information, and each text is subject to segmentation so that the subsequent feature extraction process can directly deal with the list of words after segmentation.

Balancing the number of samples: Considering that the dataset may have the problem of category imbalance, for example, the number of positive emotion samples is much larger than that of negative emotion samples, in order to avoid the bias towards most categories of samples in the model training, this study performs balanced interception according to the minimum number of positive and negative emotion samples, so as to ensure that the classifier has the ability to recognize the positive and negative emotions in the two categories of emotions.

Label generation: label 1 is assigned to the positive samples and label 0 to the negative samples, and the corresponding label vector y is generated for the model to perform supervised learning.

3.2.2 Feature Extraction

Feature representation transformation is performed on the text data, and the text after word separation is used as input to generate a sparse feature matrix based on TF-IDF. This process aims to retain the semantic information of the text, while weakening the influence of high-frequency but less informative words on the model and improving the overall generalization ability.

3.2.3 Model Training and Validation

After feature extraction is completed, the text data is converted into a feature matrix, and the sentiment labels correspond to the target variables. In order to improve the generalization ability of the model, this study adopts a multi-level training and validation strategy.

First, the dataset is divided into training and testing sets in the ratio of 8:2 to ensure that part of the data is retained for final model evaluation, and the reproducibility of the experimental results is ensured by fixing the random seeds. On this basis, a five-fold cross-validation method is adopted, where the training data are divided into five subsets, four of which are selected for training each time, and the remaining one is used for validation, cycling several times to fully utilize the limited data.

In addition, the hyperparameter search method is used to optimize the key model parameters, including:

C value (penalty coefficient): four values of 0.1, 1, 10, and 100 are tried to regulate the regularization strength of the model

class_weight (class weight): test both None and 'balanced' settings to deal with possible class imbalances

loss (loss function): compares the effects of both 'hinge' and 'squared_hinge' loss functions

max_iter (maximum number of iterations): set to 1000 and 2000 options to ensure full model convergence

During the hyper-parameter search process, F1 score, an evaluation metric that integrates precision and recall, is used to comprehensively measure the performance of the classification model on the unbalanced dataset. In addition, the parameter search process is accelerated by parallel computing to fully utilize the computational resources and improve the search efficiency.

Based on the grid search results, the optimal parameter combination is selected to train the Linear SVC model with the best performance. The model achieves efficient classification by finding the decision boundary that maximizes the distance between categories. Appropriate optimization strategies are used during training to improve the computational efficiency on large-scale datasets to ensure that the model has high computational performance while guaranteeing the classification effect.

3.2.4 Sentiment Prediction and Output

After training is completed, the trained classifier is utilized to predict the sentiment of the new stock review data. The prediction results were quantified as binary sentiment labels (1 for positive sentiment and 0 for negative sentiment).

3.2.5 Calculation of Sentiment Index

Based on the sentiment classification results, this study further construct a sentiment index to analyze the impact of sentiment fluctuations on stock market performance:

BI Index (Log Proportional Index): For each day's data, the number of positive and negative comments are counted and substituted into the formula to calculate the sentiment index.1 The additive term avoids taking the wrong value of the denominator when the number of comments of a certain category is zero.

$$\mathrm{BI} = \log\left(\frac{1+\text{positive comments}}{1+\text{negative comments}}\right) \quad (1)$$

BI_Simple Index (Simple Index of Difference): uses the ratio of the difference between the number of positive and negative comments per day to the total number of comments as a direct reflection of sentiment tendency.

$$\mathrm{BI\_Simple} = \frac{\text{positive comments}-\text{negative comments}}{\text{positive comments}+\text{negative comments}} \quad (2)$$

4. Experiments

4.1. Data set.

The data related to the SSE index from 2020 to 2021, including the opening price, the high price, the low price, the closing price and the turnover volume, were obtained by means of manual search. In addition, the data of comments posted by users, including the posting time (created_time) and the content of comments (title), were collected from the Oriental Fortune stock bar by means of web crawler. These comments reflect users' views and emotions on the stock market, and some of these samples have been manually labeled with sentiment labels for subsequent sentiment analysis research.

4.2. Data Preprocessing

This study first preprocesses the raw sentiment text data by separately loading and converting the positive and negative samples into a participle list format. Each sample is assigned a corresponding label based on its category, with positive samples marked as 1 and negative samples marked as 0. To ensure dataset balance, the number of samples in the smaller category is used as a reference, and random sampling is applied to the larger category so that both categories contain an equal number of samples.

In the feature extraction stage, a method based on term frequency and inverse document frequency (TF-IDF) is employed to select 3,000 words with high information content as features. A predefined list of stopwords is used to filter out common, meaningless terms. After feature extraction, the dataset is randomly split into training and test sets, with the training set comprising 80% of the data and the test set comprising 20%. A fixed random seed is set to ensure the stability and reproducibility of experimental results.

4.3. Evaluation Metrics

In the experiments, this study use a variety of evaluation indexes to measure the model performance comprehensively. For model selection, LinearSVC model is used and hyper-parameter optimization is performed by grid search. A 5-fold cross-validation (KFold) is used in the optimization process, and the F1-score is used as the main scoring index to better balance the imbalance of positive and negative samples.

In the model performance evaluation, the accuracy, precision, recall, and F1-score metrics are calculated. Accuracy represents the proportion of correctly predicted samples among all samples; precision denotes the proportion of true positive samples among the predicted positive samples; recall reflects the proportion of true positive samples that are correctly identified; and the F1-score, which is the harmonic mean of precision and recall, is calculated using the following formula:

$$F1-Score=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (3)$$

In addition, the classification of the model's prediction results, including True Positive, True Negative, False Positive, and False Negative, is also demonstrated through the confusion matrix.For the LinearSVC model, the study also extracted the most discriminative feature words through the coef_ attribute, and sorted them according to the feature weights. output the positive and negative feature words that have the greatest impact on the classification results to further understand the decision-making process of the model.

4.4. Analysis of model solving

In the model solving process, the study first loaded the dataset and extracted text features using the TF-IDF method. Subsequently, hyper-parameter tuning was performed using GridSearchCV, and a total of 80 fits were performed on 16 parameter combinations under 5-fold cross-validation.

The best parameter combinations were selected as C = 0.1, class_weight = 'balanced', max_iter = 1000, corresponding to the best cross-validation score of 0.8980.

The evaluation results on the test set show that the model has an accuracy of 0.9099, precision of 0.9021, recall of 0.9142, and an F1-Score of 0.9081, indicating that the model has a high classification performance in the sentiment classification task.

The confusion matrix shows the specifics of the classification (as shown in Fig. 2): 820 for true cases (TP), 857 for true-negative cases (TN), 89 for false-positive cases (FP), and 77 for false-negative cases (FN), which indicates that the model is more balanced in recognizing the positive and negative categories while maintaining a high accuracy rate.

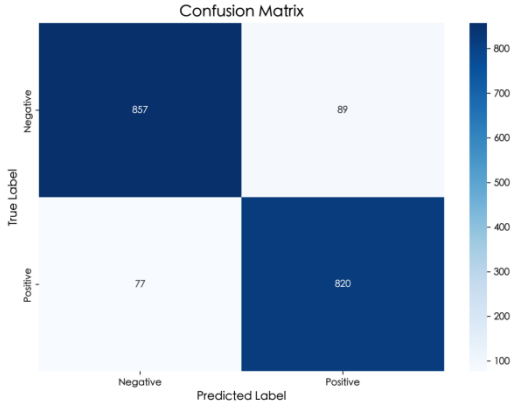

Figure 2. Confusion Matrix Visualization

In addition, by extracting the feature weights from the LinearSVC model, the study identifies the positive and negative feature words that have the greatest impact on classification outcomes. As shown in Figure 3, among the positive features, high-frequency words include "stop," "rise," "up," and "bull," indicating that these terms exhibit strong discriminative power in positive samples. Conversely, among the negative features, high-frequency words such as "down," "drop," "run," and "out" play a significant role in distinguishing negative sentiment.

After completing the model training, the optimal model along with the vectorizer is saved for subsequent use or further optimization.

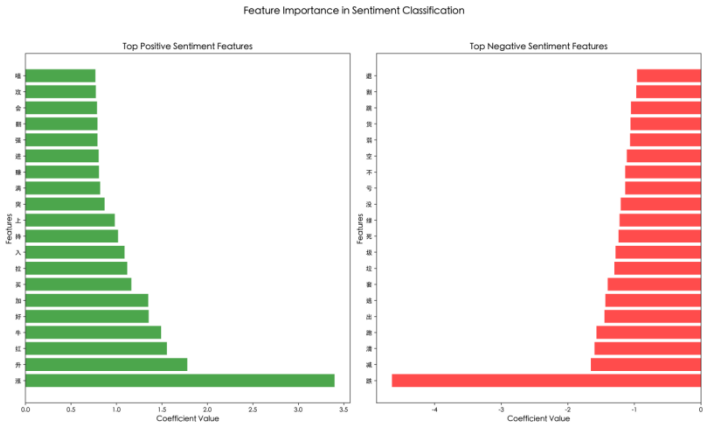

Figure 3. Sentiment Classification Feature Importance Analysis

4.5. Conclusion of the experiment

The purpose of this experiment is to analyze the relationship between the stock comment sentiment index and the closing price of the SSE index and stock trading volume. By constructing a linear support vector machine model to analyze the sentiment of stock review data and using sentiment indices (BI and BI_Simple) to compare with the data of the closing price of the SSE index and stock trading volume, the study explored the trends and potential associations between the two.

First, by calculating the sentiment polarity (positive or negative) of comments, we generated two sentiment indices: the BI index and the BI_Simple index.The BI index uses a logarithmic transformation that compares the proportion of positive and negative sentiment comments, while the BI_Simple index measures changes in sentiment through a simple difference in the proportion of positive and negative comments. Next, the sentiment index BI indicator was merged with the stock's SSE closing price and daily trading volume data, and a 10-day rolling average was applied to smooth out the fluctuations, thus providing a clearer picture of the long-term trend.

The results of merging the sentiment index and stock daily trading volume data, shown in Figure 4, illustrate the trends of the BI indicator and stock trading volume between April 2020 and April 2021. The figure reveals a correlation between the sentiment index and trading volume, with higher trading volumes generally observed when the BI indicator is positive and lower volumes when it is negative. Notably, from July to September 2020, an improvement in market sentiment coincided with an increase in trading volume, suggesting sentiment-driven activity. Conversely, in January 2021, a decline in sentiment preceded a decrease in trading volume, indicating a lagged response in the market.

The experimental results show that there is a certain synchronous and lagged relationship between BI indicator and stock trading volume, and the fluctuation of market sentiment may have an impact on stock trading behavior.

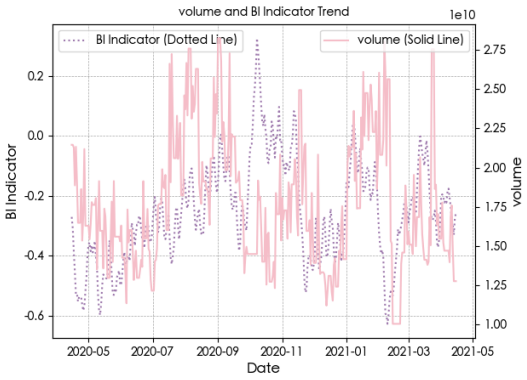

Figure 4. Correlation of Investor Sentiment (BI Indicator) with Stock Trading Volume (April 2020 - April 2021)

Figure 5 illustrates the trend relationship between the BI indicator and the closing price of the SSE index from April 2020 to April 2021. The BI indicator (purple dashed line) and the SSE index closing price (pink solid line) generally show similar volatility patterns throughout most of the period. Both indicators trended upwards in the second half of 2020, suggesting a potential positive correlation between market sentiment and stock market performance. From September 2020 to January 2021, the BI indicator exhibited increased volatility, aligning with the highs and lows of the SSE index. The sharp fluctuations of the BI indicator in early 2021 corresponded with the SSE index's downward trend, highlighting a possible link between sentiment changes and market volatility.

This correlation provides investors with a new perspective to predict market movements through sentiment analysis. Social media sentiment analysis can be used as an effective indicator of market sentiment to help investors better understand market dynamics.

Figure 5. Correlation of Investor Sentiment (BI Indicator) with Shanghai Composite Index (April 2020 - April 2021)

It is also worth noting that the BI indicator tends to lead market movements by a small amount of time, suggesting that sentiment analysis may have some predictive value for market behavior. This leading relationship was particularly evident at market turning points in late 2020 and early 2021.

5. Conclusion

This study empirically analyzes the relationship between investor sentiment and stock market performance using the Support Vector Machine (SVM) model, constructs a sentiment classification model based on Linear SVC, and verifies its effectiveness in financial market sentiment analysis. The results show that investor sentiment has a significant impact on the volatility of the SSE index and can improve the prediction ability of the market trend to a certain extent. Compared with traditional sentiment analysis methods, the machine learning model used in this study demonstrates higher classification accuracy and applicability, providing a new technical means to quantify investor sentiment.

The innovation of this study is that, on the one hand, the quantification of sentiment data is realized by combining machine learning techniques, which improves the automation level of sentiment analysis; on the other hand, the value of sentiment variables in the prediction of market performance is verified through empirical studies, which provides a more informative analytical framework for investors and policy makers.

The theoretical significance of this study is that it expands the application of sentiment analysis in financial market forecasting and provides a new perspective for understanding the intrinsic mechanism between investor sentiment and market performance. The practical significance is that this study provides an efficient and accurate sentiment quantification tool, which can help investors better understand market sentiment fluctuations and provide a reference basis for investment decisions.

Future research can further optimize the sentiment classification model, such as introducing deep learning methods to improve text feature extraction, or combining more dimensional data (e.g., news, social media, etc.) to construct more comprehensive sentiment indicators. Meanwhile, the predictive

validity of sentiment indicators under different market conditions and the method of combining sentiment analysis with traditional prediction models are explored in order to reveal the impact of investor sentiment on market dynamics in a more comprehensive way.

References

[1] Gao Yang, Shen Yiran, Xu Jiaxi. The Impact of Investor Sentiment on the Returns of the STAR Market: A Text Mining Perspective [J]. Operations Research and Management, 2022, 31(2): 184.

[2] Wang Ting, Yang Wenzhong. A Review of Research on Text Sentiment Analysis Methods [J]. Journal of Computer Engineering & Applications, 2021, 57(12).

[3] Lin Peiguang, Zhou Jiaqian, Wen Yulian. SCONV: A Method for Financial Market Trend Prediction Based on Sentiment Analysis. Journal of Computer Research and Development, 2020, 57(8): 1769-1778.

[4] Xu Xuechen, Tian Kan. A New Method for Stock Index Prediction Based on Financial Text Sentiment Analysis. Quantitative & Technical Economics Research, 2021(12): 9-22.

[5] Kumar K. S., Radha Mani A. S., Ananth Kumar T., Jalili A., Gheisari M., Malik Y., Chen H.-C., Moshayedi A. J. Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification. Applied Artificial Intelligence, 2024, 38(1): e2321555.

[6] Li Yang, Dong Hongbin. Text Sentiment Analysis Based on Feature Fusion of CNN and BiLSTM Networks. Computer Applications, 2018, 38(11): 3075-3080.

[7] Bu Hui, Xie Zheng, Li Jiahong, Wu Junjie. The Impact of Investor Sentiment Based on Stock Reviews on the Stock Market. Journal of Management Sciences in China, 2018, 21(4): 86-101.

[8] Rizinski M., Peshov H., Mishev K., Jovanovic M., Trajanov D. Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). IEEE Access, 2024, 12: 7170-7198.

[9] Peivandizadeh A., Hatami S., Nakhjavani A., Khoshcima L., Chalak Qazani M. R., Haleem M., Alizadehsani R. Stock Market Prediction With Transductive Long Short-Term Memory and Social Media Sentiment Analysis. IEEE Access, 2024, 12: 87110-87130.

[10] Xu Nan, Li Songsong, Hui Xiaofeng, et al. Research on fractal characteristics and risk measurement of China's stock market under the impact of COVID-19 pandemic [J]. Operations Research and Management Science, 2024, 33(1): 138.

[11] Sharifani K, Amini M. Machine learning and deep learning: A review of methods and applications[J]. World Information Technology and Engineering Journal, 2023, 10(07): 3897-3904.

[12] Abubakar H D, Umar M, Bakale M A. Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec[J]. SLU Journal of Science and Technology, 2022, 4(1): 27-33.