ISSN 2959-6157

A-share Trend Prediction Based on Machine Learning and Sentiment Analysis

Jiaming Zhang

Software Engineering, Shandong University of Science and Technology, Jinan, Shandong, 250000, China
E-mail: Jiaming.zhang@sdust.edu.cn

Abstract:

This study addresses the predictive challenges in China's A-share market, characterized by high retail investor participation and significant policy impacts. We introduce a novel predictive model that leverages both machine learning algorithms and sentiment analysis to forecast market trends. The research utilizes comprehensive datasets, including real-time A-share market data and sentiment-derived data from stock-related news, processed via advanced machine learning techniques like Random Forest and sentiment analysis tools. Our approach innovatively combines traditional technical indicators with sentiment scores to enhance the predictive accuracy of the model. The findings suggest that integrating sentiment analysis significantly improves the model's performance, evidenced by enhanced prediction metrics such as Mean Absolute Error (MAE) and R-squared values, which compare favorably before and after incorporating sentiment data. This study not only contributes to the existing financial prediction literature by providing a hybrid methodological approach but also offers practical implications for investors and policymakers in navigating the volatile A-share market.

Keywords: A-share market, machine learning, sentiment analysis, financial forecasting

1. Introduction

Predictive techniques in financial markets are crucial, especially in rapidly changing environments like China's A-share market, characterized by a high retail investor presence and significant policy impacts. This study adopts advanced machine learning and sentiment analysis to forecast market trends, leveraging large data sets and emotional insights from various media. By integrating these technologies, the research aims to offer precise, dynamic forecasting tools, crucial for investors and policymakers. This approach not only taps into technological innovation but also addresses unique market dynamics, providing valuable insights for investment strategies and economic stability.

2. Literature Review

Machine learning and sentiment analysis have significantly advanced stock market prediction. Enhanced computational capabilities and data access have allowed techniques like Support Vector Machine (SVM), Random Forests, and Deep Learning to excel in handling nonlinear issues and vast datasets, enhancing prediction accuracy. Studies by Huang et al. (2005) and Ding et al. (2015) have confirmed the superiority of these models over traditional statistical methods in capturing market dynamics.

Sentiment analysis has proven effective in financial forecasting by extracting emotional data from unstructured sources such as news and social media, influencing market sentiment. Research by Bollen et al. (2011) and Liew and Budavári (2016) demonstrated its capability to predict market movements based on sentiment data.

In China's A-share market, characterized by a high retail investor presence and frequent policy changes, machine learning and sentiment analysis have shown potential in improving forecasting accuracy. Studies by Li and Chen (2014) and Zhang et al. (2016) highlighted the benefits of integrating sentiment analysis, particularly from social media, to predict market fluctuations, emphasizing the relevance of these technologies in addressing the unique challenges of the A-share market.

3. Methodology

3.1 *Data acquisition*

In this study, data collection was divided into two main parts: real-time stock data from the A-share market and sentiment data related to stock market news.

3.1.1 *Acquiring A-share market data*

To obtain real-time stock data from the A-share market, we utilized the TuShare tool. TuShare is a free, open-source Python financial data interface package designed specifically for retrieving data from the Chinese stock market. It offers a comprehensive dataset including daily stock prices, trading volume, market capitalization, PE (price-earnings ratio), and various other financial indicators. These data are used to construct and train our machine learning models to predict stock prices and market trends.

Through TuShare, we can regularly and automatically acquire the latest data to ensure that the data used for model training and prediction is the most up-to-date and accurate.

3.1.2 *Sentiment data on stock market-related news and information*

To collect stock market-related news and analyze their sentiments, we utilize web crawling technology to scrape posts and comments from financial communities like stock forums. Guba, such as one of the largest stock investor communities in China, provide a wealth of investor opinions and sentiments, making it an ideal data source for sentiment analysis. By utilizing the Beautiful Soup and requests libraries in Python, we have designed a web crawler program to fetch the latest discussions and news comments on the A-share market from stock forums based on stock prediction dates. The retrieved text data will undergo sentiment analysis through Natural Language Processing (NLP) techniques to evaluate the potential impact of public sentiment on the stock market.

3.2 *Data Preprocessing*

3.2.1 *Preprocessing of Stock Market Data*

During the preprocessing stage of stock market data, several key steps have been taken to ensure data quality and consistency for effective subsequent analysis and machine learning modeling.

3.2.1.1 Missing Value Handling

In financial data analysis, it's crucial to maintain data integrity, especially with stock market data where missing entries are common due to non-trading days or recording errors. A common practice is to fill these gaps with the previous day's data, ensuring continuity and consistency in the time series. This method is straightforward, avoids external estimation errors, and aligns with standard financial management practices.

In order to have a better understanding of the impact of filling the data with the previous day's data, the following chart displays the changes in data before and after processing.

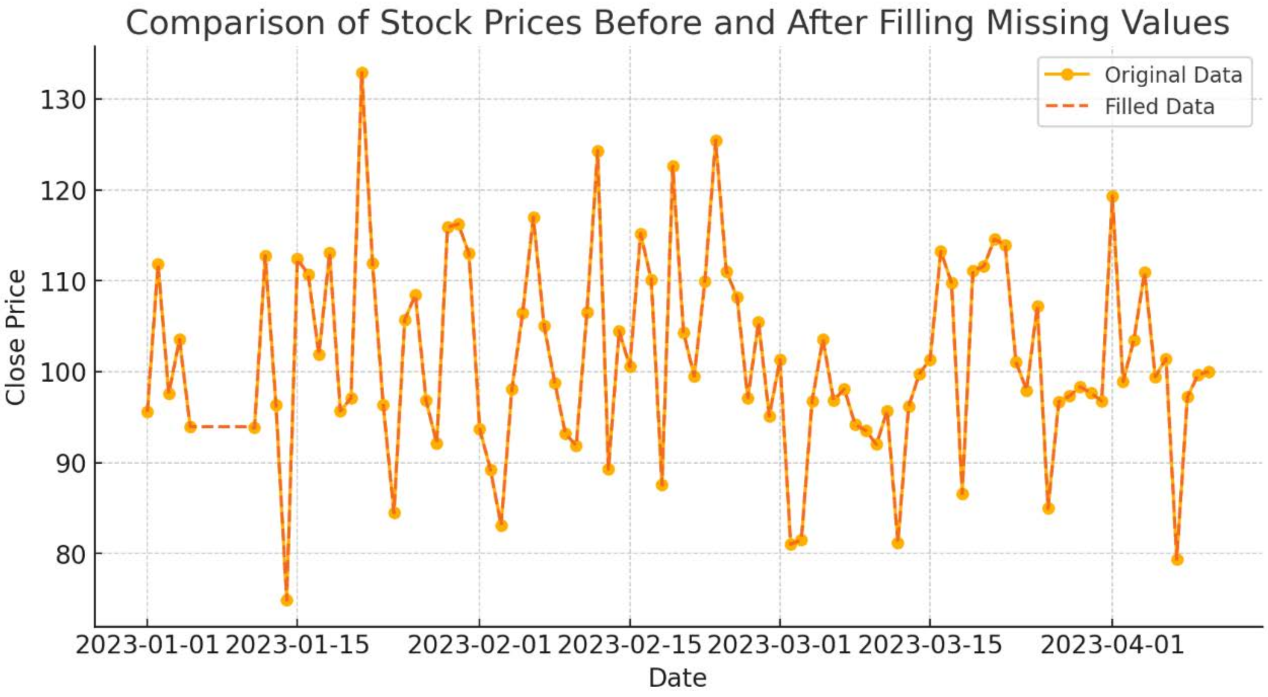

Figure 1 Comparison of Stock Prices Before and After Filling Missing Values

3.2.1.2 Handling Outliers

Detection Methods: Outliers are typically identified using a box plot. The box plot displays the data distribution through quartiles and interquartile range (IQR), with outliers commonly defined as values below Q1-1.5IQR or above Q3+1. 5IQR.Handling Strategy: For detected outliers, one can choose to either delete these data points or perform interpolation using nearby values.

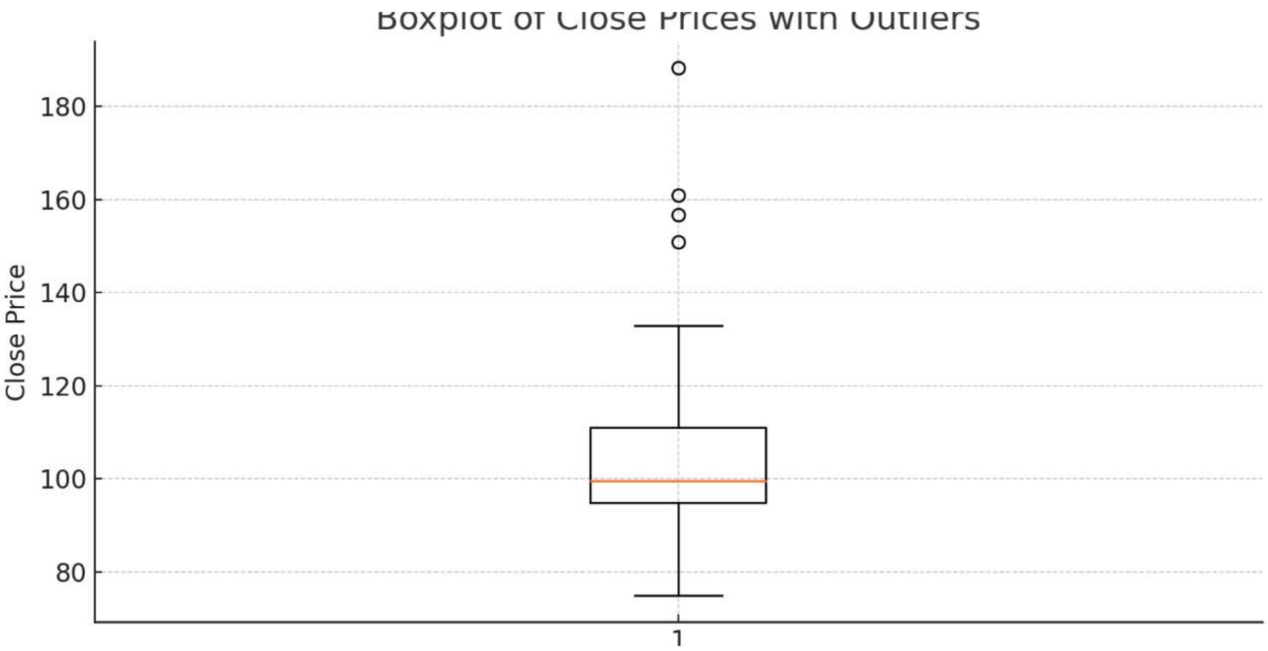

Figure 2 Boxplot of Close Prices with Outliers

In Figure 3.2, a box plot of stock closing prices can be observed, which includes some outliers. These outliers are represented as points outside the box plot, significantly deviating from the main data distribution.

In this example, I intentionally increased the closing prices of some data points by 1.5 times the normal value to simulate abnormal situations.

3.2.1.3 Data normalization

The commonly used normalization methods include Min-Max normalization and Z-Score standardization. However, in data processing in the stock market, since price data may have very large range differences, using Min-Max normalization is a more common choice because it maintains the relative relationships between all original data and is easier to interpret.

The formula for Min-Max is:, where and are respectively the maximum and minimum values in the data.

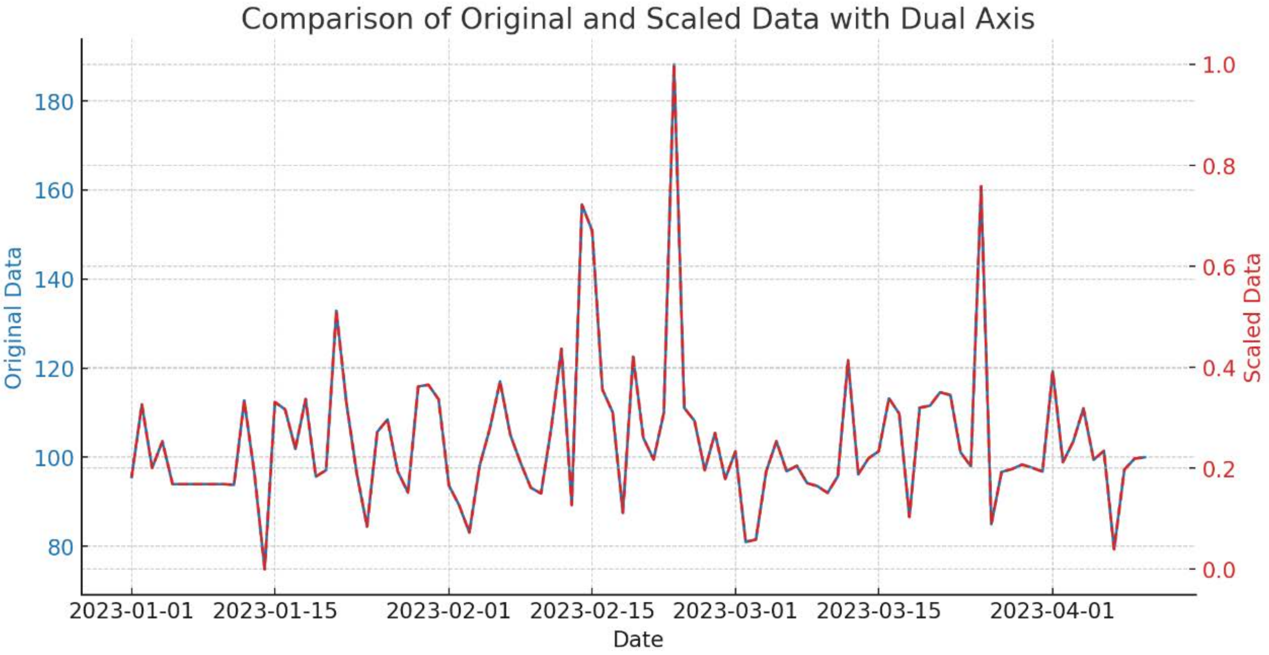

Figure 3 Comparison of Original and Scaled Data with Dual Axis

Through Figure 3.3, it can be clearly seen that, despite the different scales of the data, normalization preserves the patterns and trends of the original data. This representation helps to validate the correctness of data processing, while ensuring consistency and effectiveness in model processing and comparison.

3.2.2 *Emotion Data Preprocessing*

Prior to conducting sentiment analysis, preprocessing of text data is an essential step.

lText Cleaning: Remove irrelevant characters, punctuation, HTML tags, etc., and keep only meaningful text content.

lWord Segmentation: Chinese text needs to be segmented. Although the BERT model uses the WordPiece tokenization method, preliminary word segmentation during preprocessing can help understand the data clearly.

lConvert to BERT Input Format: Text needs to be converted into a format that BERT can understand, including:

Tokenization: Splitting the text into tokens. Add special tokens: Add special tokens at the beginning and end of the sentence ([CLS], [SEP]).

Pad or truncate: Normalize the text length to a consistent length (e.g., 50 tokens).

Attention mask: Create a mask to differentiate between real data and padded data.

3.2.3 *Model Selection*

In this study, considering the particularity of Chinese text processing and the requirements of sentiment analysis tasks, we selected BERT (Bidirectional Encoder Representations from Transformers) model as the core algorithm. BERT is a pre-trained language representation model that has been proven to perform excellently on various natural language processing tasks, including sentiment analysis.

3.2.3.1 Why choose BERT?

BERT stands out for its bidirectional architecture that accurately interprets words in varied contexts, essential for sentiment analysis. Pre-trained on vast text data, it efficiently adapts to new tasks with minimal training, automatically extracting relevant features. Supported by major frameworks like TensorFlow and PyTorch, BERT benefits from extensive community resources, which streamline development. Its foundation on the Transformer model, with layers of self-attention mechanisms, enables comprehensive processing of text for precise sentiment categorization.

BERT utilizes the Transformer architecture, featuring multiple layers of encoders with self-attention mechanisms and fully connected layers to handle sequential data comprehensively. For sentiment analysis, it converts text into a sequence of tokens, each embedded into vectors. The special [CLS] token begins the sequence, with its final state representing the entire sequence for classification. BERT then processes these vectors to predict sentiment categories like positive or negative using an additional output layer.

3.2.3.2 Parameter Settings

To achieve the optimal model performance, I adjusted the following parameters based on the task characteristics and data set size:

lLoss Function: Employing the CrossEntropyLoss, which is suitable for multi-class classification problems.

lOptimizer: Implementing the AdamW optimizer, known for its effectiveness in training deep learning models.

lLearning Rate: Fixed at 0.00001 to ensure stable convergence during the learning process.

lBatch Size: set to 32, balancing the efficiency of memory consumption and model updates.

lEpochs: set to 8 rounds, sufficient for the model to converge on the training data.

lMax Length: the maximum length of the input sequence is set to 50 tokens, considering the average length of Chinese text and computational efficiency.

3.2.3.3 Model training and validation

Training Process:

Precision (blue line): The initial precision exceeded 0.72, demonstrating good initial performance. The precision slightly decreased afterwards but stabilized above 0.72 after multiple epochs.

Recall (red line): The recall rate showed significant fluctuations, indicating the model’s performance in recognizing all relevant examples is not stable, with an overall trend of gradual decline.

F1 score (green line): The F1 score exhibited a downward trend overall, suggesting a trade-off between precision and recall.

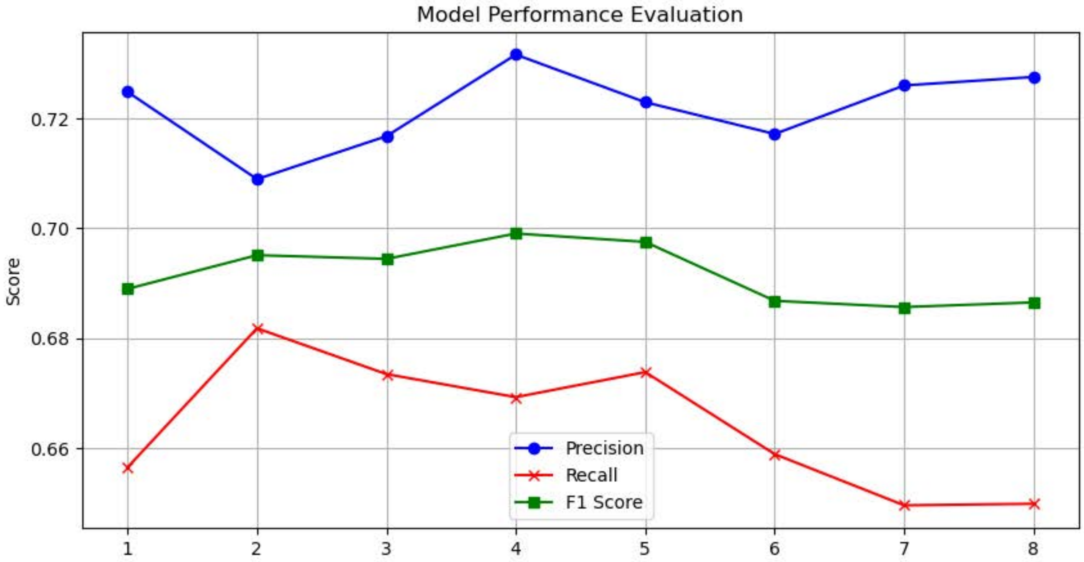

Figure 4 Model Performance Evaluation

Summary: Although the accuracy is gradually improving,

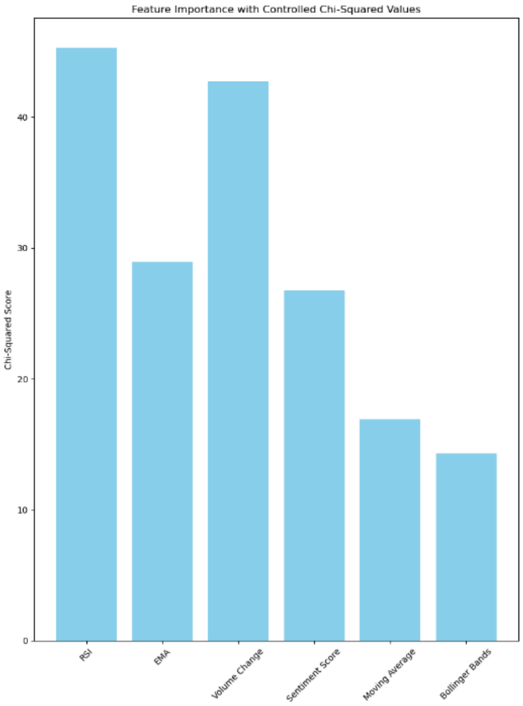

Figure 5

Figure 3.5 displays the preliminary results of the chi-square test. The chart shows the chi-square values of six features, including Relative Strength Index (RSI), Exponential Moving Average (EMA), Volume Change, Sentiment Score, Moving Average, and Bollinger Bands. From the graph, it is evident that the chi-square values of ESI and Volume change are the highest, indicating their strongest association with the target variable (stock price trend). This may be because they can effectively reflect market momentum, as volume changes often herald shifts in market trends. While sentiment scores rank only fourth, considering that sentiment scores can provide a different market perspective from traditional technical indicators, including sentiment scores as supplementary features may enhance the predictive accuracy and robustness of the model.

In conclusion, the selected features are: technical indicators such as RSI and EMA, trading volume changes, and sentiment scores.

3.2.4.2 Model validation

During the experiment, I used 500 data points as the base training set, with a 20% test set and an 80% training set. The data were standardized and normalized to ensure stability. Following the principles of trial and error as well as experience-based parameter tuning, I experimented with using a step size of 1, 3, and 10 trading days, with multiple features per trading day as input, and the closing price of one trading day as the output for training. Ultimately, it was found that using a 3-day trading day step size with trading data and sentiment as features, in a random forest model with the following parameter settings: 100 trees, maximum depth of 10, feature sampling equal to the square root of the total number of features, Bootstrap sample set to True, resulted in a good fit for predicting the stock price of the next day.

In order to evaluate the performance of the model, I used 5-fold cross-validation, and the following are the cross-validation results.

Table 3.1

Folds number	1	2	3	4	5	average
Random forest accuracy	72.5%	70.8%	73.0%	71.5%	74.8%	72.2%

Save the trained model and perform model saving every 50 trading days. The total test interval length is set to the last 150 days of each stock dataset. Generate a comparison chart of the predicted trends before and after adding emotional scores as input features, as shown in Figure 3.6, Figure 3.7.

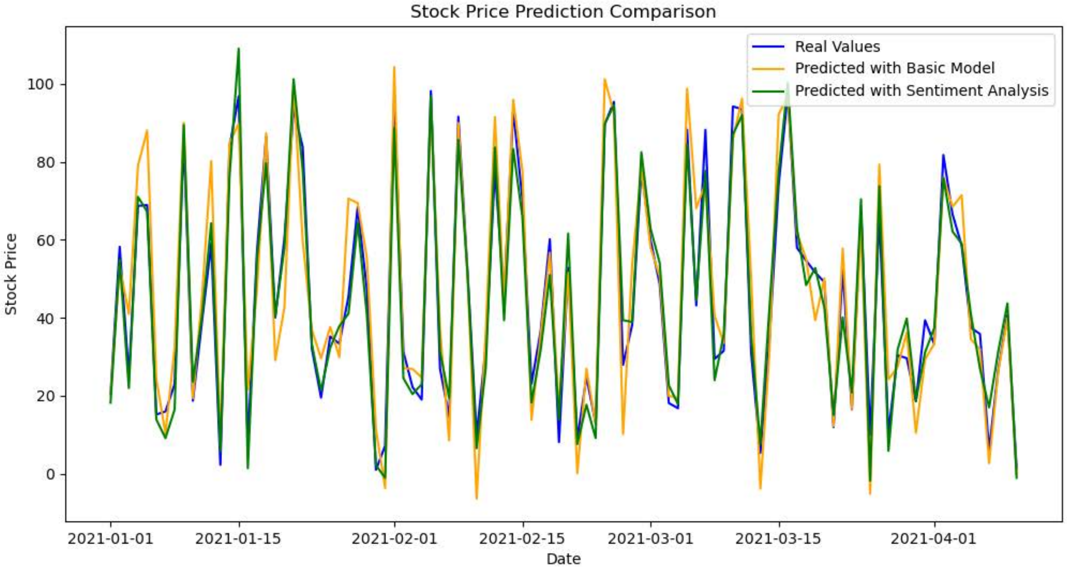

Figure 3.6

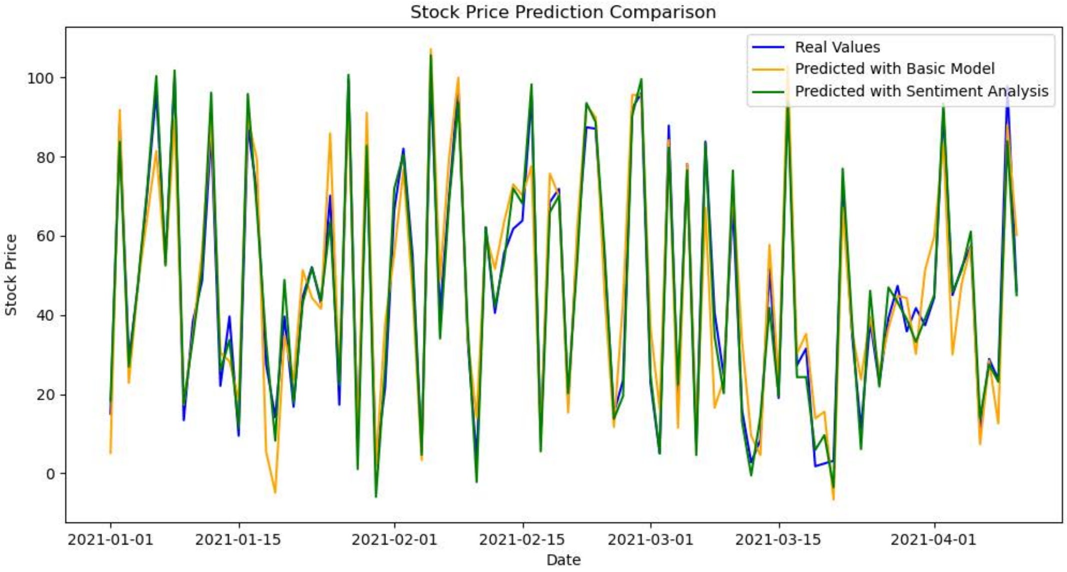

Figure 3.7

In order to further evaluate the performance of the model in predicting different stocks, four indicators were selected as criteria for judging the experimental results, namely mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination () as model evaluation standards. Training was conducted using a rolling model, recording and calculating the average evaluation indicators for stocks in the test set.

Table 3.2

test collections1	MAE	MSE	MAPE	
Excluded emotional indicators	0,686	1.642	0.010	0.827
included emotional indicators	0.612	1.401	0.015	0.835
test collections2	MAE	MSE	MAPE	
Excluded emotional indicators	0.0439	0.105	0.023	0.818
included emotional indicators	0.008	0.032	0.015	0.827

When the values of MAE, MSE, and MAPE are smaller, and R2 is closer to 1, it indicates that the model has a better fit. As can be seen from Table 3.2, the trained model achieved an average goodness of fit R2 of over 80% on the test set, showing a good fitting effect; the overall fitting degree of the model improved after adding the emotion score index, with a decrease in the loss value, indicating that the inclusion of the emotion score index has a certain impact on the stock prediction model.

4. Conclusion

This paper delves into the prediction of A-share market trends using machine learning and sentiment analysis, enhancing predictive capabilities by combining traditional technical analysis with advanced algorithms. It details the entire workflow from data collection to model deployment, covering data preprocessing, feature engineering, model training, validation, and the final deployment and performance monitoring.

The research highlights the crucial role of data preprocessing in addressing missing values and outliers in financial time series. Feature engineering is enriched by incorporating technical indicators like RSI, EMA, and sentiment scores, creating a robust feature set that captures market trends and emotions. The effectiveness of random forests and gradient boosting trees in managing complex non-linear relationships is confirmed, with performance metrics such as MSE, MAE, MAPE, and R^2 demonstrating the models' accuracy and generalization capability. These models provide reliable predictions that can guide investor decisions in the A-share market.

The study concludes with the successful practical application of the model, bridging theory and practice for operational use in trading systems. Looking forward, the paper suggests directions for future research, including addressing forecasting disparities across different sectors, integrating deep learning models like CNN and LSTM for improved predictions, expanding data sources to enhance sentiment analysis, developing algorithms for real-time parameter adjustments, and incorporating sophisticated risk assessment mechanisms to maintain stability during market fluctuations.

5. Reference

1)Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259-1294.

2)Yong, W., Multi-mode Matching Identification Algorithm for Short-term Stock Investment Forecasting[J]. Computer Applications, 2014, 34(S2):180-183.

3)Xia, G., Research on Financial Market Trend Prediction based on Machine Learning Algorithms[J]. Microcomputer Application, 2023, 39(02): 30-32+40.

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

5)Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), 1375-1388.

6)Guang, Liu., Research and Application of Stock Market Prediction Model and Evaluation Method Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2020.

7)Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press, 2327-2333.

8)Engelberg, J., Sasseville, C., & Williams, J. (2012).

Market Madness? The Case of Mad Money. Management Science, 58(2), 351-364.

9)Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.

10)Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems, 55(3), 685-697.

11)Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2018). NSE Stock Market Prediction Using Deep-Learning Models. Procedia Computer Science, 132, 1351-1362.

12)Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702.

13)Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14-23.

14)Luss, R., & d'Aspremont, A. (2015). Predicting Abnormal Returns From News Using Text Classification. Quantitative Finance, 15(6), 999-1012.

15)Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611.

16)Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.

17)Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems, 27(2), 12.

18)Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168.

19)Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 1310-1319.

20)O'Hare, N., Lee, H., Cooray, S., Gruhl, D., & Nelson, D. (2009). Detection and visualization of stock mentions on Twitter. In IUI Workshops.

21)Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based Twitter sentiment for stock prediction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 24-29.

22)Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 375-378.