The relationship between stockholder sentiment lag and stock price prediction accuracy: an empirical analysis based on LSTM and Transformer models

Haoqian Guo[1,*], Yuxin Xu[2]

[1]College of computer science, Sichuan University, Chengdu, China, 610207

[2]School of economics, Sichuan University, Chengdu, China, 610065

*Corresponding author: guohq629@gmail.com

Abstract. This comprehensive paper investigates the nuanced relationship between retail investor sentiment and stock prices in the Chinese stock market, with a special focus on the role of sentiment time lags. Using advanced time-series models, specifically Long Short-Term Memory (LSTM) and Transformer models, the study takes a detailed look at the stock price of Oriental Finance (Ticker: 300059A). The research employs varying time lags of stockholder sentiment (ranging from 0 to 4 days) as well as technical indicators to predict stock prices. Our experimental design involves comparative analysis under these two models to isolate the impact of sentiment time lags on prediction accuracy. The results reveal that the LSTM model consistently outperforms the Transformer model, particularly when a 4-day lag in stockholder sentiment is considered. Interestingly, the prediction accuracy did not uniformly improve with increased sentiment lags, suggesting a complex relationship between investor sentiment and stock prices.

Keywords: Stockholder Sentiment, Time Series Modeling, LSTM, Transformer, Stock Prediction.

1. Introduction

Investor sentiment is a subject of considerable academic interest, particularly its impact on financial markets. The China stock market provides a unique landscape for such a study, given its volatility and the significant influence exerted by retail investors. These investors are notorious for their emotionally driven decisions, which often lead to short-term market trends [1]. While previous studies have postulated behavioral theories like the herding effect and buy-the-winner behavior to explain such impact [2], there is a notable gap in exploring how the time-lagged sentiment influences stock prices.

To address this gap, this paper focuses on the effects of varying time lags (from 0 to 4 days) in investor sentiment on stock price prediction. We utilize advanced time-series models, specifically Long Short-Term Memory (LSTM) and Transformer algorithms, to predict the stock price of Oriental Finance (Ticker: 300059A). Our feature set includes not only technical indicators commonly used in stock trading but also time-lagged investor sentiments.

Our study aims to achieve two major objectives: First, to identify which among the LSTM and Transformer models delivers more accurate stock price predictions; and second, to ascertain the time lag of investor sentiment that yields optimal predictive results. Through this multi-faceted analysis, the research seeks to offer actionable insights for both market participants and policy makers, aiding in more informed investment decisions and policy planning for market stability.

2. Related Work

2.1. Previous modeling studies

Various machine learning algorithms have been applied to stock market prediction studies with some success. However, their shortcomings are also relatively obvious, and their prediction effect on the stock system is limited. With the rapid development of the big data era, deep learning technology has made breakthrough progress with its more powerful learning ability and feature extraction ability, and is widely used in stock prediction research by virtue of its ability to fit complex time series data.

FinBERT was introduced to solve the natural language processing problem in the financial field, and it was concluded that FinBERT was superior to the state-of-the-art machine learning methods [3]. Neural networks and BP algorithms were applied to classify and predict stock prices, where five consecutive days of trading data were used as input samples and the next day's values were used as output samples. After comparing the accuracy with deep learning fuzzy prediction, it was concluded that BP algorithm is more accurate and can better predict the stock price trend [4]. The MLP neural network was compared with the traditional BP neural network method by constructing the MLP neural network in the TensorFlow environment. The comparison results show that the MLP neural network has a stronger prediction ability and takes less time to predict, as well as more excellent prediction results [5]. Since previous prediction systems tend to ignore the importance of key metrics and feature engineering, hybrid GA-XGBoost prediction system was utilized for prediction tasks with enhanced feature engineering process including feature set expansion, data preparation and optimal feature set selection for hybrid GA-XGBoost algorithm [6].

RNNs (Recurrent Neural Networks) are frequently used to analyze predictive sequential data, such as the use of RNNs to predict Apple's stock price over a ten-year period and proved to have a prediction success rate of over 95% [7]. However, studies have shown that over time RNN will forget the previous state information, the general RNN model has a long memory time series data portrayal ability is weak, LSTM model in the RNN structure based on the transformation, LSTM network in addition to the application of image analysis, document summarization, speech recognition, handwriting recognition, etc., in the time series data prediction also has a good performance. It is mainly used to portray the relationship between the current data and the previous input data, and utilizes its memory ability to save the previous state information of the input network, thus solving the problem that the RNN model is unable to portray the long memory of the time series.LSTM was compared with eighteen classical models, and verified that the prediction of LSTM is the best and the prediction effect is gradually stable [8].

The Transformer model adopts the encoding-decoding framework, utilizes the self-attention mechanism to achieve more efficient feature extraction of time series, captures the dependencies between elements in long sequences, and is capable of parallel training with strong feature extraction capability, fusion of multimodal capabilities, and powerful It can realize parallel training and has powerful feature extraction, multimodal fusion and interpretation capabilities. Due to the great success of Transformer's approach in time series tasks, it is gradually being widely used in many fields such as graphical image recognition, natural language processing, and economic securities forecasting [9].

Transformer is advantageous in processing time series data such as stock market data and has been utilized in predicting future stock prices of Dhaka Stock Exchange (DSE) to support short and long term investment decisions. For the original coding position encoding of the input data, the Time2Vec scheme was introduced for the task of coding time series data as the data is time series, i.e., the predicted feature vectors depend on the feature vectors seen in the past [10].

2.2. Sentiment analysis

Investor sentiment has been shown to have a strong relationship with stock prices and has been used in stock price prediction models. A composite indicator of investor sentiment was used to verify that investor sentiment can predict market trends and that individual investor sentiment dominates market trends when the market is in an upward trend [11]. A hybrid model was proposed which combines LSTM and sentiment analysis models to predict stock prices in the Chinese A-share market. The hybrid model consists of three parts: data preprocessing, sentiment analysis and predictive model construction. Stock prices are predicted by analyzing the information more comprehensively from the aspect of time series analysis and from the aspect of investor sentiment [12].

Almost all of the literature agrees that investor sentiment has an influence on stock market trends, and the only disagreement is the positive or negative nature of the influence, which is not able to form a unified consensus in the stock markets of multiple countries. And no matter what kind of research is conducted on the impact of investor sentiment, how to measure investor sentiment is undoubtedly one of the most important prerequisites [13].

2.3. Our works

Based on the previous research on the impact of stockholders' sentiment on the stock market and the application of time-series models in stock price prediction, because stockholders' sentiment is forward-looking for the stock market in the short term [14], i.e., the stock market is lagging for stockholders' sentiment, this paper decides to investigate the impact of stockholders' sentiment lag on stock price prediction in this paper and to conduct a validation exercise on the previously validated LSTM model and the Transformer model. In particular, in the work of sentiment analysis of stockholders' comment texts, this paper uses Finbert model for the task of sentiment analysis, and for the output layer of FInbert, this paper uses manually categorized stockholders' comment texts for training.

Because the stockholder sentiment in the short term, relative to the market has a certain degree of forward-looking, this paper will be lagged operation of the stockholder sentiment data. The forward-looking nature of stockholder sentiment is mainly reflected in the short term, so the lag period chosen in this paper is 0-4 days.

3. Data

3.1. Data Collection

This article collects trading data for a single stock 300059 Oriental Fortune from January 4, 2018 to March 31, 2023, where the data is from the RESSET database, where the technical indicators include: date, opening price, high, low, closing price, trading volume.

This paper also collects stock bar comments for a single stock, 300059 Oriental Fortune, from January 4, 2018 to March 31, 2023, where the source of the data for the stock bar comments is the Oriental Fortune's stock bar forum, which is an online forum used by stockholders to discuss stock trading. The collected stock bar comments include: comment text and comment posting time.

3.2. Data Preparation

3.2.1 Data Cleaning

For the collected stock bar comment texts, this paper performs data cleaning, which is mainly used to clean some of the news texts as well as meaningless texts. In order to clean the text, this paper uses the regular matching technique to capture the features in the text that need to be cleaned and remove them.

3.2.2 Sentiment Analysis

For the stockholder comments that have been data cleaned, this paper uses Finbert for the sentiment analysis task.Finbert is a bert-based NLP model that outperforms the original bert in the prediction task after training using a small amount of text [15]. For the output layers, this paper uses a manually classfied positive sentiment text set and a negative sentiment text set for retraining, and the training texts are not involved in subsequent prediction. The data size of both text sets is around 4000 items. After retraining, Finbert is used as a binary classification model to discriminate text sentiment as either 1 (positive sentiment) or -1 (negative sentiment) when performing sentiment analysis tasks. After the sentiment discrimination is done for each text, the sentiment values of each text for each day are summed and averaged to get the stockholder sentiment score for each day.

3.2.3 Final Dataset

In the final dataset, the features are the opening price, high price, low price, closing price, trading volume, and stockholder sentiment value of the day for each trading day, and Label is the closing price of the second day. After normalizing the data, the data is rearranged into time series data . In this paper, the size of the time window chosen is 5, i.e., 5 days of data are used each time to predict the closing price on the 6th day. For the finishing time series data, this paper takes into account the problem of data size and chooses to use the first 90% as the training set, and the last 10% as the validation set.

4. Model Design

4.1. Transformer-based Modle

Our model comprises four key components: the Time2Vec layer, Fusion layer, Transformer encoder, and a feedforward neural network layer. The Time2Vec layer processes input X of shape (batch size × feature size) and embeds each input feature periodically. This output is then merged with the raw inputs in the Fusion layer and resized to the original feature dimensions. The tensor produced feeds into a 3-layer Transformer encoder with 8 attention heads, featuring a dropout rate of 0.1. The encoder's output undergoes further refinement in the feedforward layers to produce a single scalar value as the final result. Figure 1 visually presents the model, using B, S, F to represent batch_size, sequence_length, and feature_size, respectively.

Figure 1 shows the architecture of the Transformer-based model, including the shape of the inputs.

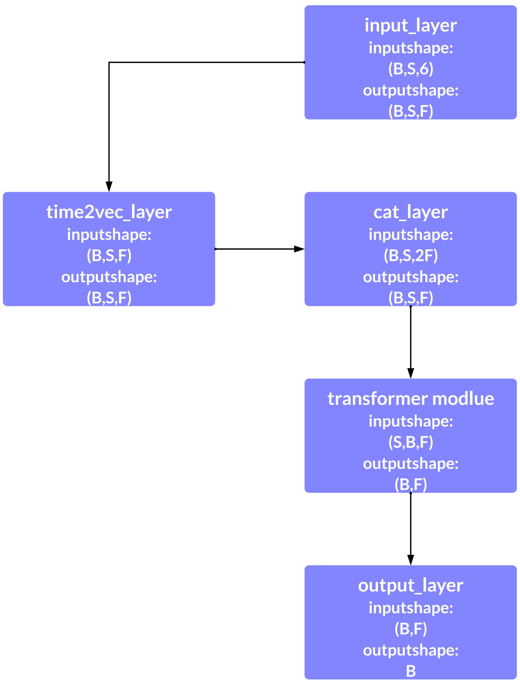

Figure 1. Transformer-based model illustration

4.2. LSTM-based Modle

The final output of the model is generated by a fully-connected layer that uses a "tanh" activation function and outputs a scalar value as the final prediction. Our model architecture comprises five essential layers: a one-dimensional convolutional layer (Conv1D), a maximum pooling layer (MaxPooling1D), a long short-term memory (LSTM) layer, a Dropout layer, and an Attention Mechanism layer. Initially, the Conv1D layer accepts inputs, applying a sigmoid activation function.

It has a kernel size of 1 and performs local feature extraction. The subsequent MaxPooling1D layer then reduces dimensionality by selecting significant features from the convolutional layer's output.

An LSTM layer follows, capturing long-term dependencies in the sequence with a "tanh" activation function. To mitigate overfitting, we include a Dropout layer that sets a portion of the LSTM output to zero during training as a regularization measure. Lastly, an Attention Mechanism layer calculates attention weights using a fully-connected layer with a "sigmoid" activation function. These weights modify the LSTM output to emphasize different time steps. The final output is generated through a fully-connected layer that employs a "tanh" activation function, producing a scalar value as the prediction.

Figure 2 shows the architecture of the LSTM-based model, including the shape of the inputs.

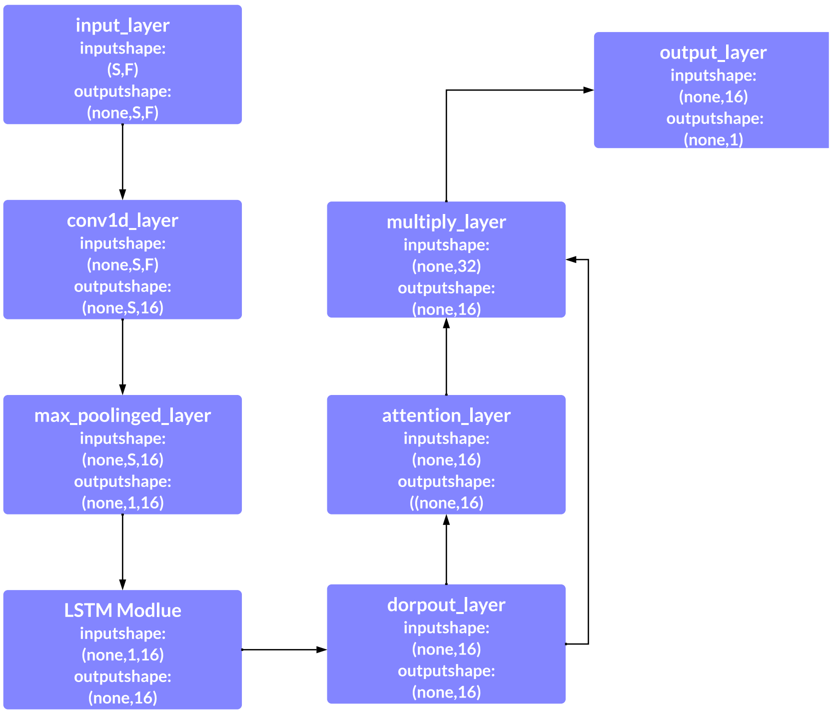

Figure 2. LSTM-based model illustration

5. Result

5.1. Metric

5.1.1 Mean Squared Error(MSE)

Mean square error is a statistical measure of the difference between the predicted value and the true value and is often used to measure the fit of a predictive model. It calculates the average of the squares of the differences between the predicted and true values.

Mathematical expression:

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{Y_i} - \widehat{\mathbf{Y}}_i\right), \tag{1}$$

The smaller the MSE, the smaller the difference between the predicted value and the true value, and the better the fit of the model.

5.1.2 Root Mean Squared Error(RMSE)

The root mean square error is the square root of the mean square error, which measures the average difference between the predicted value and the true value while retaining the same units as the original data.

Mathematical expressions:

$$\mathbf{RSME} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}, \quad (2)$$

The smaller the value of RMSE, the smaller the average difference between the predicted and true values.

5.1.3 Mean Absolute(MAE)

The mean absolute error measures the average of the absolute differences between the predicted and true values, and it is less sensitive to outliers because no squaring operation is performed.

Mathematical expressions:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|, \quad (3)$$

The smaller the MAE, the smaller the average absolute difference between the predicted and true values.

5.2. The Influence of Stockholder Sentiment Lag on Stock Price Prediction Accuracy

As for the final experimental results, two variables are set in this paper, one is the lag days of the sentiment value (1-4 days), and the other is the model used (lstm or transformer).This paper hopes to use the method of comparative experiments to analyze the impact of the lag days of the stockholder's sentiment on the accuracy of the final prediction results. Simultaneous experiments using two different architectural models are used to exclude incidental factors and ensure the authenticity of the experimental results.

The following shows the prediction results of the Transformer model and the Lstm model using data from stockholder sentiment lag 0-4, where the evaluation metrics are MSE, MAE, and RMSE.

Table 1. Predictions under two models using stockholder sentiment without lag data

Model/ Metrics	transformer	lstm
MSE	0.0003690	0.0003458
RMSE	0.0192100	0.0185900
MAE	0.0137500	0.0158900

In Table 1, the prediction results of the Transformer and Lstm models are shown under the condition of using stockholder sentiment without lagged data. In this case, the MSE of the Transformer model is 0.0003690, which is inferior to the MSE score of Lstm of 0.0003458, but the MAE of the Transformer model is 0.0137500, which is superior to the MAE score of Lstm of 0.158900.

Table 2. Predictions under two models using stockholder sentiment lagged 1 period data

Model/ Metrics	transformer	lstm
MSE	0.0003902	0.0002453
RMSE	0.0197500	0.0156600
MAE	0.0150100	0.0128800

In Table 2, the prediction results of the Transformer and Lstm models are shown under the condition of using 1-period lagged data on stockholder sentiment. In this case, the Transformer model has an MSE of 0.0003902, which is inferior to the MSE score of 0.0002453 for Lstm and inferior to the prediction results when using no lag data, but the LSTM model has an MSE of 0.0002453, which is superior to the prediction results when using no lag.

Table 3. Predictions under two models using stockholder sentiment lagged 2 period data

Model/ Metrics	transformer	lstm
MSE	0.0003448	0.0003524
RMSE	0.0185700	0.0187700
MAE	0.0135400	0.0153400

In Table 3, the prediction results of the Transformer model and the Lstm model are shown under the condition of using lagged 2-period data on stockholder sentiment. In this case, the MSE of the Transformer model is 0.0003448, which is slightly better than the MSE score of the Lstm model of 0.0003524. The accuracy of the prediction results of the Lstm model in the condition of using lagged 2-period data is inferior to the accuracy of the prediction results when using lagged 0-period data. The MSE score of the Lstm model when using lag 0 data is 0.0003458.

Table 4. Predictions under two models using stockholder sentiment lagged 3 period data

Model/ Metrics	transformer	lstm
MSE	0.0003477	0.0001712
RMSE	0.0186500	0.0130800
MAE	0.0134400	0.0102800

In Table 4, the prediction results of the Transformer model and the Lstm model are shown under the condition of using stockholder sentiment lagged by 3 periods of data. In this case, the MSE of the Transformer model is 0.0003477, which is also inferior to the MSE score of the Lstm model of 0.0001712.The accuracy of the prediction results of both the Transformer model and the Lstm model is better than the accuracy of the prediction results of the condition of using stockholders' sentiment with a lag of 3 periods than that of the condition of using stockholders' sentiment without a lag.

Table 5. Predictions under two models using stockholder sentiment lagged 4 period data

Model/ Metrics	transformer	lstm
MSE	0.0002909	0.0001644
RMSE	0.0170600	0.0128200
MAE	0.0119700	0.0101000

In Table 5, the prediction results of the Transformer model and the Lstm model are shown under the condition of using 4-period lagged data on stockholder sentiment. Among them, the Transformer model has an MSE of 0.0002909 and the Lstm model has an MSE of 0.0001644, achieving the best performance among the 5 sets of experiments. Meanwhile, the Lstm model outperforms the Transformer model with a 4-period lag in stockholder sentiment.

6. Conclusion

In this paper, the Transformer and Lstm models were used to conduct a prediction task on the stock price of 300059A Oriental Finance using stockholder data and stock technical indicators with lags of 0-4 periods as features. From the prediction results, the Lstm model's prediction results are not as good as the Transformer model's prediction results with the use of 2-day lagged stockholder sentiment, but the Lstm model outperforms the Transformer model in general, and both models obtain the best prediction results with the use of 4-periods of lagged data on stockholder sentiment.

In conducting experiments with lags of sentiment values, this paper found an interesting phenomenon that the accuracy of the prediction results did not increase with the increase of lags, but rather the results were not as good as the results obtained from the prediction task using data without lags in certain lags of performance.

In subsequent studies, the focus could be on using more general and accurate metrics to evaluate stockholder sentiment. In this paper, the stockholder sentiment is obtained by analyzing the comment titles of the Oriental Fortune website's stock bar, while it is obvious that not all stockholders communicate on Oriental Fortune, the amount of data in this case is actually insufficient. In the subsequent research, it can be considered to use the stockholders' sentiment obtained by analyzing the macro method instead of the stockholders' sentiment obtained by using the discussion text of the stock bar.

References

[1] Dong H, Gil-Bazo J. Sentiment stocks[J]. International Review of Financial Analysis, 2020, 72: 101573.

[2] Yao C Z, Li H Y. Time-varying lead–lag structure between investor sentiment and stock market[J]. The North American Journal of Economics and Finance, 2020, 52: 101148.

[3] Araci D. Finbert: Financial sentiment analysis with pre-trained language models[J]. arXiv preprint arXiv:1908.10063, 2019.

[4] Zhang D, Lou S. The application research of neural network and BP algorithm in stock price pattern classification and prediction[J]. Future Generation Computer Systems, 2021, 115: 872-879.

[5] Han Shanjie, Tan Shizhe. Design and implementation of deep learning model for stock prediction based on TensorFlow[J]. Computer Application and Software, 2018, 35(6):6

[6] Yun K K, Yoon S W, Won D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process[J]. Expert Systems with Applications, 2021, 186: 115716.

[7] Zhu Y. Stock price prediction using the RNN model[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1650(3): 032103.

[8] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM[J]. International Journal of Machine Learning and Cybernetics, 2020, 11: 1307-1317.

[9] Muhammad T, Aftab A B, Ibrahim M, et al. Transformer-based deep learning model for stock price prediction: A case study on Bangladesh stock market[J]. International Journal of Computational Intelligence and Applications, 2023: 2350013.

[10] Muhammad T, Aftab A B, Ibrahim M, et al. Transformer-based deep learning model for stock price prediction: A case study on Bangladesh stock market[J]. International Journal of Computational Intelligence and Applications, 2023: 2350013.

[11] Hu J, Sui Y, Ma F. The measurement method of investor sentiment and its relationship with stock market[J]. Computational Intelligence and Neuroscience, 2021, 2021: 1-11.

[12] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM[J]. International Journal of Machine Learning and Cybernetics, 2020, 11: 1307-1317.

[13] Hu J, Sui Y, Ma F. The measurement method of investor sentiment and its relationship with stock market[J]. Computational Intelligence and Neuroscience, 2021, 2021: 1-11.

[14] Yao C Z, Li H Y. Time-varying lead–lag structure between investor sentiment and stock market[J]. The North American Journal of Economics and Finance, 2020, 52: 101148.

[15] Araci D. Finbert: Financial sentiment analysis with pre-trained language models[J]. arXiv preprint arXiv:1908.10063, 2019.