The Trends Prediction of GEM Stock Prices Based on Time Series Analysis and Weibo Sentiment Analysis

Author: Yiran Su (20010607-T400)

Supervisor: Farrukh Javed

Essay course code: DABN01

CONTENTS

1. Introduction

1.1 Background

The study of stock market has always been a popular topic among researchers. Stock markets are naturally noisy, non-parametric, non-linear, and deterministic chaotic systems (Ahangar, Yahyazadehfar,and Pournaghshband, 2010). However, due to the economic efficiency, it has received sustained attention for hundreds of years.

Previous studies have shown that the return and volatility of stock prices is affected by many factors, such as economic, financial, political, environmental, and health crises. Fama and French (1992) predict the US stock market with internal finance variables such as book-to-market, cash-flow-yields and size. Some studies focus on macroeconomic factors that may affect stock price. Among all the potential influencing factors, investor sentiment is a very important area of study. Baker and Wurgler (2006) proposed to measure and quantify the investor sentiment. Tetlok (2007) thinks the sentiment of investors also plays a part in stock pricing and construct a measure of media content that appears to correspond to either negative investor sentiment or risk aversion.

In this thesis, the study object would be growth enterprises of China's Growth Enterprise Market (GEM). Compared to large enterprises, growth enterprises have weaker risk resistance and lower financing needs. In order to meet the financing demands of them, especially those in the technology sector, China has established the Growth Enterprise Market to provide a more convenient financing channel, like a NASDAQ-style board. While there is already a considerable amount of research using sentiment analysis to predict stock prices, most studies focus on large enterprises. In contrast, there is much less research targeting entrepreneurial or growth enterprises. Gui, Pu, Naktnasukanjn, Yu, Mu, Pan (2022) used ERINE model and BERT model to analyze sentiment index and its impact on SZSE 100 Index with sentiment index proposed by Antweiler and Frank (2004). However, the study of growth market is still limited. This thesis will narrow the gap take a deeper look at GEM's market performance.

Some studies have shown that the stock market performance of growth market plays an important role in the transmission from traditional industries into green and technology-based industries (Holmén and Wang, 2015). In addition, they are more dependent on knowledge capital. Li and Hou (2019) showed that investors are more likely to invest in the enterprises with higher

R&D investments but lower current earnings. On the one hand, they have different performance than the common stock market, and on the other hand, their research provides a basis for the transformation of Chinese enterprises and the behavior of investors. Thus, this thesis will combine time series and sentiment analysis to improve the accuracy of model prediction through comparative analysis.

1.2 Research objectives and structure

Based on previous literature, we hope to use different models to further explore the GEM stock price prediction. There are two main goals for this study:
Q1: Does social media sentiment affect the stock prices of China's GEM price?
Q2: Among the selected models, which one has the best performance?

In the next sections of the thesis, we will combine sentiment analysis, time series analysis and deep learning methods to explore the answers to these two questions. In the second section of the thesis, we will provide a comprehensive literature review, focusing on sentiment analysis related to stock prices and studies concerning GEM. The third section will detail the methodology for collecting Weibo posts and conducting sentiment analysis based on these posts. The fourth section will encompass the empirical analysis, where we will select, fit, optimize, and evaluate the performance of various models. Finally, we will discuss the results obtained and offer insights for future research directions.

2. Literature review

2.1 Methodology of prediction on stock market

The use of time series models for stock price forecasting has a long history since seventies. Box and Jenkins (1999) developed the Auto-Regressive Integrated Moving Average (ARIMA) model utilizing only the historical data of price and volume, which is one of the earliest research projects in the field. These models are proved to be too simple to make predictions due to their high requirement for data and simple structure.

Recent years, with the development of data science, new machine learning algorithms are constantly being invented. These methods can better handle dynamic, non-linear data relationships, capture potential features in the data, and process large amounts of data with higher efficiency, thus bringing greater accuracy and stability to predictions. Manish and Thenmozhi (2005) used SVM and random forest to predict the movement of S&P CNX NIFTY Market Index of the National Stock Exchange and find that SVM and found that they

have a better performance than traditional models. Kara et al. (2011) used support vector machines (SVM) and artificial neural networks to predict the daily movements of the Istanbul Stock Exchange National 100 Index from 1997 to 2007. YAO (2024) used the DLWR-LSTM model to forecast stock price volatility, which can separate volatility from trend to achieve better forecasting effect. Gülmez (2023) uses Artificial Rabbits Optimization algorithm (ARO) to optimize the hyperparameters of an LSTM model and improve the accuracy of stock market predictions. Under the influence of these new algorithms, especially neural network algorithms, the prediction accuracy of financial data has been significantly improved.

2.2 Sentiment analysis for investors

NLP (Natural Language Processing) is an important branch of the field of artificial intelligence, which strives to make computers understand and process human language. Sentiment analysis, as an important technique of NLP, has been widely applied in various filed including social media. In comparison to financial news websites, social media text reflects a more diverse and mainstream range of emotions, and it provides a larger textual dataset. So far, many researchers have used social media to capture investor's sentiment of stock market. Baberis et al. (1998) built an investor sentiment model based on the under-reaction and over-reaction of stock prices to information such as earnings announcements, and found that news with higher weight had a greater impact on investor behavior. Antweiler and Frank labelled the data with positive and negative and measures the relative bullish degree of investors. Boolen, Mao, Zeng (2010) took into account the complexity of emotions and developed the GPOMS classifier. They used OpinionLeader to categorize investor sentiment into six types: calm, alert, certain, important, kind, and happy. The study also considered the interaction between these six emotions and found that the combination of calm and happy have better model performance. Sprenger, Sandner, Tumasjan and Welpe (2014) extract more than 400,000 Twitter messages about news events and examine their effect on S&P 500 stock prices. They argued that social media is more sensitive to emergencies and more accurate in capturing public sentiment than traditional media such as publications and message boards. They sorted the news according to the type and sentiment of the event and find that not only sentiment, even the type of events may affect the market reaction, especially positive ones (e.g., M&A or Earnings). Corea (2016) applied OLS for price and LPM for trend and found that the quantity rather than the sentiment has a more significant influence on stock market, inconsistent with some other studies. Mazboudi and Khalil (2017) found that large acquirers can use corporate social media to stabilize their stock price.

For Chinese stock market, many researchers based their study on the mood of Sina Weibo, the largest social media platform in China and one of the most

visited websites in Mainland China. Fan, et al. (2014) studied the correlation of four different emotions through the messages posted on Sina Weibo and found that the posts expressing anger had the highest correlation. Li, Ann (2023) derived individual investor sentiment indexes from over 2.4 million social network posts in Sina Weibo and compared the sensitivity of both positive and negative posts and found asymmetricity patterns in the effect.

Based on the above information obtained from individual investor sentiment indexes above, the research on investor sentiment is extensive. Although researchers have different choices on research methods and data acquisition methods, it can be said that the research has been carried out to a relatively in-depth degree. Most studies confirm the effect of sentiment on stock price pricing. The research review reveals that experts use different methods and come to various conclusions. This highlights how financial markets can be unpredictable and complex. Each market behaves uniquely, and the same market can change its behavior over time. The choice of models and indices also plays a role in different results.

However, most of the studies mentioned above focus on the main board market, and there are few studies on small and medium-sized board or GEM. Since most of the companies listed on the main board market are mature and stable enterprises with good profitability, investors have higher trust in them, while the small and medium-sized board enterprises have low market liquidity and high price volatility, therefore are faced with higher risks. In addition, according to the study of Gui et al. (2021), the GEM market is not as well developed as the main board market, so the number of investors is smaller and there are less relevant posts can be collected for research, which makes sentiment analysis difficult.

3. empirical analysis

3.1 Methodology

In order to verify the statement in literature overview section, we decide to choose three different analysis methods to analyze GEM stock price: one is the traditional univariate time series method, another is the multivariate time series method with the introduction of sentiment score, and the third is the deep learning method. In the following sections, we will fit the models and compare the conclusions of the three models.

In terms of variable selection, we hope to introduce other relevant variables while studying the autocorrelation of GEM price. The explanatory variables we

introduce include Weibo sentiment score and SSEC price, hoping to explain the stock price fluctuation from both macro and micro levels.

We also need to define the methodology for model evaluation. Different from machine learning, the train-test split for time series analysis must be based on the sequence of time. In the thesis, 70% of the data set is used to train models, while 30% are used to evaluate their performance. That is, the data before 3th January is used to fit the models while the rest is used for evaluation. Three statistical metrics, including mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE) are chosen to qualitatively analyze the forecast performance of all the models. Typically, smaller values indicate better accuracy in prediction and vice versa. All of the 3 metrics are suitable for both traditional time series analysis models and deep learning models.

The metrics can be expressed as below:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

While y_i is true value, $\hat{y}_i$ is predicted value, n is the count of observations.

3.2 Descriptive statistics and data source

3.2.1 Textual data acquisition

The time range is chosen from early November 2023 to the end of February 2024 for two main considerations. First, The New Year can be seen as an environmental factor affecting investor sentiment due to holiday effect. The year-end settlement behavior of companies also has an impact on financial markets. Second, since it's close to the research time, the data is easier to obtain.

The textual data are obtained from Sina Economic, which can be regarded as a collection of news from a specific period. Compared to scraping all posts from trending topics, these posts mainly focus on social and financial issues, which have a more pronounced impact on the stock market. Additionally, the smaller volume of text makes the data easier to process and analyze. The tool *Weibo Spider*, designed and shared on GitHub, was used to gather this data.

In total, there are 6,454 posts, averaging about 70 posts per day, though the exact number may vary.

The data procession of textual data is consisted of three steps using *jieba*. Compared to other Chinese text processors like *SnowNLP*, *jieba* has a more neutral attitude toward words. First, the text is segmented into words to help the model better understand the text. In the next step, special characters and punctuation are removed. Stop words, which do not contribute to sentiment, are also extracted from the text to improve the model's efficiency. For this, we use the stop words dictionary designed by Fudan University, a widely-used stop words list in sentiment analysis. It contains function words like auxiliary verbs and prepositions that do not have actual meaning in Chinese. Finally, the count of positive and negative words in each post is determined using cnsenti, a Chinese sentiment analysis tool.

To calculate the daily overall score of sentiment, I use the metric below:

$$Score = \left(N_{pos} - N_{neg}\right) / \left(N_{pos} + N_{neg}\right)$$

While $Score$ is daily sentiment score, N_{pos} is the count of positive words and N_{neg} is the count of negative words.
The score is a decimal between 0 and 1. The closer it is to 1, the more positive the sentiment and vice versa. Rather than categorizing, scoring based on the number of emotion words makes the data easier for subsequent time series analysis.
The curve of the time series is shown in figure 1:

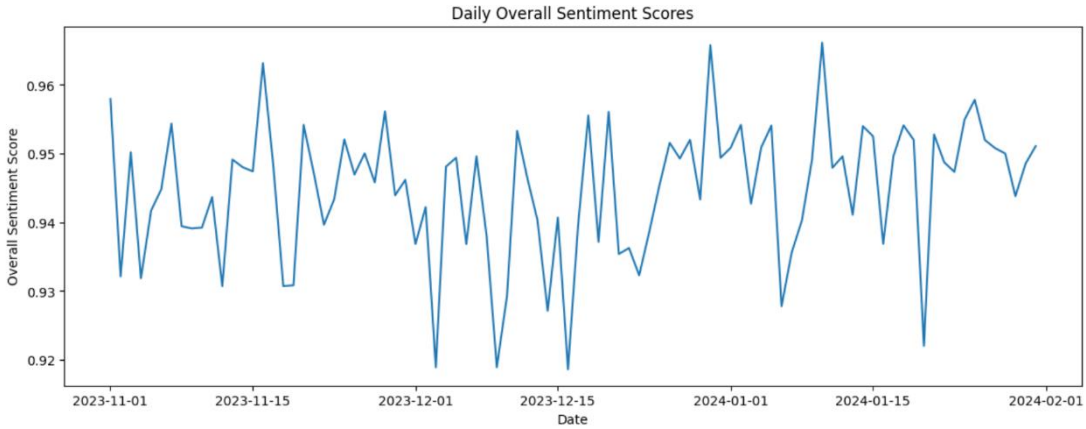


Figure 1 Daily overall sentiment score

The daily sentiment score for the chosen time span can be regarded as time series with 92 observations. For most days, the scores are between 1 and 0.9, indicating that the sentiment is positive in most situations.

It can be seen from the plot that the time series is non-stationary. The overall score in December is lower that November and January. Besides, it has a

larger volatility in January and February. This may be due to the increased use of social media as the Spring Festival approaches. As a result, feedback to both positive and negative news is more pronounced.

Augmented Dickey-Fuller test (ADF test) is a statistical method used to detect the existence of unit root in time series data, also known as unit root test. For daily sentiment score, the P-value is 2.05, larger than the critical value (-2.89). This implies the presence of unit roots and the non-stationarity of the time series.

3.2.2 Stock prices acquisition

Due to the reason that the sentiment on a daily basis, the GEM daily stock price is represented with closing price for each day, which can be obtained from eastmoney, a website with information of Chinese economics and finance. As a time series, it's non-stationary with seasonal differences and trends. Overall, the price shows a decreasing trend and the volatility is higher in January. This may be the result of annual effect. Investors tend to adjust their portfolios at the start of a year, potentially driving up certain sectors. Besides, the investors tend to have an optimistic attitude about the new year, thus increasing their investment.
This result is also shown in the P-value of the ADF test, which is 0.98, much higher than the 5% critical value. The time series also gets to stationary after first differencing.

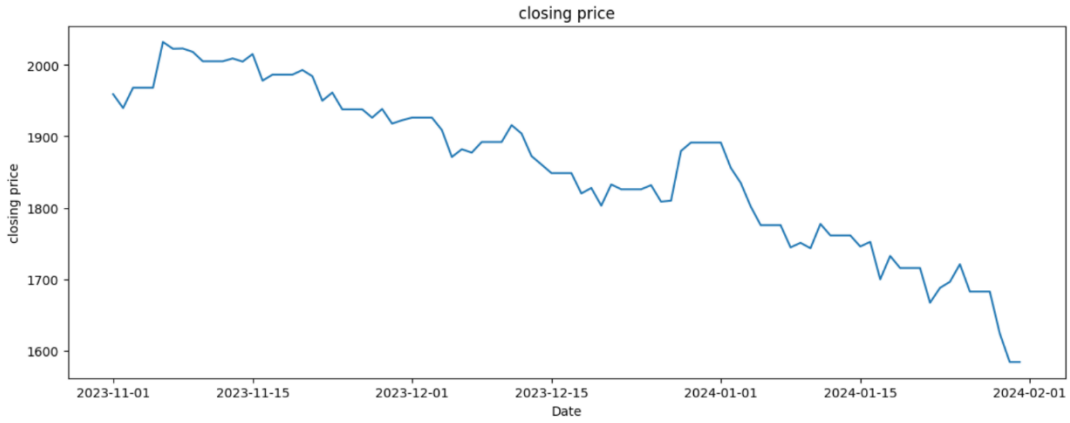


Figure 2 GEM daily closing price

In order to make the model more generalized, more explanatory variables need to be introduced. the researchers made different choices in terms of the selection of explanatory variables. The Shanghai Composite Index (SSEC) represents the performance of the entire Shanghai stock market, which reflects the overall trend and sentiment of the market. While macroeconomic index like GDP and GNP are not suitable for monthly analysis, SSEC can help capture the overall movement of the market, thereby improving the accuracy of the prediction of individual stock prices. Fluctuation in the Shanghai Index

may have an impact on the stock prices of SSE market, as the overall trend of the market affects all stocks. However, when making prediction, only historical information can be used since the observation for the same period remains unknown. After the same way of preprocessing, the trend of SSEC closing price is show in figure 3. In terms of trends, the two prices are broadly in line.

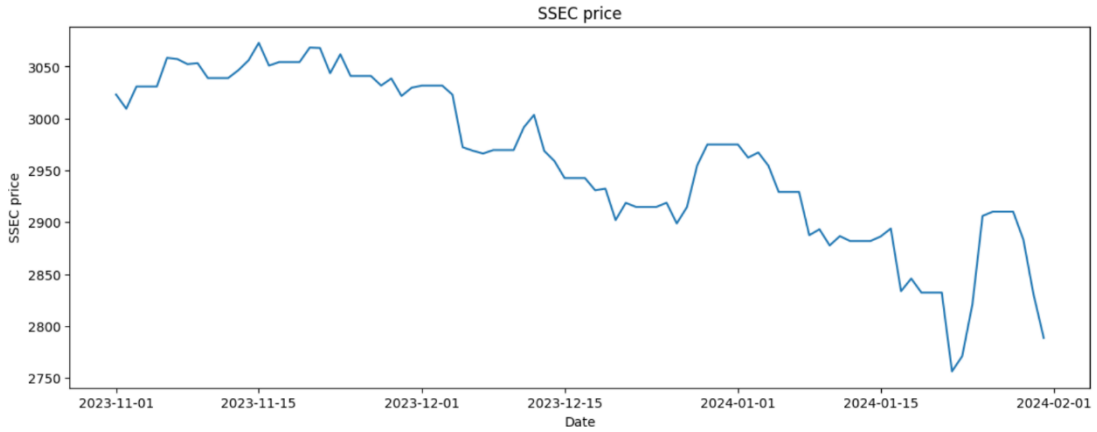


Figure 3 SSEC daily closing price

Finally, the statistical information of all the variables is listed in table 1. Through comparison between SSEC price and GEM price, we can see that the mean of SSEC price is significantly larger than that of GEM price. In addition, the standard deviation of the SSEC price is also larger than the GEM price, which means that the SSEC price is more volatile.

Table 1 Statistical information

variable	count	mean	std	min	25%	50%	75%	max
SSEC price	92	2960.266	678.173	2756.340	2909.193	2968.005	3031.655	3072.830
GEM price	92	1856.991	111.445	1583.770	1771.975	1874.775	1938.740	2032.340
sentiment score	92	0.945	0.010	0.919	0.939	0.947	0.951	0.966

3.3 Univariate time series prediction

3.3.1 Performance of ARMA model

ARMA model is the most common model used for time series analysis. It's consisted of autoregressive (AR) component and moving average (MA) component. The model is based on both its own lag and the past noises. Due to the simple structure, it has a higher calculating efficiency. On the other hand, the model is only compatible for stationary time series and only take the historical information into consideration without other potential factors. As a

result, it has lower accuracy and not suitable for long term prediction.

The parameter p, q for ARMA model can be estimated with PACF and ACF of the series. ACF describes the autocorrelation between one observation and another, including both direct and indirect correlation information, while PACF only describes the direct autocorrelation with another observation.

Figure 4 shows ACF and PACF of the first difference of closing price:

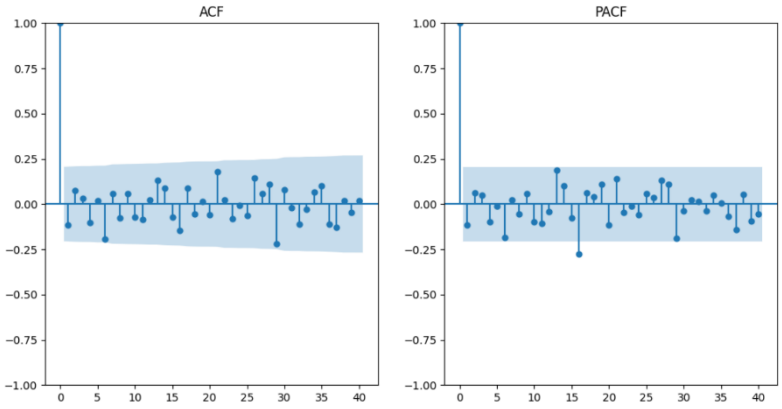


Figure 4 ACF and PACF for GEM price

Both ACF and PACF decrease sharply to 0 after 1 lag, indicating that ARMA(1,1) may have the best performance. The model can be expressed as below:

$$x_t = \varphi_1 x_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon$$

while x_t represents the closing price at the moment t, ε represents the white noise, φ_1 represents the autoregressive coefficient and θ_1 represents the moving average coefficient.

The plot of the fitting can be show as below:

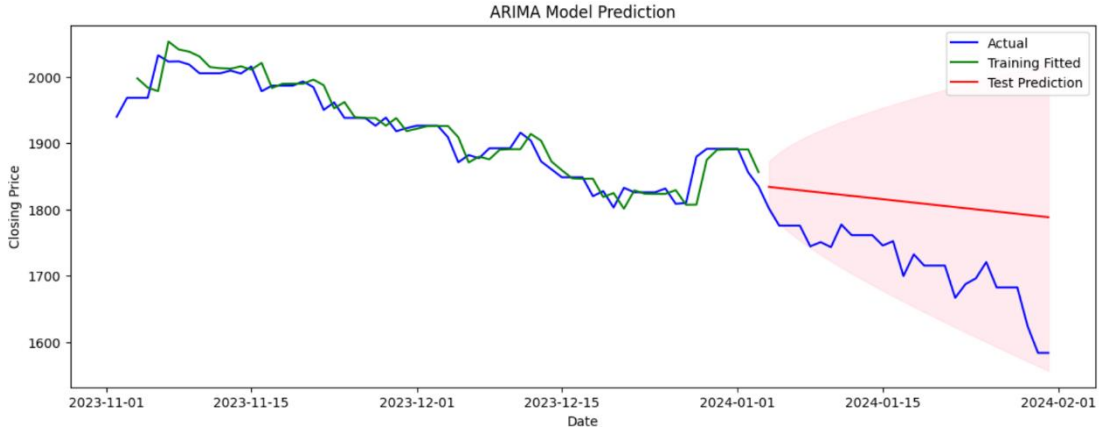


Figure 5 performance for ARMA model

The MSE, RMSE and MAE are 9778.340, 98.885 and 89.804 for each. From the images and metrics, we infer that the result is not ideal. Although the prediction stays within the 95% prediction range, the real data is always higher than the predicted value, and the difference becomes larger over time. Considering only the impact of the model itself, we find that the ARMA model does not accurately capture all the features of the input data. Besides, as shown in table 2, the P-value for each coefficient is larger than 0.05, indicating that neither of them is significant. Since the stock price soared on January 1th, ARMA model could not handle it reasonably, and the data of each subsequent period were predicted again on the basis of the previous high forecast value, leading to the result of increasing deviation. More complex models are required to make better prediction.

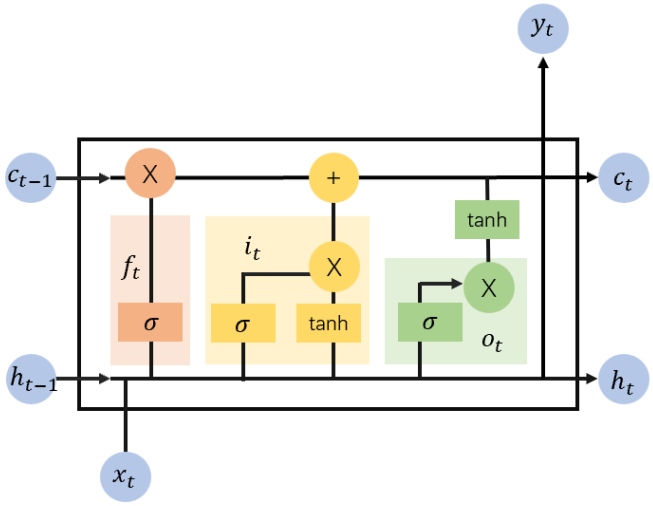
Table 2 ARMA model significant testing result

	coef	std.error	z	P>|z|
ar.L1	-0.3805	1.362	-0.279	0.78
ma.L1	0.2918	1.372	0.213	0.832
sigma2	460.7217	57.406	8.026	0

3.3.2 Performance of LSTM model

LSTM, a special type of neural network, has a much more complex structure than traditional time series analysis models like ARMA model. As a deep learning model, it can be used to describe non-linear relationships and can store long-term historical information effectively. The model is oriented from RNN (Recurrent Neural Networks). The difference from a standard feedforward neural network is that an RNN has connections not only to the inputs and outputs but also has loops within its network units. This loop allows information to be passed from one step of the network to the next. LSTM has a more flexible performance than traditional RNN models in processing memory. LSTM's ability to effectively remember long-term information and suppress unwanted information interference allows it to capture complex patterns and dependencies in longer sequences. It can also effectively deal with the problems of gradient vanishing and gradient explosion. (Md et al., 2023).

The structure of the model has 3 layers in total: input gate, forget gate and output gate. The input gate is responsible for the inflow of new information. The forget gate uses a sigmoid activation function to output a value between 0 and 1, controlling the proportion of input retained. The output gate determines the output value of the LSTM unit. It uses a sigmoid function to determine which parts of the cell state will be output, and then activates the cell state through a tanh function. The structure can be shown below:



While f_t is forget gate, i_t is input gate, o_t is output gate, c_t is cell state.

As deep learning model, LSTM model has more tuning parameters than other models: ①Time step. The time step determines how many previous data are used to predict the next data. The larger the time step, the larger the impact of the long-term data. ② Count of hidden layer. It decides the structure of the model. Model with more hidden layers has better ability in capture complex relationship, while also faced with more possibility of overfitting. ③batch size. Batch size indicates the number of data passed to the model for training at a time. When using gradient descent, the proper batch size can make the loss more accurate. 4. Count of iterations. The more iterations, the smaller the loss. However, overfitting should be avoided when adjusting the number of epochs.

After fitting and tunning the model, we find that the model has the best performance with 3 layers, with 50, 10 and 1 units in each layer. The batch size, time step and epoch are 40,4 and 50 for each. This indicates that 40 samples are used for each update, each input sequence consists data for 4 days, and a total of 50 rounds of iteration are performed. The model with the best performance can be plot as below:

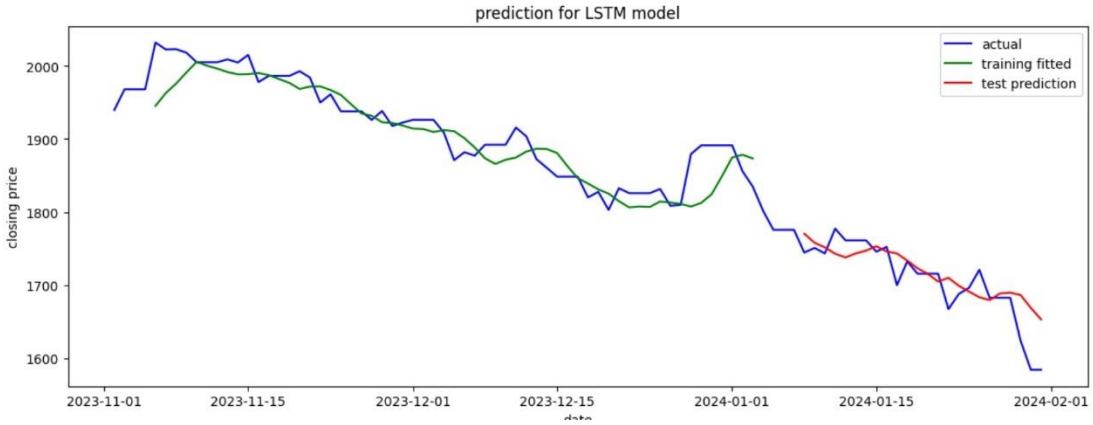


Figure 6 performance for LSTM model

Compared with ARMA model, LSTM has a significantly better performance. The figure shows that the predicted values from the LSTM are very close to the actual values in the training set. Evaluation metrics confirm this, with the LSTM model achieving an RSE of 1507.87, an RMSE of 38.831, and an MAE of 28.558, each of which is lower than those for the ARMA model.

3.4 Multivariate time series prediction

3.4.1 Cointegration test

Although several time series may not be stationary, some linear combination of them may be stationary. Cointegration means that a linear combination of multiple variables is stationary in the long run. When conducting multivariate time series analysis, it is necessary to conduct co-integration test. Normally, Johansen cointegration test is used for more than 2 variables.

The trace test for Johansen test is a joint significance test with 3 hypotheses based on the number of cointegrate relationships. There are up to K-1 cointegration relationships for k independent variables, and up to 2 cointegration relationships for three variables.
For the first hypothesis, H0: There exists no integration relationships.
For the second hypothesis, H0: There is at most one cointegration relationship.
For the third hypothesis, H0: There are at most two null hypotheses for cointegration relationships.

Table 3 Johansen test result

Eigenvalue statistics	Critical values (90%)	Critical values (95%)	Critical values (99%)
0.36673905	32.0645	35.0116	41.0815
0.20603088	16.1619	18.3985	23.1485
0.10222109	2.7055	3.8415	6.6349

All the eigenvalue statistics are significantly smaller than the critical value for all the critical values, indicating that all of the H0 hypothesis cannot be rejected, that is, there is no stationary linear combination of variables in the long term.

3.4.2 Granger causality test

Before fitting multivariate model, Granger causality test is needed to be carried out with stock price and sentiment scores. Granger causality tests are used to test whether one set of time series is the cause of another. The H0 assumption of the Granger causality test is that if the variable A Granger causes the variable B, then the past value of A should help explain the present change in B. If P-value is smaller than the critical value, then H0 hypothesis is rejected, and the two variables can be used to explain each other. Here we applied the test directly to the first differences of the 3 variables.

In the first test, the H0 hypothesis is that sentiment score does not granger cause GEM stock price. The P-value for each lag is shown in table 4:

Table 4 Granger test result between sentiment score and GEM closing price

Number of lags	F	P-value
1	0.0001	0.9922
2	0.261	0.7709
3	0.1798	0.9098
4	1.1567	0.3365
5	2.2746	0.0556
6	2.2765	0.0455*

The correlation between GEM and SSEC price also need to be find out. Applying the same Granger causality test to SSEC price and GEM price, H0 hypothesis is that SSEC stock price does not granger cause GEM price. The result is shown in table 5:

Table 5 Granger test between GEM and SSEC price

Number of lags	F	P-value
1	3.9136	0.0511
2	2.979	0.0562
3	2.4553	0.069
4	2.1089	0.0536
5	2.6846	0.0146*
6	2.3604	0.0183*

In table 4, P-value gets less than the significance level after the sixth lag, indicating that there exists a significant causal relationship between sentiment scores and prices. In table 5, p values are less than the significance level in both the fifth and sixth lag periods, which means that SSEC granger affects

GEM in the fifth and sixth lag periods. Therefore, it is possible to use the historical sentiment scores and the market performance of SSEC to predict the trend of GEM. P-value. However, we cannot say that the fluctuation of emotion caused the change of stock price, that is, the direction of causality between the two variables cannot be determined.

3.4.3 Performance of VAR model

There is no long-term integration between variables. However, there still exists granger causality between them and historical SSEC price and sentiment score can still be used to make predictions on GEM price.

When selecting traditional multivariate time series analysis models, it is common to determine the model type based on the presence of long-term cointegration relationships. If cointegration relationships exist among the variables, the Vector Error Correction Model (VECM) should be used; if not, the Vector Autoregressive (VAR) model should be applied. Compared to the VAR model, the VECM model adjusts the relationships between variables using error correction terms to maintain their long-term equilibrium. Based on the results obtained in previous sections, it is appropriate to establish a VAR model in this context.

Derivate from AR model, VAR (variant autoregression) model can capture dynamic relationships between multi variables. Each time series variable is considered to be a linear combination of its own lags and is also affected by the lag value of other time series variables. It requires all the variables to be stationary but does not need cointegration, so the first difference is applied to all the variables before fitting the model.

The count of lags needs to be determined before regression. Although the Granger causality test shows that there is a causal relationship between variables at the fifth and sixth lags, this cannot be directly used to fit a VAR model. The Granger causality test focuses on the causal relationships of individual lags, whereas the VAR model determines the lag order based on the overall performance of the model. Therefore, while causality tests may identify several significant lags, the VAR model might select a lower lag order to avoid overfitting and improve forecasting accuracy. When fitting a VAR model, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are typically used to determine the optimal lag order. For the VAR model in this study, both AIC and BIC indicate that the model performs best with 1 lag.
The model with 1 lag has expression as below:

$$GEM\_price = \varphi_0 + \varphi_i Y_{t-1} + \varepsilon_t$$

$$Y_{t-1} = \begin{bmatrix} GEM\_price \\ Scores \\ SSEC\_price \end{bmatrix}, \quad \varphi_i = \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} & \varphi_{1,3} \\ \varphi_{2,1} & \varphi_{2,2} & \varphi_{2,3} \\ \varphi_{3.1} & \varphi_{3.2} & \varphi_{3.3} \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

While φ_0 represents constant term, φ_i represents coefficient, ε_t represents white noise, GEM_price represents GEM closing price, $Score$ represents daily Weibo sentiment, $SSEC\_price$ represents SSEC closing price.

In order to make the time series stationary and facilitate the fitting of VAR model, we carry out difference for each variable. However, the metrics cannot be directly applied to differenced fitting results. In the following steps, it is essential to reverse the differencing when evaluating the overall performance to ensure fairness and comparability of the results. Compared with the difference model which only shows the predict for volatility, the reverse differenced model can reflect the prediction of the long-term trend. The reverse differencing of the observed values is calculated as the cumulative sum of the predicted values obtained after differencing, starting from the first observed value. The figure below shows the fitting curves after differencing and after reversing the differencing:

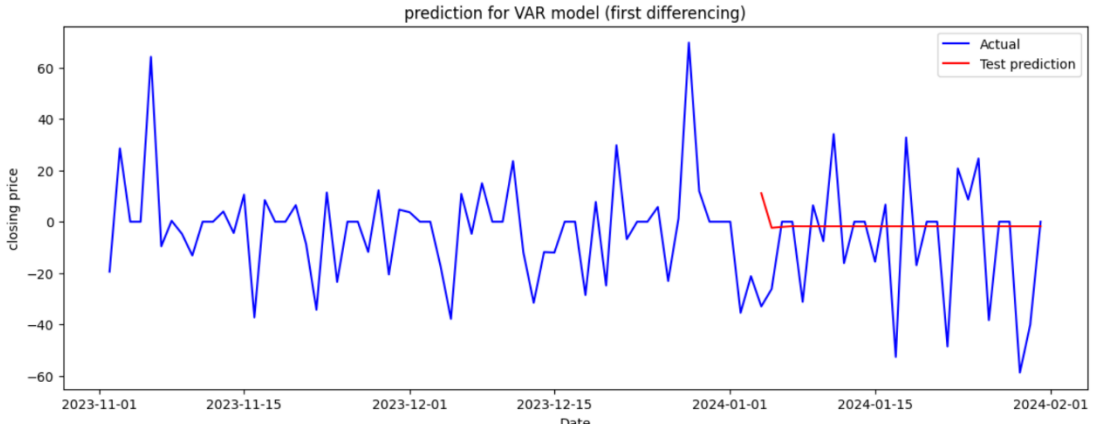


Figure 7 performance for VAR model (differenced)

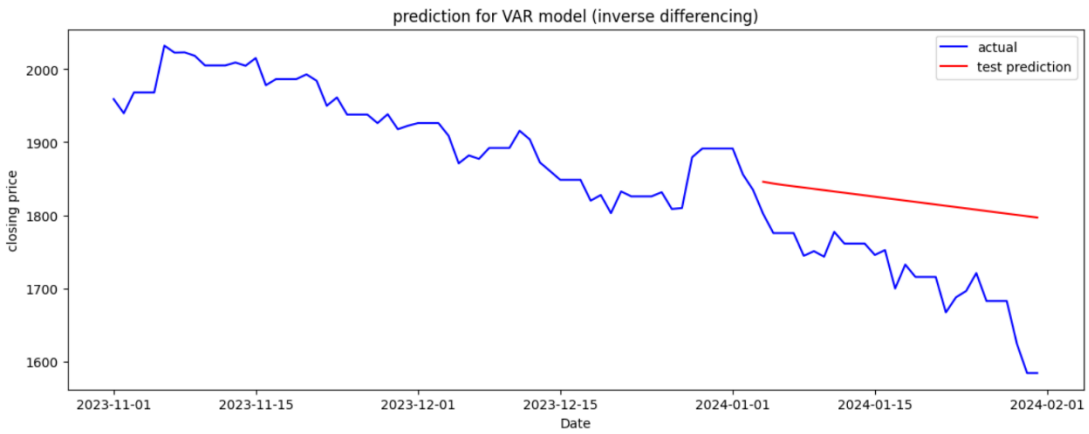


Figure 8 performance with VAR model (reverse differenced)

From the graph, we can see that the forecast price of GEM tends to be straight after the first two forecast periods. Like the ARMA model, the VAR model does little to capture the volatility of variables. In addition, sentiment scores and SSEC prices have almost no impact on GEM prices. From the numerical point of view, it has the lowest comprehensive performance among the three models, with the three metrics are MSE (10360.188), RMSE (101.785) and MAE (93.651) for each.

We also need to compare the result for significant testing for each variable. All of the coefficient for equation of GEM price is larger than 0.05, illustrating that they do not have a significant impact on the result. Among the test results for all the coefficients, only the coefficient between sentiment score and its own lag is significant.

Table 6 significant testing results for equation GEM price

	coefficient	std.error	t-stat	prob
const	-1.80076	2.517314	-0.715	0.474
L1.X	0.11701	0.267683	0.437	0.662
L1.Y	-10.3634	190.5519	-0.054	0.957
L1.Z	-0.30781	0.348266	-0.884	0.377

Table 7 significant testing results for equation sentiment score

	coefficient	std.error	t-stat	prob
const	0.000252	0.001469	0.171	0.864
L1.X	-0.00097	0.000156	-0.618	0.536
L1.Y	-0.46257	0.111205	-4.16	0
L1.Z	0.000275	0.000203	1.354	0.176

Table 8 significant testing results for equation SSEC price

	coefficient	std.error	t-stat	prob
const	-0.640399	1.934852	-0.331	0.741
L1.X	0.123795	0.205746	0.602	0.547
L1.Y	2.732623	146.4615	0.019	0.985
L1.Z	-0.169933	0.267684	-0.635	0.526

Finally, the performance of all the models is integrated and compared. The table shows the performance of the three models using different metrics. Among all these models, LSTM model has the significantly much superior performance across all criteria, while the ARMA and VAR models show less satisfactory results. The ARMA and VAR models both use a lag of 1, indicating they do not utilize much historical information. Additionally, the relationships between the three variables might be weak or non-linear, which the VAR model struggles to capture. Therefore, even though the VAR model is more complex, it does not perform better than the ARMA model, as the ARMA model adequately describes the time series characteristics and its prediction errors are mainly influenced by the historical values of individual series.

Table 5 Model comparison based on valuation metrics

Model	MSE	RMSE	MAE
ARMA (1,1)	9778.340	98.885	89.804
LSTM	1507.870	38.831	28.558
VAR (1)	10360.188	101.785	93.651

4. Discussion

In the thesis, we mainly include 2 tasks: the first is to analyze Weibo sentiment with cnsenti and jieba, while the second is model fitting and comparison. For the first task, we find that sentiment for most days is positive and the daily fluctuation is not obvious, influenced by both the way the data is obtained and analyzed.

For the second task, we select the potential variables and model for comparison. We compared three models in total: the first is a traditional univariate ARMA model, the second is a univariate LSTM model, and the third is a VAR model that incorporates sentiment analysis from Weibo posts, including three time series—SSEC closing price, GEM closing price, and sentiment score. We evaluated the models using MSE, RMSE, and RSE metrics, as well as by examining the fitting curves. The results show that the LSTM model significantly outperforms the other two models, likely due to the following reasons:

1. Complex relationships: The ARMA and VAR models are too simplistic to capture the complex relationships between variables. Financial markets exhibit intricate behaviors, and both ARMA and VAR models are limited to linear combinations of variables, unable to capture non-linear relationships or complex dynamic patterns. In contrast, the LSTM model's complex structure allows it to handle these complexities more effectively.

2. Flexibility in handling relationships: LSTM models are better equipped to manage both long-term and short-term time series. Unlike ARMA and VAR models, which rely on fixed lag periods, LSTMs has 3 gates and a memory unit to retain long-term information and quickly update short-term data, resulting in more accurate modeling.

3. Data Processing and Model Performance: The superior performance of the LSTM model is partly due to its ability to capture the intricate relationships between variables. Additionally, the performance can be influenced by data processing methods, choice of time span, and data obtain techniques.

5. Conclusion and limitation

Overall, the advanced performance of LSTM in modeling complex and dynamic relationships contribute to its better performance compared to traditional models. This result shows the superiority of deep learning models in financial market forecasting. The success of LSTM shows that deep learning methods can overcome the limitations of traditional time series models, especially when dealing with real-world problems that have complex dependency structures. This predictive capability has considerable economic value for both individual investors and businesses, making LSTM one of the most popular models in stock market prediction nowadays. Whether the LSTM model retains good generalize ability in other sectors of the Chinese financial market or in markets of other countries is a question to be answered by researchers.

However, there is still no clear conclusion about the selection of external variables. In the most popular opinion, the fluctuations in the stock market can be a complex multivariate process. However, there are also many researchers demonstrate that whether it makes sense to introduce eternal information is still remains to be discussed. Although sentiment is recognized as a significant factor that leads to stock market fluctuations in numerous studies, there are also other studies that only based on time series. Some researchers regard it as an indicator for the soundness of the market. Studies by Chen et al. (1997) and Zhang et al. (1999) show that the Chinese stock market has reached weak efficiency in the late 1990s, and the introduction of market external information does not necessarily significantly improve the prediction accuracy. For this thesis, we did not capture or elaborate on such relationships. It is worth exploring whether deep learning models can describe these relationships when multivariate linear models have no effect.

From another perspective, 4, which is of great significance to both government departments and enterprises.

As for the methodology and model selection for the thesis, there remain some limitations to be questioned for the thesis. Here we list two of them:

1. More complex and cutting-edge models may be selected. In this thesis, we only use the general LSTM model. However, on the basis of LSTM model,

many models with more complex structure and more complete function have been developed, which have been briefly introduced in the literature review. Whether these models have the ability to generalize in GEM market is worth further exploration.

2. There are two possibilities for the reason not to catch the multivariate relationships: ① GEM prices are very little affected by external factors. ②Failure to introduce more effective predictors. If we only consider the second influencing stock prices include the macro market, political environment, exchange rates, and more, which we do not deeply explore in the thesis. Therefore, future research could consider including additional variables to explore more complex interactions between them. For instance, incorporating industrial indices, GDP, or policy factors. Due to the constrain of daily data, this study could only include the SSEC as a variable to measure the overall market trend, which has certain limitations. Given that the GEM is more influenced by macro policies, introducing other variables might have better forecasting results.

Reference

Ackert, L.F., Jiang, L., Lee, H.S., Liu, J., 2016. Influential investors in online stock forums. International Review of Financial Analysis 45, 39–46. https://doi.org/10.1016/j.irfa.2016.02.001

Ahangar: The comparison of methods artificial neural... https://scholar.google.com/scholar_lookup?title=The%20comparison%20of%20methods%20artificial%20neural%20network%20with%20linear%20regression%20using%20specific%20variables%20for%20prediction%20stock%20price%20in%20tehran%20stock%20exchange&publication_year=2010&author=R.G.%20Ahangar&author=M.%20Yahyazadehfar&author=H.%20Pournaghshband (accessed 8.13.24).

Antweiler, W., Frank, M.Z., n.d. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards on JSTOR [WWW Document]. URL https://www-jstor-org.ludwig.lub.lu.se/stable/3694736 (accessed 3.6.24).

Baker, M., Wurgler, J., n.d. Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance.

Bao, Y., Li, J., Zhu, Y., 2018. Mammary Analog Secretory Carcinoma With ETV6 Rearrangement Arising in the Conjunctiva and Eyelid. The American Journal of Dermatopathology 40, 531. https://doi.org/10.1097/DAD.0000000000001062

Barberis, N., Shleifer, A., Vishny, R., 1998. A Model of Investor Sentiment. Journal of Financial Economics 49, 307–343.

Bhandari, H.N., Rimal, B., Pokhrel, N.R., Rimal, R., Dahal, K.R., Khatri, R.K.C., 2022. Predicting stock market index using LSTM. Machine Learning with Applications 9, 100320. https://doi.org/10.1016/j.mlwa.2022.100320

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Broadstock, D.C., Zhang, D., 2019. Social-media and intraday stock returns: The pricing power of sentiment. Finance Research Letters 30, 116–123. https://doi.org/10.1016/j.frl.2019.03.030

Corea, F., 2016. Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector. Big Data Research 4, 70–74. https://doi.org/10.1016/j.bdr.2016.05.001

Fama, E.F., French, K.R., 1992. The Cross-Section of Expected Stock Returns. The Journal of Finance 47, 427–465.

Fan, R., Zhao, J., Chen, Y., Xu, K., 2014. Anger is More Influential Than Joy: Sentiment Correlation in Weibo. PLoS ONE 9, e110184. https://doi.org/10.1371/journal.pone.0110184

Gao, Z., Zhang, J., 2023. The fluctuation correlation between investor sentiment and stock index using VMD-LSTM: Evidence from China stock market. The North American Journal of Economics and Finance 66, 101915. https://doi.org/10.1016/j.najef.2023.101915

Gui, J., Pu, J., Naktnasukanjn, N., Yu, X., Mu, L., Pan, H., 2022. Measuring investor sentiment of China's growth enterprises market with ERNIE. Procedia Computer Science, International Conference on Identification, Information and Knowledge in the internet of Things, 2021 202, 1–8. https://doi.org/10.1016/j.procs.2022.04.001

Holmén, M., Wang, P., 2015. Pyramid IPOs on the Chinese Growth Enterprise Market. Emerging Markets Finance & Trade 51, 160–173.
Kara, Y., Acar Boyacioglu, M., Baykan, Ö.K., 2011. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications 38, 5311–5319. https://doi.org/10.1016/j.eswa.2010.10.027

Kumar, M., Thenmozhi, M., 2006. Forecasting stock index movement: A comparison of support vector machines and random forest, in: Indian Institute of Capital Markets 9th Capital Markets Conference Paper.
Li, J., Ahn, H.-J., 2024. Sensitivity of Chinese stock markets to individual investor sentiment: An analysis of Sina Weibo mood related to COVID-19. Journal of Behavioral and Experimental Finance 41, 100860. https://doi.org/10.1016/j.jbef.2023.100860

Li, X., Hou, K., 2019. R&D based knowledge capital and future firm growth: Evidence from China's Growth Enterprise Market firms. Economic Modelling 83, 287–298. https://doi.org/10.1016/j.econmod.2019.07.005

Mazboudi, M., Khalil, S., 2017. The attenuation effect of social media: Evidence from acquisitions by large firms. Journal of Financial Stability 28, 115–124. https://doi.org/10.1016/j.jfs.2016.11.010

Sprenger, T., Sandner, P.G., n.d. News or Noise? Using Twitter to Identify and Understand Company-Specific News Flow | Semantic Scholar [WWW

Document]. URL https://www.semanticscholar.org/paper/News-or-Noise-Using-Twitter-to-Identify-and-News-Sprenger-Sandner/a9ff381c2d6ec2d1ff69dde323eea42c383e4640 (accessed 3.6.24).

Stambaugh, R.F., Yu, J., Yuan, Y., 2012. The short of it: Investor sentiment and anomalies. Journal of Financial Economics 104, 288–302. https://doi.org/10.1016/j.jfineco.2011.12.001

Tetlock, P.C., n.d. Giving Content to Investor Sentiment: The Role of Media in the Stock Market.
Time Series Prediction With Genetic-Algorithm Designed Neural Networks: An Empirical Comparison With Modern Statistical Models - Hansen - 1999 - Computational Intelligence - Wiley Online Library [WWW Document], n.d. URL https://onlinelibrary.wiley.com/doi/10.1111/0824-7935.00090 (accessed 8.13.24).

Vijayalakshmi, V., 2024. 12 - Implementation of sentiment analysis in stock market prediction using variants of GARCH models, in: Hemanth, D.J. (Ed.), Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications. Morgan Kaufmann, pp. 227–249. https://doi.org/10.1016/B978-0-443-22009-8.00002-1

Zhang, X., Li, G., Li, Y., Zou, G., Wu, J.G., 2023. Which is more important in stock market forecasting: Attention or sentiment? International Review of Financial Analysis 89, 102732. https://doi.org/10.1016/j.irfa.2023.102732

Zhao, C., Wu, M., Liu, J., Duan, Z., li, J., Shen, L., Shangguan, X., Liu, D., Wang, Y., 2023. Progress and prospects of data-driven stock price forecasting research. International Journal of Cognitive Computing in Engineering 4, 100–108. https://doi.org/10.1016/j.ijcce.2023.03.001