Received 7 May 2023, accepted 17 May 2023, date of publication 22 May 2023, date of current version 1 June 2023.
Digital Object Identifier 10.1109/ACCESS.2023.3278790

RESEARCH ARTICLE

A Stock Price Prediction Model Based on Investor Sentiment and Optimized Deep Learning

GUANGYU MU[1,2], NAN GAO[1], YUHAN WANG[1], AND LI DAI[1]
[1]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130000, China
[2]Key Laboratory of Financial Technology of Jilin Province, Changchun 130000, China
Corresponding author: Guangyu Mu (guangyumu@jlufe.edu.cn)

This work was supported in part by the National Social Science Fund of China under Grant 19BJY246, in part by the Natural Science Fund Project of the Science and Technology Department of Jilin Province under Grant 20190201186JC, and in part by the Science and Technology Development Plan Project of Changchun Science and Technology Bureau under Grant 21ZY61.

ABSTRACT Accurate prediction of stock prices can reduce investment risks and increase returns. This paper combines the multi-source data affecting stock prices and applies sentiment analysis, swarm intelligence algorithm, and deep learning to build the MS-SSA-LSTM model. Firstly, we crawl the East Money forum posts information to establish the unique sentiment dictionary and calculate the sentiment index. Then, the Sparrow Search Algorithm (SSA) optimizes the Long and Short-Term Memory network (LSTM) hyperparameters. Finally, the sentiment index and fundamental trading data are integrated, and LSTM is used to forecast stock prices in the future. Experiments demonstrate that the MS-SSA-LSTM model outperforms the others and has high universal applicability. Compared with standard LSTM, the R^2 of MS-SSA-LSTM is improved by 10.74% on average. We found that: 1) Adding the sentiment index can enhance the model's predictive performance. 2) The LSTM's hyperparameters are optimized using SSA, which objectively explains the model parameter settings and improves the prediction effect. 3) The high volatility of China's financial market is more suitable for short-term prediction.

INDEX TERMS Deep learning, LSTM model, stock price prediction, sentiment analysis, sentiment dictionary, sparrow search algorithm.

I. INTRODUCTION

With the maturity of China's stock market and the rapid growth of Internet finance, many people realize the importance of investment and choose to enter the financial market. However, the stock market is characterized by massive data and enormous volatility. Many retail investors need more data-mining skills to make money. Therefore, accurate stock price prediction can reduce investment risks and improve investment returns for investors and enterprises.

Early scholars used statistical methods to construct a linear model to fit the stock price time series trend. The traditional methods contain ARMA, ARIMA, GARCH, etc. The ARMA is established to conduct a time series stock analysis [1]. The ARIMA model is developed based on the ARMA and predicts the trend of stock price changes [2]. The ARIMA model can also introduce wavelet analysis to improve the fitting accuracy of the Shanghai Composite Index [3]. The GARCH model provides innovative ideas for stock time series prediction through a time window [4]. At the same time, some scholars have combined ARMA and GARCH to build a new prediction model, which provided theoretical support for the volumetric price analysis of multivariate stocks [5]. Generally speaking, these classical methods only capture regular and structured data. However, traditional forecasting methods require assumptions that are uncommon in real life. Therefore, It is challenging to describe nonlinear financial data using statistical methods.

Subsequently, many researchers attempt to anticipate stock prices using machine learning approaches such as Support Vector Machines (SVM) and Neural Networks. Machine learning's core idea is to use algorithms to parse data, learn from it, and make predictions about new data. Because the SVM shows unique benefits in dealing with limited samples,

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Ching Ying.

high-dimensional data, and nonlinear situations, many scholars use it in stock forecasting. Hossain and Nasser [6] found that the SVM method is superior to the statistical ones in stock prediction accuracy. Chai et al. [7] suggested a hybrid SVM model to anticipate the HS300 index's ups and downs and found that the least squares SVM combined with the Genetic Algorithm (GA) performed better. However, when the SVM applies to large-scale training samples, much memory and computing time will be consumed, which may limit its development space in predicting a large amount of stock data. Then, Artificial Neural Networks (ANN) and multi-layer ANN address financial time series issues. According to the experimental data, ANN has the benefits of quick convergence and high accuracy [8], [9], [10]. Moghaddam and Esfandyari [11] evaluated the effect of several feedforward artificial neural networks on the market stock price forecast through experiments. Liu and Hou [12] improved the BP (Back Propagation) neural network using the Bayesian regularization method. Nevertheless, the traditional neural network method has the following areas for improvement. Generalization ability is not strong, quickly leads to overfitting, and falls into local optimization. Since many samples need to be trained, better models must be found to solve these problems.

A new model for predicting stock prices is proposed in this paper (MS-SSA-LSTM), which matches the characteristics of multi-source data with LSTM neural networks and uses the Sparrow Search Algorithm. The MS-SSA-LSTM stock price forecast model can forecast the stock price in advance and help investors and traders make more informed investment decisions. Investors and traders obtain the data of individual stocks they want to invest in, including historical transaction data and comment information of stock market shareholders, and input them into the MS-SSA-LSTM model. The model automatically outputs a stock price trend chart and forecasts the stock price for the next day.

Here, we can get the motivation for this paper.

(1) Adding sentiment indicators to the model features will improve prediction accuracy. However, applying a general dictionary in the financial field cannot achieve good results. Therefore, We need to construct a sentiment dictionary specific to individual stocks.

(2) Stock price series has complicated characteristics such as nonlinearity, high noise, and strong time-variability. LSTM is effective at handling comprehensive time series data. So LSTM network, a deep learning method, is adopted.

(3) In the LSTM network, hyperparameter variation directly affects the model's prediction accuracy. Artificial selection of appropriate hyperparameters for the network model will cost many resources. Therefore, LSTM can be optimized using the Sparrow Search Algorithm proposed in 2020.

Here are the main contributions of this paper.

(1) The SSA-LSTM model incorporates multi-source data that affects stock prices, such as historical trade data and stock forum comments. The MS-SSA-LSTM model outperforms the single data source model regarding prediction accuracy.

(2) Analyze the stock forum comments text information. Then, we construct a sentiment dictionary based on authoritative dictionaries in the financial investment field suitable for individual stock sentiment analysis and sentiment indicator calculation.

(3) A Sparrow Search Algorithm-optimized LSTM model can acquire better hyperparameter values and forecast stock prices than a single LSTM network.

II. REVIEW OF THE LITERATURE

A. PREDICTING STOCK PRICE WITH DEEP LEARNING

With the advancement of artificial intelligence, deep learning provides a new research method for stock price prediction. Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN) are the most common deep learning technologies. RNN can mine time series information in data and predict stock price [13]. However, there are gradient dissipation and gradient explosion in the RNN model, which makes training challenging. The LSTM model realizes the time memory function through the cell gate switch, which can solve the problem that the RNN model cannot describe time series with a long memory [14]. Therefore, the LSTM model predicts the volatility of the stock market index and performs well [15].

In addition, Yan et al. [16] developed a high-precision prediction model based on the LSTM deep neural network in a short-term financial market. LSTM neural network has a higher prediction accuracy than BP neural network and standard RNN and can successfully forecast stock prices. Nabipour et al. [17] selected ten technical indicators as the prediction model's input. The experiment revealed that LSTM outperformed other algorithms in terms of model-fitting abilities, including decision tree, random forest, Adaboost, XGBoost, ANN, and RNN. Aksehir and Kilic [18] suggested a CNN-based model predicting the next-day trading behavior of Dow Jones 30 Index equities. Technical indicators, gold, and oil price data are fed into the model. The accuracy of the results is 3-22% higher than other models based on CNN.

However, scholars usually need to artificially set the hyperparameter values of the LSTM network, which is intensely subjective. The setting of hyperparameters directly controls the topology of the network model, which will affect the model's prediction performance. Artificial selection of the appropriate hyperparameters for the network model will consume many human and computing resources. To solve the above problems, scholars try to use some Swarm Intelligence (SI) algorithms to optimize the hyperparameters of neural networks. SI can search globally to a certain extent and find the approximate value of the optimal solution. Ji et al. [19] demonstrated that the LSTM optimized by the Improved Particle Swarm Optimization model (IPSO) is superior to the LSTM model. Zeng et al. [20] optimized

LSTM by adopting an Adaptive Genetic Algorithm (AGA) and improved the prediction accuracy. At the same time, other intelligent algorithms are also widely used in optimizing neural networks. Sparrow Search Algorithm (SSA) was put forward and inspired by the sparrows' foraging and anti-predation behavior [21]. Compared with the traditional particle swarm and gray wolf optimization algorithms, the SSA algorithm is relatively novel and has a robust optimization ability in price prediction. In addition, the algorithm considers the randomness of individuals in the search process, so it has a solid global search capability. At the same time, the algorithm can also adaptively adjust the search parameters to meet the needs of different problems.

B. FACTORS AFFECTING THE STOCK PRICE

Besides studying the methods of stock price forecasting, scholars also constantly explore factors affecting the stock price. Previous prediction models often take stock of historical data [22], technical analysis indicators [23], and financial information [24] as input features. However, behavioral finance theory research shows investors can only be partially rational when making complicated decisions involving risks and uncertainties. Individual investors account for a large proportion of China's stock market, and their investment decisions are easily influenced by emotions and background factors, causing abnormal fluctuations in the stock market [25], [26], [27]. With the development of text mining technology, more scholars pay attention to quantifying investor sentiment in unstructured texts. Investors' complex sentiments significantly affect the stock prices and the volatility cycle of stock prices [28], [29]. Based on this research, more and more scholars have established different stock prediction models by mining investor sentiment in text data sources [30], [31], [32], [33].

The content of the stock forum is updated in real-time, and the information mined from it is closer to the actual state of investor sentiment than other data sources. However, it will significantly affect investors' investment propensities and a subsequent linkage effect on stock price fluctuations. So it is crucial to analyze the text information of stock forums for our research. Scholars mainly adopt machine learning and semantic analysis methods regarding sentiment classification methods. The machine learning method has high classification accuracy but relies on financial personnel to manually classify training sets, which leads to increased time costs. The semantic analysis method is easier to use in economic and financial analysis, but the standard dictionary is challenging to apply to the economic context [34]. The key lies in constructing a unique financial dictionary.

III. RESEARCH METHODS

A. SENTIMENT ANALYSIS TECHNOLOGY

A sentiment dictionary refers to an emotional corpus that contains words that can identify the emotional characteristics of sentences. In the sentiment dictionary, the categories of words include positive, negative, denial, and degree. Positive words are words with positive emotions; negative words are words with negative emotions; denial words reverse the emotional inclination of sentences, and the degree words increase or decrease emotional intensity [35].

Based on the constructed sentiment dictionary, the sentiment indicator of the stock review text is calculated. We judge the screened text words and the sentiment words in the sentiment dictionary and assign the weights of the sentiment words accordingly. Weight represents the emotional intensity of words. Positive, negative, denial, and degree words have their value ranges. We can adjust the weights of words within the specified range as needed [36]. Then, we determine the sentiment tendency according to the sum of the weights of all words in the sentence [37]. When the sentiment score of the sentence is greater than 0, it represents that the shareholders have positive reviews and bullish attitudes toward the stock, which is considered an excellent opportunity to buy. However, when the score is less than 0, the investors have negative reviews, complaints, and bearish attitudes toward the stock, which may lead to selling the stock. Here, sentiment is divided into bullish, bearish, and neutral.

The stock forum's comments sentiment tendencies on each trading day are calculated. In this article, sent is used as a sentiment indicator, and the formula is as follows:

$$\mathrm{sent} = (\mathrm{pos} - \mathrm{neg})/(\mathrm{pos} + \mathrm{neg}) \tag{1}$$

As shown in (1), pos indicates the number of comments or articles with a positive emotional tendency in a trading day, and neg represents the number of comments with a negative sentiment tendency in a trading day.

B. SPARROW SEARCH ALGORITHM

The Sparrow Search Algorithm (SSA) simulates the sparrows' feed and preemption behavior by constantly updating its position. The SSA divides sparrows into detectors, followers, and monitors [38]. Each position corresponds to a solution. According to the algorithm, the proportion of monitors in the population is 10%-20%. However, detectors and followers dynamically change. In other words, one individual becoming a detector necessarily means that the other individual will become a follower. Detectors offer populations foraging directions and areas. Followers follow the detector for food. Monitors keep an eye on the foraging site and abandon food if danger is detected. By constantly updating the location, sparrows can get the best resources in the foraging process [39].

Assume there are n sparrows within the population. Then the population of all sparrows can be represented as $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]^T$. The fitness function is $\boldsymbol{F} = [\mathrm{f}(x_1), \mathrm{f}(x_2), \cdots, \mathrm{f}(x_n)]^{\mathrm{T}}$.

The formula is generally shown in (2) and (3).

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \tag{2}$$

$$\boldsymbol{F} = \begin{bmatrix} \mathrm{f}([x_{1,1} & x_{1,2} & \cdots & x_{1,d}]) \\ \mathrm{f}([x_{2,1} & x_{2,2} & \cdots & x_{2,d}]) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{f}([x_{n,1} & x_{n,2} & \cdots & x_{n,d}]) \end{bmatrix} \tag{3}$$

where d denotes the variable's dimension to be optimized.

The detector location is updated as (4).

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} \cdot \exp[-i/(a \cdot \mathrm{iter}_{\max})], & R_2 < \mathrm{ST} \\ x_{i,j}^{t} + Q \cdot L, & R_2 \geq \mathrm{ST} \end{cases} \tag{4}$$

where t represents the number of the current iteration. $x_{i,j}^{t}$ denotes the i sparrow's location in the j dimension in the t generation. a is a 0 to 1 arbitrary number. $\mathrm{iter}_{\max}$ is the utmost iteration count. R_2 and ST represent respective alarm values and safety thresholds. Q is a random number drawn from a standard normal distribution. L is the $1 \times d$ matrix, and each of its elements is 1.

The formula for a follower position update is shown in (5).

$$x_{i,j}^{t+1} = \begin{cases} Q \cdot \exp[(x_{worst}^{t} - x_{i,j}^{t})/i^2], & i > n/2 \\ x_p^{t+1} + \left| x_{i,j}^{t} - x_p^{t+1} \right| \cdot A^{+} \cdot L, & i \leq n/2 \end{cases} \tag{5}$$

where x_{worst}^{t} denotes the location of the individual in generation t with the lowest fitness. x_p^{t+1} indicates the location of the $t+1$ generation individual with the best fitness. A shows a $1 \times d$ matrix, in which each member is stochastically preset -1 or 1. $A^{+} = A^{T}(AA^{T})^{-1}$. When $i > n/2$, it implies that the i follower has low adaptability, is not qualified to compete with the detector for food, and must fly to other areas to find food. However, when $i \leq n/2$, followers feed near the best individual x_p^{t+1}.

The formula for a monitor position update is shown in (6).

$$x_{i,j}^{t+1} = \begin{cases} x_{best}^{t} + \beta \cdot \left| x_{i,j}^{t} - x_{best}^{t} \right|, & f_i > f_g \\ x_{best}^{t} + k \cdot (x_{i,j}^{t} - x_{best}^{t})/(|f_i - f_w| + \varepsilon), & f_i = f_g \end{cases} \tag{6}$$

where x_{best}^{t} represents the optimal global location in the t generation. The control step size is β, which follows the standard normal distribution. $k\epsilon[-1, 1]$, ε is specified as a constant to prevent the denominator from being 0. f_i indicates the present individual's fitness value. f_g and f_w indicate the current best and worst global fitness values separately. When $f_i > f_g$ it means that sparrows are outside the colony, extremely vulnerable to predators, and constantly change their positions to obtain higher fitness. When $f_i = f_g$, this sparrow is in the middle of the population and keeps approaching nearby companions to stay away from the dangerous area.

C. LONG AND SHORT-TERM MEMORY NETWORK

The Long and Short-Term Memory Network (LSTM) is an additional Recurrent Neural Network (RNN) variant. In the backpropagation process, the RNN model faces with lack of extended memory, gradient disappearance, and explosion. To address the problems above, the LSTM model introduces a gate mechanism to increase the cellular structure for judging whether the information is valid [40]. As a result, LSTM is suitable for processing and predicting significant events with relatively long periods in practice sequences.

Specifically, a neuron has a cell state and three gate mechanisms in the LSTM model. The memory space of the LSTM model is the cell state, which resembles memory and is the model's critical factor. Cell states change over time, and the gate mechanism determines and updates the recorded information [41]. The gate mechanism allows information to pass selectively, which the dot product operation and sigmoid function realize. Here, the dot product defines the amount of transmitted information, and the sigmoid value is between 0 and 1. When the sigmoid is set to 0, it indicates discarding information. When the sigmoid is set to 1, it means complete transmission [42].

Where the sigmoid function is:

$$\sigma(\mathrm{c}) = 1/(1 + \mathrm{e}^{-\mathrm{c}}) \tag{7}$$

The hyperbolic tangent function is:

$$\tanh(c) = (e^{c} - e^{-c})/(e^{c} + e^{-c}) \tag{8}$$

LSTM has three gates. The forgetting gate determines how much the previous cell state is kept for the present moment [43]. Cells that comply with the rules stay, and states that do not conform are forgotten through the forget door.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{9}$$

Then, the input gate will update the state and determine the amount of altered network input. The sigmoid function determines the value be modified, whereas the tanh function builds a vector. Furthermore, the two are dotted to get the new candidate value.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{10}$$

$$c_t' = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{11}$$

The old cell state is dotted with the forgetting gate f_t. Then, dot-multiply the input gate i_t with the current input state c_t' and add value to update the cell state.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c_t' \tag{12}$$

Finally, the output gate decides the output information. The current cell state treated by the hyperbolic tangent function is dotted with the output part determined by the sigmoid function to obtain the final output result.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

$$h_t = o_t \times \tanh(c_t) \tag{14}$$

IV. MODEL CONSTRUCTION

A. HISTORICAL TRADING INDICATORS

This paper focuses on individual stocks as the research object and selects indexes that can fully reflect stock price changes. The fundamental stock trading indicators include the previous closing price, opening price, closing price, highest price, lowest price, complex price, daily amplitude, trading volume, and turnover rate. The opening and closing prices represent a stock's starting and final prices on a specific day. Trading volume is the number of shares transacted. The stock's closing price usually determines investment profit and daily losses. Therefore, the closing price is used as a predicted target. The complex price is the stock price after the transaction price is repaired. Daily amplitude refers to the fluctuation range from the lowest to the highest point of the stock's current price. The turnover rate is the percentage of the daily turnover of a stock in its circulation, which reflects how frequently the stock is traded. The specific calculation formula is shown below:

$$\text{Complex Price} = C * \beta \tag{15}$$

$$\text{Daily Amplitude}_{\text{n}} = (H_n - L_n)/C_{n-1} \tag{16}$$

where C represents the closing price. β is the cumulative stock price adjustment multiplier. C_{n-1} represents the closing price on the $n-1$st day. L_n is the lowest price on the nth day. H_n is the highest price on the nth day.

The above historical indicators can directly reflect the changes in stock prices and are indispensable for predicting stock prices. Because the deep learning method can mine the stock price rule from the fundamental indicator data [45], the paper only considers historical indicators of the stock.

B. SENTIMENT INDEX BUILDING

Investors' decisions are influenced by not only the technical indicators of the stock but also the sentiments generated by the comments. Therefore, we integrate the sentiment index with fundamental trading data to forecast the stock price. This paper chooses East Money Net as the source of the text information. We use the Octoparse tool to crawl data. Octoparse is a powerful Web data searcher based on Visual Windows, which covers all search requirements and can simulate human operations to interact with web pages. East Money Net provides rich emotional data for research as the stock forum community with the longest development time, the most significant average attention, and the most extensive user activity in China. On the East Money Net (http://guba.eastmoney.com/), we use the Octoparse tool to mine comment data related to stock within a specific time range. The collected target fields include stock code, stock name, post title, and time. The title can get more information than the content [44]. Therefore, this paper's research object is the post's title.

Firstly, we preprocess the collected text information. Empty titles, identical text content, announcements, and forwarded articles are all noisy data and must be deleted. Then, we process the obtained non-noise text data. Jieba in the Python library is used for word segmentation, and the stop word is removed to facilitate subsequent text analysis [45].

Part of the sentiment dictionary constructed in this paper comes from GooSeeker software, whose content is rich and comprehensive, including 22,215 sentiment words. The dictionary provided by the software is used as the essential sentiment dictionary. The other part is to obtain sentiment words in the financial field through statistical analysis of literature related to a stock emotional dictionary. However, the sentiment dictionary of the above two parts needs to be revised for stock sentiment analysis. Stock's comments in the stock bar come from netizens' remarks, and there are many Internet buzzwords. Therefore, this paper builds a stock-specific thesaurus based on the crawled text data to conduct a more accurate sentiment analysis. Firstly, word frequency statistics are performed on the text after word segmentation, and the top words in the frequency of the text are extracted. Then, we manually classify and assign values to the removed words. Additionally, sentiment words and particular sentiment words for the stock forum are added to the sentiment dictionary of GooSeeker software to establish a new and more accurate stock-specific sentiment dictionary. Finally, the sentiment indicator corresponding to each comment is obtained through (1).

C. CONSTRUCTION OF MS-SSA-LSTM STOCK PRICE PREDICTION MODEL

We construct a novel MS-SSA-LSTM stock price prediction model, and the model framework can be seen in Fig. 1.

The specific implementation process is shown below.

Step 1: Obtain data. Get the fundamental trading indicators of the stock and crawl the text of the East Money forum posts.

Step 2: Analyze sentiment. Build a unique sentiment dictionary in the financial field based on authoritative dictionaries. The sentiment indicator is calculated by sentiment analysis.

Step 3: Preprocess the data. The stock's historical data and emotional indicators are integrated to form a multi-source index matrix and normalize the data. Then, the training set and the test set are divided proportionally.

Step 4: Initialize parameters. One is the parameters to be optimized for LSTM, including the learning rate, iterations, and the neurons' number in hidden layers. The other is the parameters of SSA, including the sparrow's position, each parameter's upper and lower limits, and the maximum iteration number

Step 5: Optimize the LSTM model's hyperparameters. The error value obtained by training LSTM is the sparrow population's fitness value, and the objective function is to minimize the error value. Update the sparrow's location based on the objective function's result. When SSA reaches the initial set number of iterations, it will jump out of the loop and construct the LSTM model based on the optimal hyperparameters.

Step 6: Train and evaluate the model. The training model obtains the prediction results. Then, evaluate the model's prediction performance using various evaluation metrics.

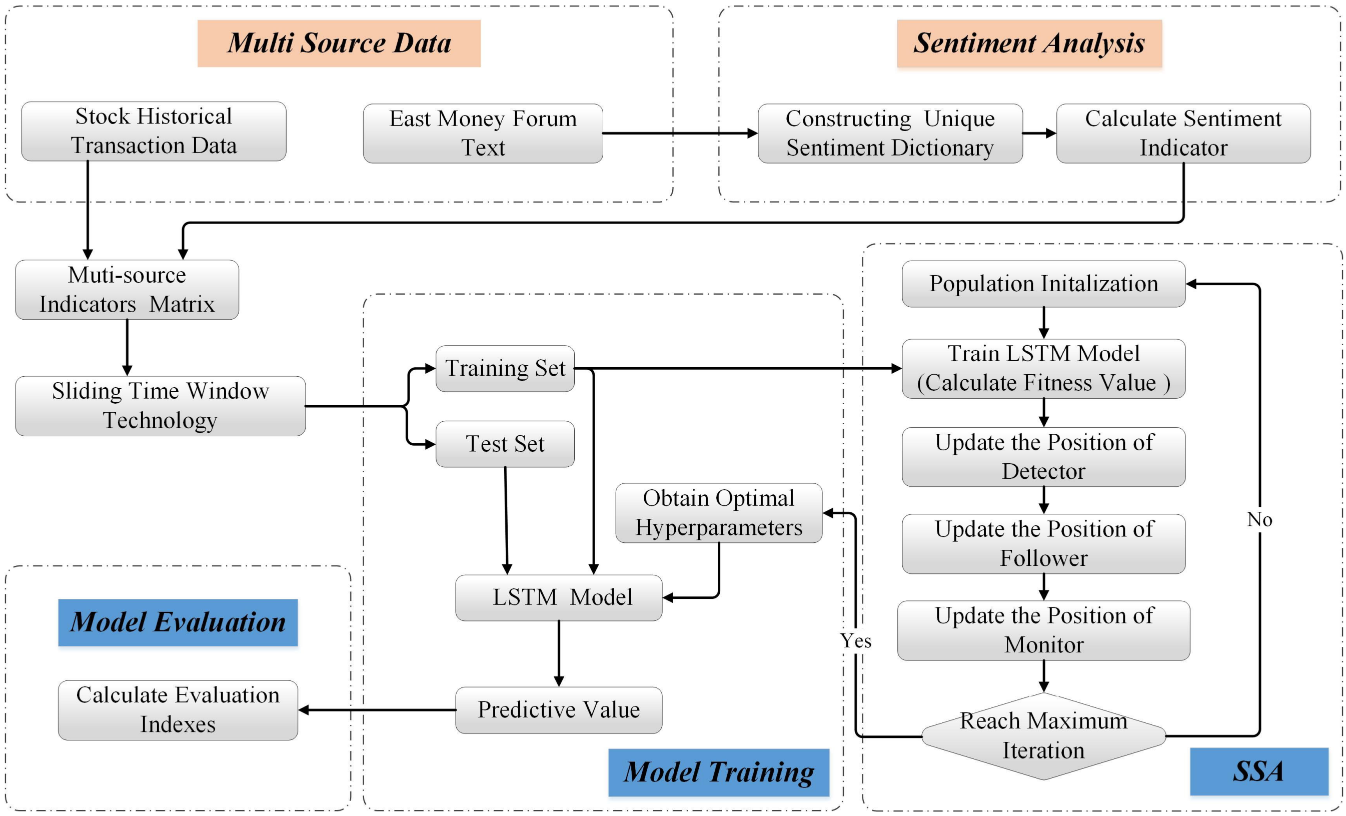

FIGURE 1. MS-SSA-LSTM model framework.

D. MODEL EVALUATION CRITERIA

Here, The model's validity is evaluated using average absolute percentage error (MAPE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The formula for calculating the evaluation criteria is listed below:

$$\mathrm{MAPE} = [\sum_{i=1}^{N} (|\hat{y}_i - y_i| / y_i)]/N \tag{17}$$

$$\mathrm{RMSE} = [\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2]^{1/2} \tag{18}$$

$$\mathrm{MAE} = (\sum_{i=1}^{N} |\hat{y}_i - y_i|)/N \tag{19}$$

$$R^2 = 1 - (\sum_{i=1}^{N} (y_i - \hat{y}_i)^2)/(\sum_{i=1}^{N} (y_i - \bar{y}_i)^2) \tag{20}$$

Here, N is the number of predicted samples, y_i is the actual price, and $\hat{y}_i$ is the predicted price. MAPE, RMSE, and MAE calculate the distinction between the actual and predicted values, whose range of values is $[0, +\infty)$. The closer to 0, the superior the model's prediction ability. R^2 measures the degree of model fitting; its entire range is [0, 1]. The closer the R^2 is to 1, the better goodness-of-fit.

V. EXPERIMENT AND RESULTS

A. DATA ACQUISITION

The prospective set of the research object comprises all stocks traded on China's A-share market. ST stocks with delisting warnings and those with less than five years on the market are excluded from the candidate set. Thus, we select six representative stocks from different industries for empirical research to verify the model's universality. Table 1 provides the six stocks' details.

TABLE 1. Description of the selected stocks.

Stock Code	Stock Abbreviation	Stock Name	Industry	Launch Date
601857	PetroChina	PetroChina Company Limited Co., Ltd.	petroleum	2007/11/5
600030	CITIC Securities	CITIC Securities Co., Ltd.	security	2003/1/6
002424	Guizhou Bailing	Guizhou Bailing Group Pharmaceutical Co., Ltd.	medicine	2010/6/3
002168	HiFuture Technology	Shenzhen HiFuture Information Technology	internet	2007/9/19
300013	Xinning Logistics	Henan Xinning Modern Logistics Co., Ltd.	storage	2009/10/30
300035	Zhongke Electric	Hunan Zhongke Electric Co.,Ltd.	electric power equipment	2009/12/25

Firstly, this paper obtains six stocks' fundamental trading data from the Ruisi Financial database from July 1, 2016, to June 30, 2022. The data set contains nine indexes, including the former closing price, opening price, closing price,

TABLE 2. Description of the selected stocks.

Date	Previous Closing Price	Opening Price	Closing Price	Highest Price	Lowest Price	Complex Price	Daily Amplitude	Trading Volume	Turnover Rate
2022-06-30	5.4	5.37	5.34	5.39	5.3	7.7133	1.6667	165082240	0.102
2022-06-29	5.34	5.36	5.39	5.41	5.34	7.7855	1.3109	164277004	0.1015
2022-06-28	5.29	5.34	5.36	5.36	5.3	7.8811	1.1342	145428839	0.0898
2022-06-27	5.36	5.37	5.31	5.4	5.3	7.8076	1.8657	170405974	0.1052
2022-06-24	5.31	5.28	5.3	5.31	5.28	7.7929	0.565	122909343	0.0759

TABLE 3. Sentiment dictionary.

Dictionary	Vocabulary
Positive	look wonderful, buy-ins, boon, increase, growth, buy up, red plate, excellent, rise suddenly and sharply, buy a little, percentage gain, success, profit, capital increase, open a position, excessive profit, bull market, take off, make a profit, hung up, pull up, latch-key kids, hit a new high, pursue high, buy in more shares, market leader, pull rose, rallied on, get on, good, pull high, fully invested, all warehouse, add some invested, rise, upward trend, climbing more than, expectability, revive, recovery, make money, excessive growth, bullish, in sight, want to buy, up-regulation, earn profit, rise up, rise in price, enormous potential, huge opportunities, upriver, lift up, get warm again, great sun, entrance, universal rise, rocket, turn into, enter the arena, proceed with, go up, rise rapidly, get involved, king, multi-win, keep rising, rise violently, turn red, hopeful, positive line, make a fortune...
Negative	rubbish, fall, collapse, liars, lost, green, throw, suffer great losses, clear out, sell, clearance sale, quilt cover, nosedive, new low, lower order, fall endlessly, give up, slump, bear market, waste, be broken, sell short, loss, down-regulation, deceptive, scourge, steep fall, deficit, kiting stocks, reduction of Shares, fraudulent, be slaughtered, plummet, slash, fall, count backwards, thunderstorm, chaos, bottom of the table, go private, very weak, arrest people, tantivy, lose money, long fall, fraud, arrest, exit, retreat market, leave, foam, refuse, sent shares, slow down, flow, dead loan, throwing dishes, weaker tendency, fell down, empty win, retreat, smashing, bearish, default, bite at a bait, step-down, line of downward trend, exotic, guy, to death, pledge, short squeeze, sell out, low, lower, below, suffer the crash, bust...

highest price, lowest price, complex price, daily amplitude, trading volume, and turnover rate. Taking PetroChina as an example, Table 2 displays the sample information.

Then, we crawl the East Money forum posts for stocks' comment information (http://guba.eastmoney.com/). We obtain 193000, 322609, 79162, 138300, 42505, and 50652 comments in the stock forum comments of the six stocks. After data preprocessing, 189878, 316926, 73690, 132049, 41209, and 49362 valid comments are retained. We manually screen out words with a high frequency related to the rise and fall of stock prices.

This paper collects sentiment words from stock forum comments and builds a sentiment dictionary, as shown in Table 3 below. Based on this sentiment dictionary, we compute the sentiment score of each statement. Then, according to (1), we obtain the sent value of each trading day.

B. ANALYSIS OF THE STOCK PRICE PREDICTION MODEL

Fig. 2 shows the closing price changes of six different stocks. As can be seen from the figures, the closing prices are non-stationary series. The first 70% of the data is the training sample set, while the remaining 30% is applied to the test sample set. The blue line represents the training set, and the orange one is the test set. We set the time sliding window, each containing historical and predicted values. The training and test sets are continuously constructed by moving the sliding window forward.

We compare the improved model with other models to validate its efficacy, as shown in Table 4. The training samples were substituted into the following six models for training.

The setting of the LSTM network's hyperparameters affects the model's prediction results. In SSA-LSTM and MS-SSA-LSTM models, the optimal hyperparameters are obtained by SSA, including learning rate, iteration number, and neuron number in hidden layers. The model's loss value decreases rapidly when the number of iterations is less than 20. When the number of iterations exceeds 20, the loss value tends to be flat, and the error curve of the test set also converges synchronously.

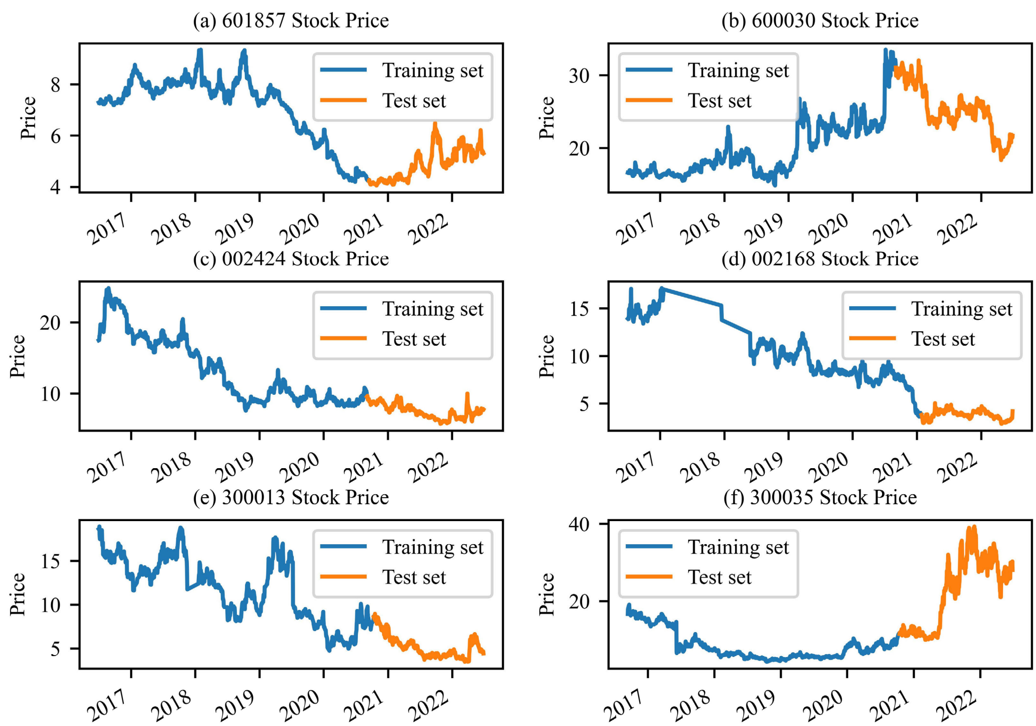

FIGURE 2. **Closing price sequence of six stocks. (a) PetroChina's closing price sequence; (b) CITIC Securities's closing price sequence; (c) Guizhou Bailing's closing price sequence; (d) HiFuture Technology's closing price sequence; (e) Xinning Logistics's closing price sequence; (f) Zhongke Electric's closing price sequence.**

TABLE 4. **Prediction model.**

Model	Detail
MLP	Input vector: fundamental stock trading indicators; Prediction using the MLP method
CNN	Input vector: fundamental stock trading indicators; Prediction using the CNN method
LSTM	Input vector: fundamental stock trading indicators; Prediction using the LSTM method
MS-LSTM	Input vector: multi-source data (stock forum sentiment indicator + fundamental stock trading indicators); Prediction using the LSTM method
SSA-LSTM	Input vector: fundamental stock trading indicators; The hyperparameter values of the LSTM method are optimized using the SSA; LSTM uses the hyperparameter values obtained by optimization to make predictions
MS-SSA-LSTM	Input vector: multi-source data (stock forum sentiment indicator + fundamental stock trading indicators); The hyperparameter values of the LSTM method are optimized using the SSA; LSTM uses the hyperparameter values obtained by optimization to make predictions

In this paper, the six models in Table 4 forecast the closing prices of the six stocks. Fig. 3 to 8 intuitively show these six stocks' predicted and actual values. In these figures, the X-axis illustrates the number of trading days in the test set, and the Y-axis indicates the stock closing price. In addition, the red line represents the actual closing price, and the blue line is the predicted one. The prediction results display that the predicted value curve of the LSTM model corresponds to the actual value of the stock price in the trend. However, there is always a particular divergence between the actual and predicted values, which has an evident lag phenomenon. From the prediction results of other models, the MS-SSA-LSTM model has a prediction curve closer to the natural stock price curve. When the stock price fluctuates sharply, the MS-SSA-LSTM model's prediction effect is significantly better than other models.

To further verify the prediction performance of this model, Table 5 reveals the assessment criterion result of each prediction model. The bold and underlined data in the table are the best experimental results among all models. Compared with MLP and CNN models, the LSTM model's predicted

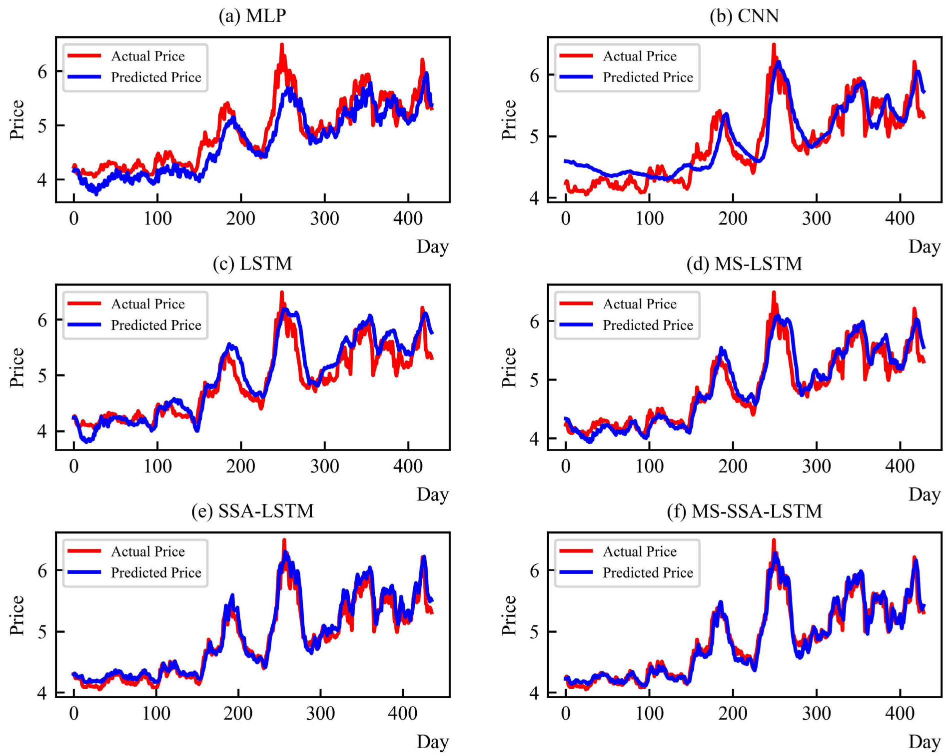

FIGURE 3. PetroChina share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.

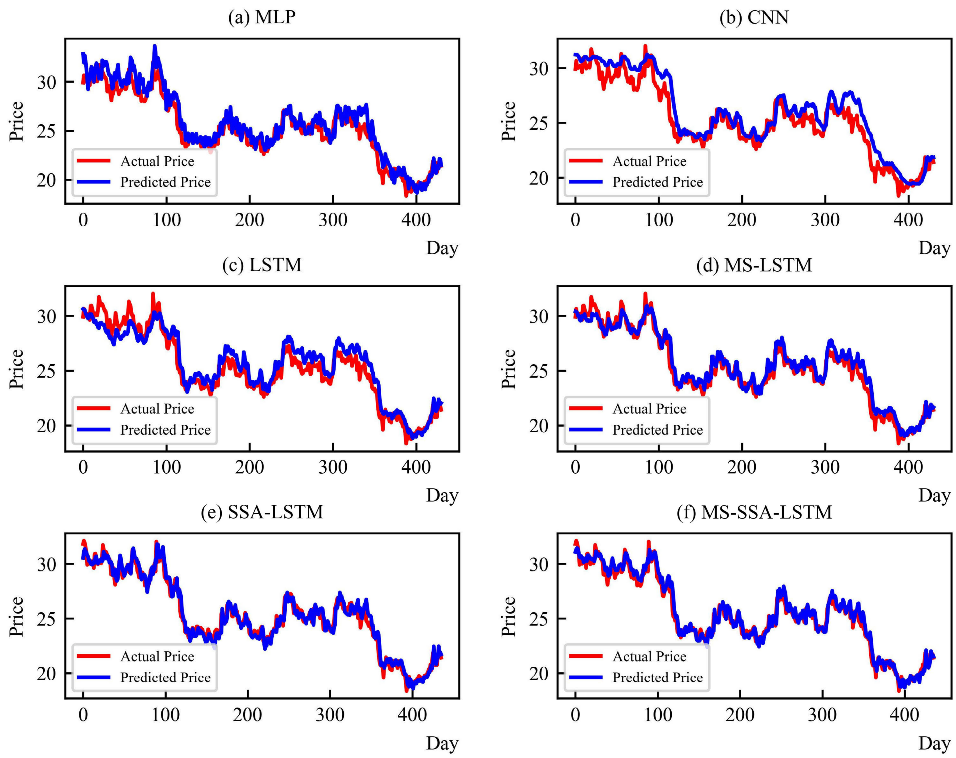

FIGURE 4. CITIC Securities share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.

value is closer to the actual value. Thus, the LSTM model is superior to MLP and CNN models. Compared with the LSTM model of a single data source, the MS-LSTM model with multi-source data has a better prediction effect. Compared

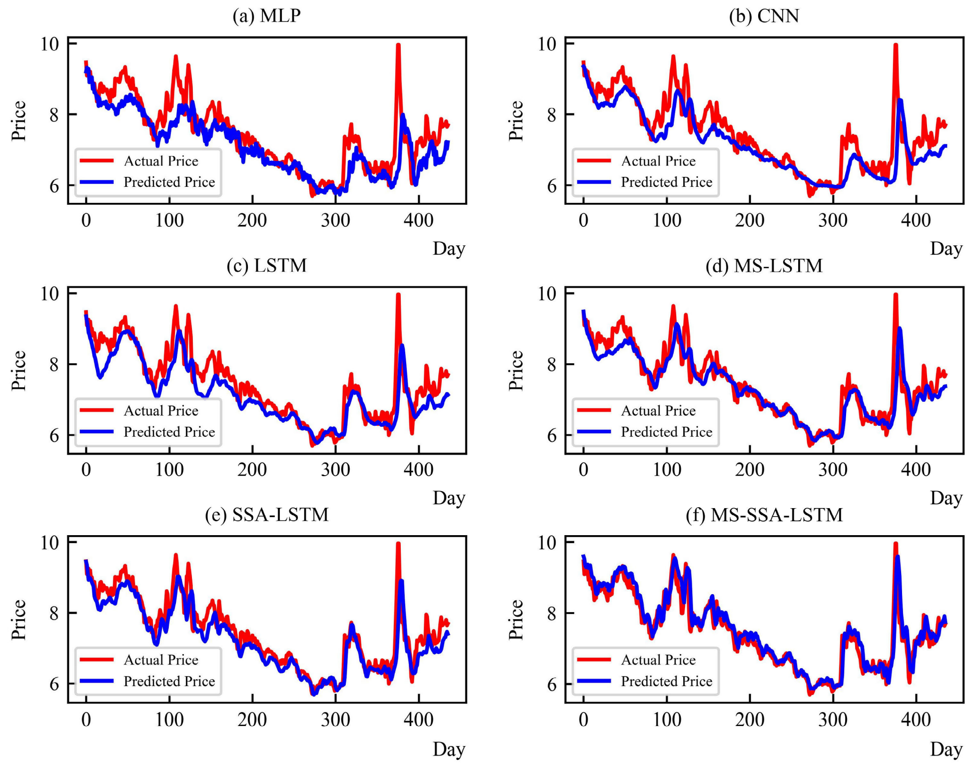

FIGURE 5. **Guizhou Bailing share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.**

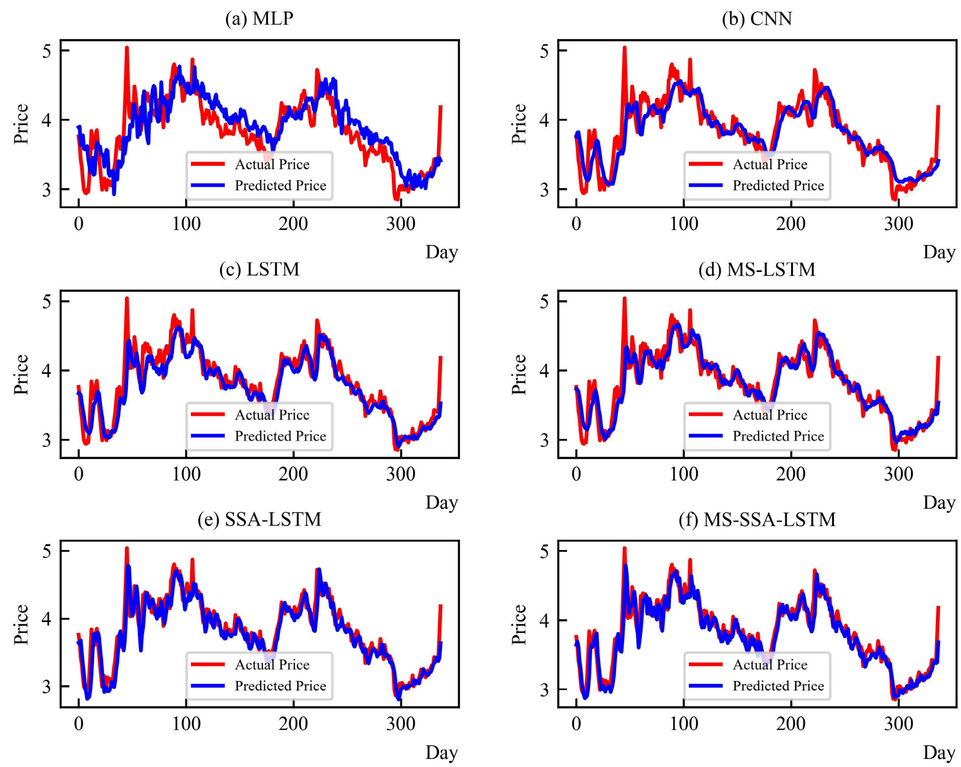

FIGURE 6. **HiFuture Technology share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.**

with the SSA-LSTM model with a single data source, the MS-SAA-LSTM model with sentiment indicators in input variables has higher prediction accuracy. The values of MAPE, MAE, and RMSE are minor, and R^2 is closer to 1. The

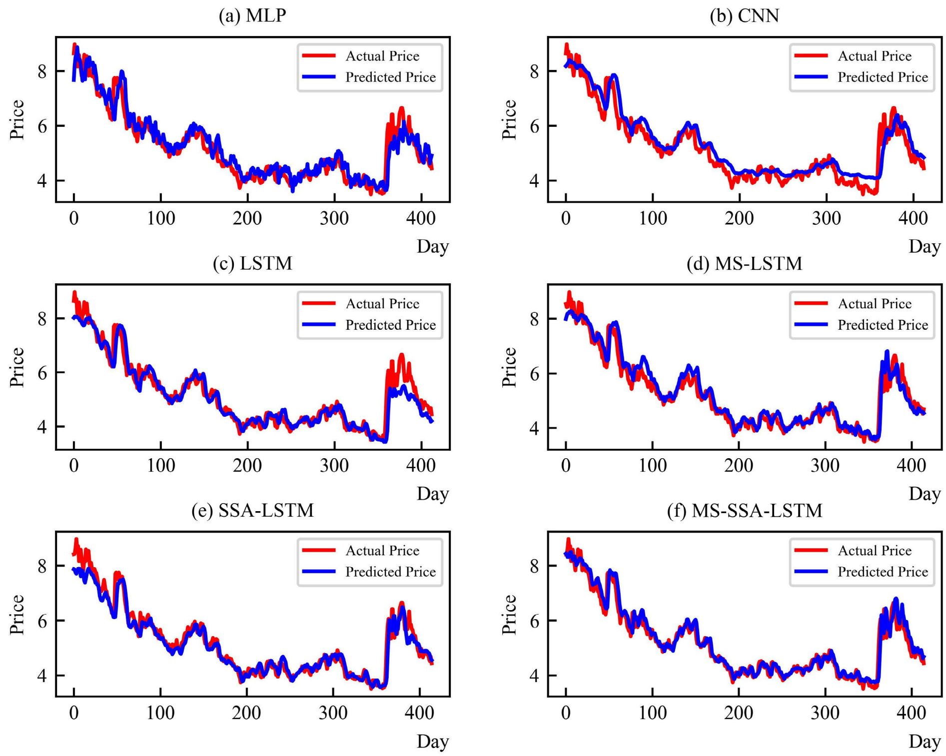

FIGURE 7. Xinning Logistics share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.

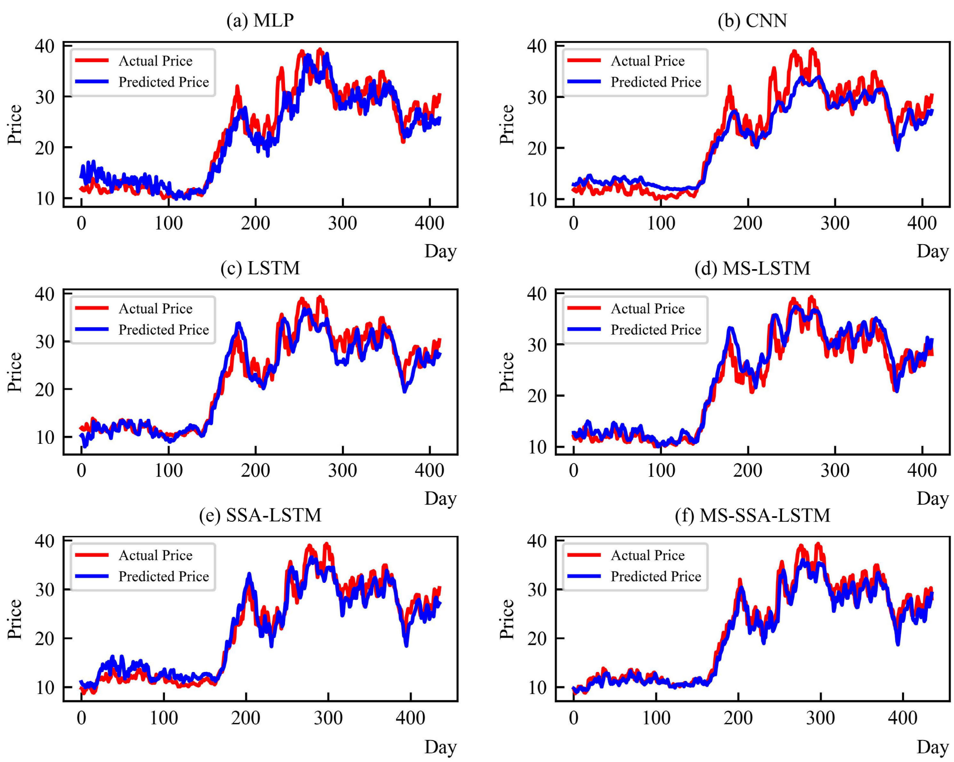

FIGURE 8. Zhongke Electric share price forecast. (a) MLP; (b) CNN; (c) LSTM; (d) MS-LSTM; (e) SSA-LSTM; (f) MS-SSA-LSTM.

experiment results indicate that the prediction outcomes are more precise when the model's input variables include a stock forum and news sentiment analysis results. At the same time, combining the emotional indicators obtained from the text information of stock reviews and news with the fundamental stock trading data is necessary.

TABLE 5. Model evaluation criteria analysis.

Model	Evaluation Criteria	PetroChina	CITIC Securities	Guizhou Bailing	HiFuture Technology	Xinning Logistics	Zhongke Electric
MLP	MAPE	0.046866	0.034079	0.053357	0.060139	0.055361	0.105973
	RMSE	0.292755	1.087889	0.599585	0.280839	0.399571	2.972397
	MAE	0.232282	0.859182	0.415673	0.224187	0.287691	2.308264
	R^2	0.753651	0.885055	0.618064	0.636735	0.896207	0.894828
CNN	MAPE	0.046411	0.036230	0.048235	0.040358	0.061223	0.099523
	RMSE	0.272681	1.160317	0.554630	0.222655	0.402674	2.721890
	MAE	0.222982	0.904479	0.375419	0.154106	0.310117	2.112973
	R^2	0.786275	0.869239	0.673189	0.771665	0.894589	0.911808
LSTM	MAPE	0.037605	0.031962	0.046887	0.033928	0.046360	0.080497
	RMSE	0.248887	0.991646	0.512350	0.190621	0.375462	2.548278
	MAE	0.185710	0.810142	0.365334	0.132328	0.249871	1.892625
	R^2	0.821947	0.904493	0.721116	0.832641	0.908354	0.922700
MS-LSTM	MAPE	0.028418	0.022524	0.036460	0.032475	0.045429	0.077738
	RMSE	0.197691	0.721064	0.433228	0.180148	0.333892	2.242278
	MAE	0.141268	0.557175	0.283547	0.125234	0.242056	1.702037
	R^2	0.887664	0.949502	0.800601	0.850525	0.930593	0.940003
SSA-LSTM	MAPE	0.020661	0.018782	0.036605	0.031119	0.038470	0.089932
	RMSE	0.134624	0.608499	0.404247	0.167820	0.304169	2.172967
	MAE	0.103214	0.469336	0.283440	0.120174	0.210899	1.743758
	R^2	0.946517	0.965252	0.826387	0.870283	0.941478	0.946007
MS-SSA-LSTM	MAPE	**0.018216**	**0.017453**	**0.029119**	**0.029589**	**0.035686**	**0.058062**
	RMSE	**0.123077**	**0.568586**	**0.320423**	**0.157564**	**0.274097**	**1.869229**
	MAE	**0.091298**	**0.438209**	**0.220938**	**0.114505**	**0.189724**	**1.365726**
	R^2	**0.956459**	**0.969661**	**0.890922**	**0.885653**	**0.952478**	**0.960072**

TABLE 6. Prediction effect of the MS-SSA-LSTM model with different time steps.

Model	Evaluation Criteria	PetroChina	CITIC Securities	Guizhou Bailing	HiFuture Technology	Xinning Logistics	Zhongke Electric
MS-SSA-LSTM(5)	MAPE	**0.016019**	**0.015339**	**0.019017**	0.027517	**0.029746**	**0.038965**
	RMSE	**0.115017**	**0.510695**	**0.224353**	0.146532	**0.213963**	**1.274311**
	MAE	**0.081445**	**0.384959**	**0.144908**	0.106437	**0.155798**	**0.895620**
	R^2	**0.961975**	**0.975525**	**0.946525**	0.901106	**0.971043**	**0.981443**
MS-SSA-LSTM(10)	MAPE	0.016341	0.024942	0.026477	**0.027209**	0.033208	0.042126
	RMSE	0.117011	0.789871	0.286236	**0.146284**	0.258109	1.344521
	MAE	0.083021	0.647192	0.202313	**0.104895**	0.177852	0.957299
	R^2	0.960645	0.941451	0.912956	**0.901439**	0.957861	0.979342
MS-SSA-LSTM(20)	MAPE	0.017581	0.020590	0.026992	0.028497	0.299261	0.066385
	RMSE	0.122042	0.669828	0.336038	0.153892	0.225044	1.942409
	MAE	0.088456	0.524833	0.208484	0.110745	0.156675	1.426487
	R^2	0.957188	0.957895	0.880032	0.890921	0.967966	0.956884

In addition, to verify the SSA's efficacy in optimizing the LSTM model parameters, this paper compares the model optimized by SSA with the model of artificially set parameters. The results show that the SSA-LSTM model outperforms the LSTM model regarding prediction effect and closeness to the actual value and has a minor prediction error. The MS-SSA-LSTM model surpassed the MS-LSTM model in prediction ability and impact. The experimental outcomes show that SSA has a powerful ability to train parameters, which can avoid the over-fitting of model training and prevents the gradient from disappearing or exploding. Therefore, SSA improved the stock price forecast model's performance.

From the predictions of the latter four models, the fitting degree of the models is the MS-SSA-LSTM, SSA-LSTM, MS-LSTM, and LSTM prediction model in order from high to low. The prediction and evaluation criteria obtained by the MS-SSA-LSTM model on six stock data sets are smaller than other models, and R^2 values are closer to 1. Taking PetroChina as an example, in the MS-SSA-LSTM model, the value of MAPE, RMSE, MAE, and R^2 reaches 0.018216, 0.123077, 0.091298, and 0.956459, respectively. The six stocks' price predicted by the MS-SSA-LSTM model is 10.74% greater than the standard LSTM model's average. Generally speaking, the proposed MS-SSA-LSTM model is universal and robust in forecasting stock prices.

C. INFLUENCE OF TRADING DAY ON THE STOCK PRICE PREDICTION

In order to evaluate the result of the time step size on the MS-SSA-LSTM model's prediction effect, three different time steps were chosen, namely 5, 10, and 20. Table 6 depicts the experimental results corresponding to various time steps. The best-predicted results are identified in bold and underlined.

First, we analyze HiFuture Technology. When the time step is 10, the stock price trend can be better predicted. When the time step is 20, the forecast accuracy is lower, even worse than the previous five trading days. In addition, some trading days are too far away from the forecast day, and the data does not help but negatively interfere with the stock price forecast effect on the forecast data.

Then, we analyze the results of another five stocks. Setting the time step to 5 has the best result in PetroChina, CITIC Securities, Guizhou Bailing, Xinning Logistics, and Zhongke Electric. If the time step is long enough, the useless data will influence the model's training and significantly reduce the training efficiency. If the time frame is exceptionally long, relatively irrelevant data will affect the model's training.

Overall, the prices of six stocks change rapidly, and using data over more extended periods cannot improve the prediction effect of stock prices.

VI. CONCLUSION AND FUTURE RESEARCH

The stock price is affected by shareholder emotion, and the hyperparameters in the LSTM network are frequently chosen based on subjective experience. Therefore, this paper proposes the MS-SSA-LSTM model of stock price prediction. Six representative data sets of individual stocks in the Chinese financial market are selected to train and test the model. Moreover, four assessment indicators are employed to check the model's prediction performance. Through comparative analysis, we can draw the following results.

First of all, the MS-SSA-LSTM model considers multiple data sources. On the one side, the data source is the characteristics of the historical transaction data, including the previous closing price, opening price, closing price, etc. On the other side, the data source is the sentiment of shareholders in the market. Adding a sentiment indicator enhances the model's predictive performance compared with only using fundamental stock trading indicators as input. Thus, the stock bar can be used as a guiding platform for investor sentiment. The platform managers can conduct public opinion management and early warning, reasonably guide shareholders to invest rationally and take countermeasures in advance for possible panic or riot. Ultimately, we should establish a good and orderly stock market environment.

Secondly, the parameters of LSTM need to be adjusted artificially, which challenges obtaining the best prediction results. Therefore, the SSA is chosen to optimize the LSTM model's hyperparameters. This approach not only objectively explains the network structure and parameter setting of the model but also improves the model's adaptability and forecasting capabilities. In a challenging financial market, the LSTM model optimized by SSA can quickly and precisely comprehend data properties, provide high-precision price prediction, and decrease investors' risk.

Thirdly, the comparative experiments of six individual stocks in different models verify that the MS-SSA-LSTM model has high accuracy, reliability, and adaptability to the stock market. In the experiment, 5-10 time steps can make the prediction effect of the model reach optimal, indicating that the enormous volatility of China's financial market is more suitable for short-term prediction. Meanwhile, this model can be applied to other time series problems.

Finally, the MS-SSA-LSTM model is universal. Although we tested the model only on China's financial market, it is also suitable for foreign stock markets. The input values to the model are multi-source data, one is the technical stock data, and the other is the sentiment index of shareholders. These multi-source data exist in both Chinese and foreign financial markets. Social media platforms for discussing stock prices are diversified and open. As long as we can dig out the comments on the stock, we can calculate the sentiment index and use it as one of the model's input features. In addition, establishing a comment platform for the stock market helps us collect comments easily and quickly.

In the MS-SSA-LSTM model, our multi-source data still have limitations. Regarding sentimental analysis, this paper only divides emotions into positive and negative. In addition, other variables, such as macroeconomic conditions and policy shifts, will also impact the stock price forecast. In the future, sentimental analysis needs to be further refined. We should extract different emotional indicators into our research, such as sadness, fear, anger, and disgust. Meanwhile, more data sources, such as Weibo and WeChat official accounts, may be introduced to estimate market sentiments. In addition, we should use mining techniques to find other factors that predict stock prices.

REFERENCES

[1] M. M. Rounaghi and F. N. Zadeh, "Investigation of market efficiency and financial stability between S&P 500 and London stock exchange: Monthly and yearly forecasting of time series stock returns using ARMA model," *Phys. A, Stat. Mech. Appl.*, vol. 456, pp. 10–21, Aug. 2016, doi: 10.1016/j.physa.2016.03.006.

[2] G. Bandyopadhyay, "Gold price forecasting using ARIMA model," *J. Adv. Manage. Sci.*, vol. 4, no. 2, pp. 117–121, 2016, doi: 10.12720/joams.4.2.117-121.

[3] H. Shi, Z. You, and Z. Chen, "Analysis and prediction of Shanghai composite index by ARIMA model based on wavelet analysis," *J. Math. Pract. Theory*, vol. 44, no. 23, pp. 66–72, 2014.

[4] H. Herwartz, "Stock return prediction under GARCH—An empirical assessment," *Int. J. Forecasting*, vol. 33, no. 3, pp. 569–580, Jul. 2017, doi: 10.1016/j.ijforecast.2017.01.002.

[5] H. Mohammadi and L. Su, "International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models," *Energy Econ.*, vol. 32, no. 5, pp. 1001–1008, Sep. 2010, doi: 10.1016/j.eneco.2010.04.009.

[6] A. Hossain and M. Nasser, "Recurrent support and relevance vector machines based model with application to forecasting volatility of financial returns," *J. Intell. Learn. Syst. Appl.*, vol. 3, no. 4, pp. 230–241, 2011, doi: 10.4236/jilsa.2011.34026.

[7] J. Chai, J. Du, K. K. Lai, and Y. P. Lee, "A hybrid least square support vector machine model with parameters optimization for stock forecasting," *Math. Problems Eng.*, vol. 2015, pp. 1–7, Jan. 2015, doi: 10.1155/2015/231394.

[8] A. Murkute and T. Sarode, "Forecasting market price of stock using artificial neural network," *Int. J. Comput. Appl.*, vol. 124, no. 12, pp. 11–15, Aug. 2015, doi: 10.5120/ijca2015905681.

[9] D. Banjade, "Forecasting Bitcoin price using artificial neural network," Jan. 2020, doi: 10.2139/ssrn.3515702.

[10] J. Zahedi and M. M. Rounaghi, "Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran stock exchange," *Phys. A, Stat. Mech. Appl.*, vol. 438, pp. 178–187, Nov. 2015, doi: 10.1016/j.physa.2015.06.033.

[11] A. H. Moghaddam, M. H. Moghaddam, and M. Esfandyari, "Stock market index prediction using artificial neural network," *J. Econ., Finance Administ. Sci.*, vol. 21, no. 41, pp. 89–93, Dec. 2016, doi: 10.1016/j.jefas.2016.07.002.

[12] H. Liu and Y. Hou, "Application of Bayesian neural network in prediction of stock time series," *Comput. Eng. Appl.*, vol. 55, no. 12, pp. 225–229, 2019.

[13] A. M. Rather, A. Agarwal, and V. N. Sastry, "Recurrent neural network and a hybrid model for prediction of stock returns," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3234–3241, Apr. 2015, doi: 10.1016/j.eswa.2014.12.003.

[14] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306, doi: 10.1016/j.physd.2019.132306.

[15] G. Ding and L. Qin, "Study on the prediction of stock price based on the associated network model of LSTM," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 6, pp. 1307–1317, Nov. 2019, doi: 10.1007/s13042-019-01041-1.

[16] X. Yan, W. Weihan, and M. Chang, "Research on financial assets transaction prediction model based on LSTM neural network," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 257–270, May 2020, doi: 10.1007/s00521-020-04992-7.

[17] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and S. Shahab, "Deep learning for stock market prediction," *Entropy*, vol. 22, no. 8, p. 840, Jul. 2020, doi: 10.3390/e22080840.

[18] Z. D. Aksehir and E. Kiliç, "How to handle data imbalance and feature selection problems in CNN-based stock price forecasting," *IEEE Access*, vol. 10, pp. 31297–31305, 2022, doi: 10.1109/ACCESS.2022.3160797.

[19] Y. Ji, A. W. Liew, and L. Yang, "A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast," *IEEE Access*, vol. 9, pp. 23660–23671, 2021, doi: 10.1109/ACCESS.2021.3056713.

[20] X. Zeng, J. Cai, C. Liang, and C. Yuan, "A hybrid model integrating long short-term memory with adaptive genetic algorithm based on individual ranking for stock index prediction," *PLoS ONE*, vol. 17, no. 8, Aug. 2022, Art. no. e0272637, doi: 10.1371/journal.pone.0272637.

[21] J. Xue and B. Shen, "A novel swarm intelligence optimization approach: Sparrow search algorithm," *Syst. Sci. Control Eng.*, vol. 8, no. 1, pp. 22–34, Jan. 2020, doi: 10.1080/21642583.2019.1708830.

[22] J. Borade, "Stock prediction and simulation of trade using support vector regression," *Int. J. Res. Eng. Technol.*, vol. 7, no. 4, pp. 52–57, Apr. 2018, doi: 10.15623/ijret.2018.0704009.

[23] X. Li and P. Tang, "Stock price prediction based on technical analysis, fundamental analysis and deep learning," *Stat. Decis.*, vol. 38, no. 2, pp. 146–150, 2022, doi: 10.13546/j.cnki.tjyjc.2022.02.029.

[24] J. Heo and J. Y. Yang, "Stock price prediction based on financial statements using SVM," *Int. J. Hybrid Inf. Technol.*, vol. 9, no. 2, pp. 57–66, Feb. 2016, doi: 10.14257/ijhit.2016.9.2.05.

[25] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise trader risk in financial markets," *J. Political Economy*, vol. 98, no. 4, pp. 703–738, Aug. 1990, doi: 10.1086/261703.

[26] H. Cui and Y. Zhang, "Does investor sentiment affect stock price crash risk?" *Appl. Econ. Lett.*, vol. 27, no. 7, pp. 564–568, Jul. 2019, doi: 10.1080/13504851.2019.1643448.

[27] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decis. Support Syst.*, vol. 53, no. 3, pp. 458–464, Jun. 2012, doi: 10.1016/j.dss.2012.03.001.

[28] M. Nofer and O. Hinz, "Using Twitter to predict the stock market," *Bus. Inf. Syst. Eng.*, vol. 57, no. 4, pp. 229–242, Jun. 2015, doi: 10.1007/s12599-015-0390-4.

[29] P. Fan, Y. Yang, Z. Zhang, and M. Chen, "The relationship between individual stock investor sentiment and stock yield-based on the perspective of stock evaluation information," *Math. Pract. Theory*, vol. 51, no. 16, pp. 305–320, 2021.

[30] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Comput. Appl.*, vol. 32, no. 13, pp. 9713–9729, Sep. 2019, doi: 10.1007/s00521-019-04504-2.

[31] X. Xu and K. Tian, "A novel financial text sentiment analysis-based approach for stock index prediction," *J. Quantum Technol. Econ.*, vol. 38, no. 12, pp. 124–145, 2021, doi: 10.13653/j.cnki.jqte.2021.12.009.

[32] C.-R. Ko and H.-T. Chang, "LSTM-based sentiment analysis for stock price forecast," *PeerJ Comput. Sci.*, vol. 7, p. e408, Mar. 2021, doi: 10.7717/peerj-cs.408.

[33] Y. Li and Y. Pan, "A novel ensemble deep learning model for stock prediction based on stock prices and news," *Int. J. Data Sci. Anal.*, vol. 13, no. 2, pp. 139–149, Sep. 2021, doi: 10.1007/s41060-021-00279-9.

[34] C. Kearney and S. Liu, "Textual sentiment in finance: A survey of methods and models," *Int. Rev. Financial Anal.*, vol. 33, pp. 171–185, May 2014, doi: 10.1016/j.irfa.2014.02.006.

[35] T. Wang and Z. Zhang, "Research on the construction method of emotional lexicon for movie review," *Comput. Digit. Eng.*, vol. 50, no. 4, pp. 843–848, 2022, doi: 10.3969/j.issn.1672-9722.2022.04.031.

[36] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, vol. 17, no. 4, pp. 723–742, Jun. 2013, doi: 10.1007/s11280-013-0221-9.

[37] L. Yang and T. Zhai, "Research on sentiment tendency analysis of video reviews based on sentiment dictionary," *Netw. Secur. Technol. Appl.*, vol. 255, no. 3, pp. 53–56, 2022.

[38] A. Fathy, T. M. Alanazi, H. Rezk, and D. Yousri, "Optimal energy management of micro-grid using sparrow search algorithm," *Energy Rep.*, vol. 8, pp. 758–773, Nov. 2022, doi: 10.1016/j.egyr.2021.12.022.

[39] Y. Chen, Z. Liu, C. Xu, X. Zhao, L. Pang, K. Li, and Y. Shi, "Heavy metal content prediction based on random forest and sparrow search algorithm," *J. Chemometrics*, vol. 36, no. 10, Sep. 2022, Art. no. e3445, doi: 10.1002/cem.3445.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[41] V. Gupta and M. Ahmad, "Stock price trend prediction with long short-term memory neural networks," *Int. J. Comput. Intell. Stud.*, vol. 8, no. 4, p. 289, 2019, doi: 10.1504/ijcistudies.2019.10025266.

[42] V. R. Madireddy, "Stock market prediction in BSE using long short-term memory (LSTM) algorithm," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 6, no. 1, pp. 561–565, Jan. 2018, doi: 10.2139/ssrn.4289422.

[43] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep learning with long short-term memory for time series prediction," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 114–119, Jun. 2019, doi: 10.1109/MCOM.2019.1800155.

[44] T. Matsubara, R. Akita, and K. Uehara, "Stock price prediction by deep neural generative model of news articles," *IEICE Trans. Inf. Syst.*, vol. E101.D, no. 4, pp. 901–908, 2018, doi: 10.1587/transinf.2016iip0016.

[45] G. Song, D. Cheng, S. Zhang, W. Liu, and X. Ding, "A model of textual emotion score calculation based on the emotion dictionary," *China Comput. Commun.*, vol. 33, no. 22, pp. 56–58 and 62, 2021.

GUANGYU MU received the bachelor's degree in computer science from Jilin Engineering University, in 1994, and the master's and Ph.D. degrees in computer science from Jilin University, in 2004 and 2011, respectively. She is currently a Professor, a Faculty Member with the School of Management Science and Information Engineering, and the Director of the Institute of Economic Information Management, Jilin University of Finance and Economics. She is a Peer Reviewer of the China National Social Science Fund Projects, an Expert on Humanities and Social Science Fund Projects of the Ministry of Education of China, and an Expert on the Outstanding Contribution of Jilin Province. Her research interests include information resource management, data mining, and financial engineering.

NAN GAO received the degree in engineering, in 2021. She is currently pursuing the master's degree in management science and engineering with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. Her current research interests include financial market forecasting and stock portfolio. Her awards and honors include the National Inspirational Scholarship, the First-Class Scholarship, the title of "Three Good Students," and the National Certificate of Mathematical Modeling.

YUHAN WANG received the degree in engineering, in 2019. She is currently pursuing the master's degree in management science and engineering with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. Her current research interest includes financial information engineering. Her awards include the Second-Class Scholarship.

LI DAI received the degree in engineering, in 2021. She is currently pursuing the master's degree in management science and engineering with the School of Management Science and Information Engineering, Jilin University of Finance and Economics. Her current research interests include financial market forecasting and green finance. She received the Third-Class Scholarship from the School of Management Science and Information Engineering, Jilin University of Finance and Economics.