ORIGINAL ARTICLE

Stock closing price prediction based on sentiment analysis and LSTM

Zhigang Jin[1] · **Yang Yang**[1] · **Yuhong Liu**[2]

Abstract

Stock market prediction has been identified as a very important practical problem in the economic field. However, the timely prediction of the market is generally regarded as one of the most challenging problems due to the stock market's characteristics of noise and volatility. To address these challenges, we propose a deep learning-based stock market prediction model that considers investors' emotional tendency. First, we propose to involve investors' sentiment for stock prediction, which can effectively improve the model prediction accuracy. Second, the stock pricing sequence is a complex time sequence with different scales of fluctuations, making the accurate prediction very challenging. We propose to gradually decompose the complex sequence of stock price by adopting empirical modal decomposition (EMD), which yields better prediction accuracy. Third, we adopt LSTM due to its advantages of analyzing relationships among time-series data through its memory function. We further revised it by adopting attention mechanism to focus more on the more critical information. Experiment results show that the revised LSTM model can not only improve prediction accuracy, but also reduce time delay. It is confirmed that investors' emotional tendency is effective to improve the predicted results; the introduction of EMD can improve the predictability of inventory sequences; and the attention mechanism can help LSTM to efficiently extract specific information and current mission objectives from the information ocean.

Keywords Stock market prediction · Long short-term memory · Attention mechanism · Empirical mode decomposition

1 Introduction

The stock market is a place where stocks can be transferred, traded and circulated. It has been around for 400 years and has become an important channel for large companies to raise funds from investors. On the one hand, through the issuance of stocks, a large amount of capital flows into the stock market, which enhances the organic composition of corporate capital by promoting capital concentration, and greatly promotes the development of commodity economy. On the other hand, through the circulation of stocks, funds are pooled and the accumulation of capital is effectively promoted. Therefore, the stock market is considered as a barometer of the economic and financial activities in a country or region. In particular, the trading price of the stock market often serves as an indicator for the price and quantity of the stock as it can objectively reflect the supply and demand relationship of the stock market.

However, the formation mechanism of stock prices is quite complicated. The combined use of various factors and the special behavior of individual factors, including political, economic and market factors as well as technology and investor behavior, will all lead to changes in stock prices. As a result, stock prices are constantly changing, and this change provides a living space for speculative activities and increases the risk of the stock market. This kind of risk not only may bring economic losses to investors, but also may bring certain side effects to the economic construction of the enterprises and countries.

Accurately analyzing and predicting stock prices in a timely way are critical to investor choices and national

✉ Yuhong Liu
   yhliu@scu.edu

   Zhigang Jin
   zgjin@tju.edu.cn

   Yang Yang
   2018234158@tju.edu.cn

1   School of Electrical and Information Engineering, Tianjin University, Tianjin, People's Republic of China

2   Department of Computer Science and Engineering, Santa Clara University, Santa Clara, USA

economic stability. The analysis of the stock market, which includes the collection, sortation and integration of various relevant information, will help to understand and predict the trend of stock prices and to make corresponding investment decisions to reduce risks and obtain higher benefits [1]. If we can predict the rise and fall of stocks in an accurate and timely way, corresponding regulation and healthy guidance of the stock market will provide a solid support for the sustainable development of the economy. Therefore, the study of stock forecasts can guide investors to make beneficial investments, not only to provide profits for individuals, but also to contribute to the development of the national economy.

However, stock market prediction is generally regarded as one of the most challenging problems in time-series prediction due to its characteristics of noise and volatility [2]. How to accurately predict the stock movement is still an open question in modern social economy and social organization. Many related studies have emerged, as a research hot spot in economic science, coupled with the concerns of investors and the attraction of high returns. The conventional method of stock market prediction mainly focuses on time-series analysis. De Gooijer et al. [3] have reviewed the papers published at journals managed by the International Institute of Forecasters (Journal of Forecasting 1982–1985; International Journal of Forecasting 1985–2005) and found that over one-third of all the papers published at these journals focused on time-series forecasting. In particular, conventional time-series analysis methods include autoregressive model (AR) [4], moving average model (MA) [5], autoregressive and moving average model (ARMA) [6] and autoregressive integrate moving average model(ARIMA) [7]. All these approaches mainly focus on the time series itself, while ignoring other influencing factors such as the context information. Specifically, they assume the previous data and the later data as independent and dependent variables, respectively, aiming to obtain the quantitative relationship between them. Moreover, these methods often require some assumptions and pre-knowledge, such as the underlying data distribution, valid ranges for various parameters and their connections. However, as a complex system with many influential factors and uncertainties, the stock market tends to exhibit strong nonlinear characteristics, which makes the conventional analytical methods ineffective. In addition, the amount of information processed by the modeling and forecasting of the stock market is often very large, raising great challenges for the algorithm design. These characteristics make the prediction of the stock market based on conventional methods inadequate.

Recently, financial field has been widely using machine learning models in time-series analysis. Support vector regression (SVR) [8] and artificial neural networks (ANNs) [9] both gained considerable results [10]. In addition, deep learning becomes a new trend of machine learning, due to its excellent ability to map nonlinear relationships and adopt limited background knowledge. Deep learning has powerful data processing capabilities that can solve the problems caused by the complexity of financial time series. Therefore, the combination of deep learning and finance has very broad prospects [11], but the work in this area is not enough.

In addition, human irrational behavior makes the stock market not completely objective and does not necessarily conform to scientific principles. Their emotional psychological and behavioral characteristics are very important in the economic system. Recent researches also showed that the investors' sentiment may take an important role in stock market investment. For example, Antweiler and Frank [12] through experiments on the content of the e-mail and the Dow Jones index found that the content of the e-mail help predict stock market changes and confirmed the correlation between online reviews and stock trading volume. In addition, reference [13] also emphasized the important role of emotion in investors' decision making. Although there are many studies showing the strong correlations between sentiment and stock price, few works have considered sentiment analysis for stock prediction, which is exactly what we focus on in this work.

In this work, we propose to predict the stock market closing prices by adopting investors' sentiment, empirical modal decomposition (EMD) and a revised long short-term memory (LSTM) model with the attention mechanism. The major contributions of this paper are summarized as follows.

Firstly, we propose to involve investors' sentiment for stock prediction, which can effectively improve the model prediction accuracy. Specifically, we calculate the sentiment index through sentiment analysis on a large number of stock market comments, which are classified into either bullish or bearish. These comments reflect the overall investors' sentiment orientation. Inspired by this idea, we adopt sentiment index, which directly reflects the behavior of shareholders, together with the historical data of stock price as the input of the stock market prediction model.

Secondly, we propose to adopt EMD to extract the trend term of the stock price sequence. Generally, simple sequences are more predictable than complex sequences. EMD is a common method to process time series and has been applied in various fields [14]. The main purpose is to gradually decompose the fluctuations of different scales of a stationary or non-stationary complex signal to obtain a series of intrinsic mode functions (IMF) with better performance and a single residual term [15]. Henceforth, we decompose the complex sequence of stock price and

predict the decomposed simple sequences to obtain better prediction effect.

Thirdly, we propose to apply LSTM to predict the stock closing prices, which could be influenced by various influential factors and show high uncertainty and nonlinear characteristics. Although some scholars have used deep learning for financial field, most of them used it for classification tasks, and very few works have used it for prediction of specific values in financial time series. In addition, among different commonly used deep learning models, such as convolutional neural networks (CNN), deep belief networks and LSTM, we chose LSTM due to its advantages of analyzing relationships among time-series data through its memory function.

Last, we propose to improve the LSTM with attention mechanism. Few efforts have been made to investigate how to use LSTM to effectively extract meaningful information from noisy financial time-series data. Therefore, this paper aims to contribute to this area. In particular, we combine LSTM with attention mechanism and propose a stock market prediction model, which can focus more on more critical information. With the improved LSTM model, we can not only improve prediction accuracy, but also reduce time delay.

In summary, a stock market prediction model, which adopts sentiment analysis, EMD and LSTM with attention mechanism, is proposed in this paper. The rest of the paper is organized as follows. Section 2 reviews the related work; Sect. 3 presents the proposed scheme; Sect. 4 discusses the experiments and results, followed by conclusion in Sect. 5.

2 Related work

In this section, we provide a detailed discussion on prior relevant literature on stock market prediction.

2.1 Conventional time-series analysis

There have been many attempts to predict the stock market through conventional time-series analysis methods. Tang [16] studied the application of the ARMA-GARCH model extended by the AR-GARCH model in stock price prediction. In [17], the author constructed an autoregressive dynamic Bayesian network (AR-DBN) based on dynamic Bayesian network (DBN) [18] and inferred the market index, which improves the predictability of stock market volatility. The authors in [19] put the conventional ARMA model with SVMs, which was combined to give full play to the advantages of both to conduct stock market prediction, providing a model with better explanatory power. However, most of the conventional time-series analysis studies relied on the linear relationship between stock prices and were more suitable for sequences with stable trends and laws, which made them inadequate to handle more complex nonlinear relationships. Moreover, the stock market has many influential factors and the impact is complex, which is ignored in simple time-series analysis methods and makes the prediction less effective.

2.2 Long short-term memory model

LSTM is a long short-term memory network, a kind of time loop neural network, which is suitable for processing and predicting important events with relatively long interval and delay in time series. In this section, we mainly discuss works on the development of LSTM and its application in time-series prediction.

In the field of deep learning, the conventional feed-forward neural network represented by CNN has excellent performance in solving classification tasks, but cannot handle the complex time correlation between information. RNN introduces a directional loop, where the output of the neuron can be directly applied to itself at the next timestamp. This directional loop enables RNN to handle the problem of before and after the input. In order to solve the long-term dependence in neural networks and the disappearance and outburst of the conventional RNN model gradients, in 1997 Hochreiter and Schmidhuber [20] proposed LSTM, an RNN architecture that better stores and accesses information and serves as the basis for other models.

There has been a lot of work to prove that LSTM can achieve better results in time-series prediction. Shi et al. [21] used the LSTM improved by convolution for the time-series prediction problem of precipitation, and it is proved that the algorithm is superior to other existing precipitation prediction models. Ma [22] used LSTM to capture nonlinear traffic dynamics for short-term traffic prediction, achieving the best predictive performance in terms of accuracy and stability. Liu [23] used LSTM to predict wind speed. Ding et al. [24] constructed a deep convolutional neural network to study the relationship between the occurrence of events and stock prices over different time horizons. Fischer [25] confirmed that LSTM is more suitable for time-series analysis than the no-memory classification method, such as random forest (RAF), deep neural network (DNN) and logistic regression classifier (LOG).

The existing literature inspires our work to adopt LSTM to predict stock market. However, the specific characteristics of the stock market, such as nonlinear, unstable and complicated influencing factors, cannot be well handled by existing literature. Therefore, considering that EMD can decompose complex sequences into simple sequences, and LSTM has a powerful ability to process nonlinear sequences, in this work, we propose to revise the LSTM

model by adopting attention mechanism that can help to focus more on information which relates more with the sequences to be predicted. In addition, we propose to adopt EMD to extract the trend term of the complex stock pricing sequence by decomposing it into more predictable simple sequences.

2.3 Sentiment analysis

Although it is generally accepted that stock market prices are largely driven by new information and follow a random pattern, many studies have shown the effectiveness of predicting stock market behavior based on behavioral economics that emphasizes the important role of emotion in decision making [26, 27]. With the maturing of the behavioral finance, the influence of investors' irrational factors on stock market has attracted more attention. For example, the frustration caused by making mistakes in decision making and the personality traits in the investment process will all have an impact on stock movements.

As social networks become more and more important to people, there is growing evidence that investors are not completely rational, and the interactions among investors in stock market get more and more convenient and frequent. Therefore, the sentiment expressed by other investors and the opinions expressed by the social network may affect the investor's mood and change their decision making, and further influence the stock market to a certain extent. For example, Kaminski and Globe [28] found that there is a positive correlation between the number of investors' bullish comments and the stock closing price. Baker and Wurgler [29, 30] constructed a sentiment index that reflects investor sentiment changes. The results showed that changes in the index not only affected individual firms, but also the changes in the entire stock market. Gilbert and Karahalios [31] proposed that people's emotional state influenced their choices. They demonstrated that sentiment extracted from Web-related text contained information that could predict stock prices.

Although there have been some researches confirming the correlation between the sentimental tendency of online commentary and the trend of stocks, very few works have been proposed to predict the specific stock price based on sentiment analysis. For example, the authors in [32], based on the user review data of Snowball, a stock review Web site, proposed a method based on the hot optimization path (TOP), which analyzed the relationship between investor sentiment and the stock market. Zhou et al. [33] based on the SVM-ES model and other online emotions, such as disgust, joy, sadness and fear, predicted the stock market and achieve an accuracy of 64.15%.

In this work, we propose our sentiment analysis model based on CNN, which is used to classify stock market comments into bearish and bullish. In order to improve the accuracy of sentiment classification, we first preprocess the short text of stock reviews, such as acronym changes, spelling correction, root restoration and symbol replacement. The daily sentiment index is calculated based on the classification results, which is then fed into the stock prediction module as an input feature. Different from other sentiment analysis works, we combine CNN with word2vec to improve the effectiveness of classification. In addition, by calculating the index, the emotional analysis of social groups is realized, which enables the consideration of emotions from as many investors as possible.

3 Proposed scheme

In this section, we discuss the proposed scheme in details. Specially, we provide an overview of the proposed scheme and the theoretical basis of our models, followed by detailed discussions on the three proposed key modules: (1) sentiment index model, (2) trend item extraction by EMD and (3) improved LSTM algorithm with attention mechanism. A summary of the proposed scheme is presented at the end.

3.1 Overview

In this work, we propose a sentiment and empirical mode decomposition-based LSTM with attention mechanism (i.e., S_EMDAM_LSTM) for stock market closing price prediction. The proposed algorithm contributes in three major aspects: a calculation of sentiment index module based on CNN, a time-series decomposition module based on EMD and a prediction module based on LSTM with attention mechanism.

Conventional time-series analysis mainly adopts historical data as input, while ignoring other stock influential factors and their complicated influencing mechanisms. In this work, the proposed scheme considers investors' sentiment by computing binary sentiment indices as bearish and bullish.

Additionally, as simple sequences are more predictable than complex sequences in general, in this work, we propose to adopt EMD to decompose the closing price series (i.e., the complex sequence) into simple sequences and handle them separately to predict the trend item. Furthermore, we propose an improved LSTM model to predict the stationary sequences obtained by EMD.

The nonlinear and non-stationary nature of stock prices limits the applicability of conventional models, such like AR [34]. In this work, we propose to adopt LSTM to predict the stock market due to its advanced capabilities to model nonlinear relationship among output and various

input factors. Furthermore, in order to focus on information that is more critical to the current task objective, we propose to integrate attention mechanism with the LSTM model by adding an attention layer.

Figure 1 demonstrates the structure of the proposed S_EMDAM_LSTM model. The design details will be discussed in the rest of this section.

3.2 Preliminaries

3.2.1 Convolutional neural network

The CNN model proposed by Kim [35] is adopted as a basis for the sentiment analysis module in this work. The opening of the text classification and the structure of the model are shown in Fig. 2.

The model is divided into three parts: input layer, convolution layer and classification layer. The input layer is an $r * u$-dimensional text word vector matrix, where r is the number of characteristic phrases for each text and u is determined by data processing. The convolutional layer first passes through the convolution kernel w of the length h

to convolve the word vector matrix. In particular, the convolution process is:

$$t_i = f(w * s_{i:i+h-1} + b) \tag{1}$$

where $s_{i:i+h-1}$ is a continuous text segment consisting of the ith phrase to the $i + 1$th phrase; $*$ is a convolution operation; b is a bias term; and f is a nonlinear activation function. Then, the batch normalization (BN) algorithm is used to normalize the training speed. The maximum value pooling is used to reduce the dimensionality, so that the number of features is consistent.

The classification layer prevents data distribution changes through the BN algorithm and calculates the classification probability through the softmax layer. The stock market comments are classified by probability, which is calculated as:

$$P_j = P(y = j | X, b) = \frac{e^{X^{\mathrm{T}} W_j + b_j}}{\sum_{i=1}^{L} e^{X^{\mathrm{T}} W_i + b_i}} \tag{2}$$

where P_j is the probability that the text belongs to the jth class; X, W, b_i and b_j are the inputs of the classification layer, the weight matrix, the ith element of the offset term
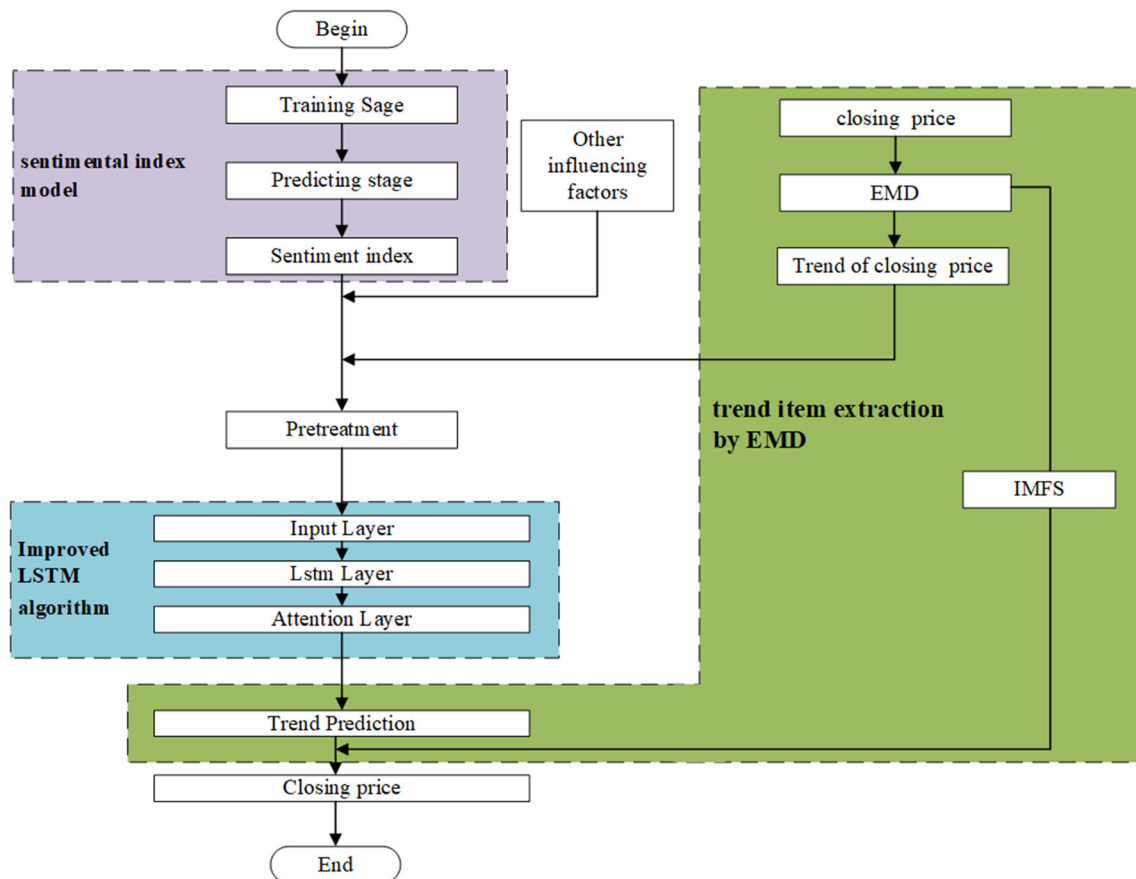

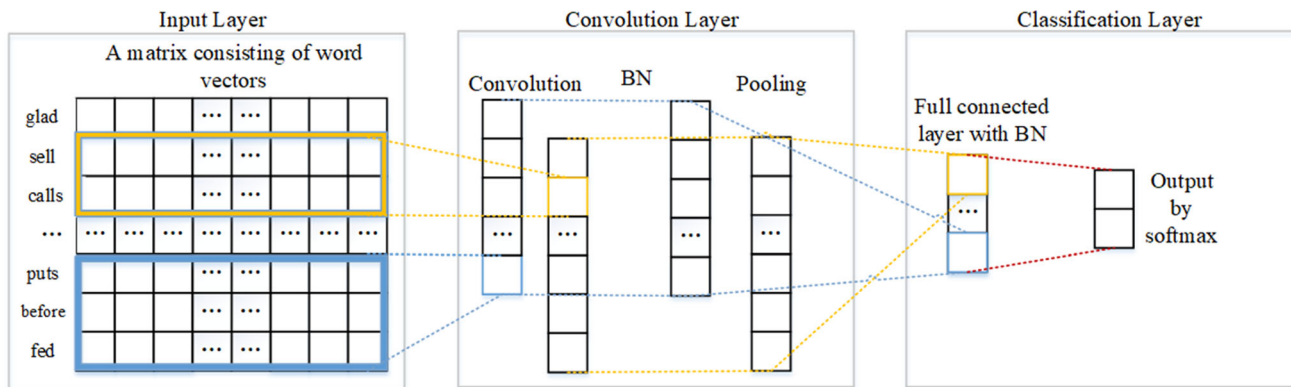**Fig. 1** The flowchart of S_EMDAM_LSTM

Fig. 2 Structure of CNN adopted in this work

and the jth element of the bias term, respectively; and L is the number of categories.

3.2.2 Empirical mode decomposition (EMD)

EMD is a signal analysis method proposed by Huang [36], in 1998. The key of EMD is that it can decompose a complex signal into a finite intrinsic mode function (IMF) and the sum of the residual waves. The decomposed IMF components contain local characteristic signals of different time scales of the original signal. The EMD method can be applied theoretically to the decomposition of any type of time-series signal and thus has a significant advantage over the previous stationary method in dealing with non-stationary and nonlinear data. Therefore, it has been applied quickly and effectively in different engineering fields, such as focal electroencephalogram signals [37], wind speed [38] and automatic bearing fault diagnosis [39]. Specifically, the EMD decomposition method is based on the following assumptions.

1. The data have at least two extreme values, a maximum value and a minimum value.
2. The local time-domain characteristics of the data are uniquely determined by the time scale between the extreme points.
3. If the data do not have an extreme point but an inflection point, the extreme value can be obtained through one or multiple data differentials. The decomposition result can then be calculated through integration.

The obtained IMF should satisfy the following two conditions. (1) For a component signal, the offset between the number of extreme points and the number of zero crossings is not greater than 1. (2) The mean of the upper and lower envelopes defined by the extremum is zero at any time. After a non-stationary time series is decomposed by EMD, multiple random terms, periodic terms and a residual trend

term will be obtained. The decomposed subsequences have stronger regularity and are beneficial to the improvement in prediction accuracy. The extracted residual term is used in this work as a trend term to predict.

The detailed decomposition algorithm is briefly described in Algorithm 1.

Algorithm 1 EMD Algorithm

Input: $x(t)$
Output: $c_i(t)$
    Initialization. $r_1(t) = x(t)$, $i = 1$, $k = 0$
    while if the margin $r_{k+1}(t) = r_k(t) - c_{k+1}(t)$ is a monotonic function or a constant **do**
        Initialization. $h_1(t) = r_1(t)$
        Find all the maxima and minima points of the data sequence $h_k(t)$
        Use the cubic spline interpolation function to fit the upper and lower envelopes, $e_+(t)$ and $e_-(t)$
        Calculate the mean of the upper envelope and the lower envelope $m_k(t)$
        $h_{k+1}(t) = h_k(t) - m_k(t)$
        if SD $= \frac{\sum[h_k(t) - h_{k-1}(t)]^2}{\sum[h_{k-1}(t)]^2}$ is less than a given threshold **then**
            $c_i(t) = h_k(t)$
        else
            $k = k + 1$
        end if
    end while

3.2.3 Long short-term memory

The core idea of LSTM is to control the increase or deletion of information through three "gates." RNN based on this LSTM has four neural network layers, and the interaction of these four neural network layers enables it to solve the long-term dependence problem in RNN model training. An LSTM model contains many LSTM units, each of which contains an input gate i, an output gate o, a forgetting gate f and a memory unit c.

At time t, given the input vector x_t and the hidden state h_{t-1} of the previous moment, the LSTM unit calculates the implicit state h_t of the current moment by internal looping and updating:

$$i_t = \sigma\left(U^i x_i + W^i h_{t-1} + b^i\right) \quad (3)$$

$$f_t = \sigma\left(U^f x_i + W^f h_{t-1} + b^f\right) \quad (4)$$

$$o_t = \sigma\left(U^o x_i + W^o h_{t-1} + b^o\right) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \varphi(U^c x_i + W^c h_{t-1} + b^c) \quad (6)$$

$$h_t = o_t \odot \varphi(c_t) \quad (7)$$

Among them, the parameter set $\{U^i, U^f, U^o, U^c, W^i, W^f, W^o, W^c\}$ corresponds to the weight matrix of different gates, $\{b^i, b^f, b^o, b^c\}$ denotes the corresponding offset term, and σ and φ are, respectively, sigmoid. And the tanh is activation function; $\odot$ represents the point-by-point multiplication between vectors.

3.3 The Proposed S_EMDAM_LSTM model

3.3.1 Sentiment analysis based on CNN

In this paper, at the first step we aim to conduct a group sentiment analysis that integrates users' sentiment preferences for the stock along with the stock's historical data as one factor in predicting the closing price of the stock. The sentiment index is calculated based on the number of daily bullish/bearish comments made by lots of integrated users. Therefore, in order to calculate the sentiment index and get the group sentiment tendency, we first realize the correct sentiment classification of the single stock review.

Different from base model of CNN, we integrate it with word2vec by changing initialization of word vector. Word2vec is a model to learn semantic knowledge from a large amount of text in an unsupervised way. The key of word2vec is to map words from the original space to the new multidimensional space. Specifically, by learning the text, the semantic information of the word is represented by the word vector, and by embedding the space, the semantically similar words are mapped to similar distances. In this paper, we use the Skip-gram model in word2vec to calculate the cosine similarity between the input vector and the target word's output vector, and perform softmax normalization.

In this project, word2vec is introduced to first train large-scale stock comment corpora and learn high-dimensional vector representations of phrases. Then, the word vector representation of the stock comments to be classified is calculated by word2vec. If the phrase in the stock comments to be classified exists in the trained large-scale stock comments corpus, the result is directly used. Otherwise, it is randomly initialized by word2vec. Then, the word vectors, which represent the preprocessed text, will be provided as the input for CNN. After that, based on the

CNN model improved by word2vec, we calculate the sentiment index for sentiment analysis of the group.

The sentiment index of the day is calculated based on the number of daily bullish and bearish comments. It can reflect the investor's overall emotional tendency, which is the group sentiment analysis. We use the method proposed by Antweiler and Frank [40] to calculate the daily sentiment index of the stock.

$$BI_t = \ln\frac{1 + M_t^{\text{bullish}}}{1 + M_t^{\text{bearish}}} \quad (8)$$

where BI_t is the sentiment index of day t and M_t^{bullish} and M_t^{bearish} are bullish stock weights and bearish stock weights, respectively, which are calculated by the number of daily bullish and bearish comments. The index takes the impact of bullish and bearish comments on investor sentiment over a certain period of time into account, and the index changes in direction are proportional to the ratio between the bullish and bearish stock valuation weights.

When more stock comments show bullish sentiment, the index BI_t is positive and the overall propensity of the sentiment is identified as bullish. In the opposite way, if the number of bearish sentiment comments is larger than that of the bullish comments, the sentiment index BI_t is negative, and the overall sentiment is manifested as bearish. The positive or negative sign of the sentiment index BI_t indicates the category of sentimental orientation, and the magnitude of the sentiment index indicates the degree of inclination to a certain category.

In order to further confirm that the sentiment index calculated by Eq. (8) can guide the stock closing price prediction, this paper conducts a Granger causality test [41] on the causal relationship between the sentiment index BI and the closing price. Granger causality test is a method developed by economists to analyze the causal relationship between variables. Given a time series containing the past information of two variables X and Y, if the prediction effect on the variable Y is better than that on the past information of Y alone, X is considered to be the Granger cause of Y. That is, the variable X helps to explain the future changes in the variable Y. The Granger causality test assumes that the information about the predictions for each of y_t and x_t is contained in the time series of these variables. The inspection requirements are:

$$y_t = \sum_{i-1}^{q} \eth_i x_{t-i} + \sum_{j-1}^{q} \beta_j y_{t-j} + u_{1t} \quad (9)$$

$$x_t = \sum_{i-1}^{s} \lambda_i x_{t-i} + \sum_{j-1}^{s} \delta_j y_{t-j} + u_{2t} \quad (10)$$

The white noises u_{1t} and u_{2t} are assumed to be irrelevant.

Equation (9) assumes that the current y is related to the past value of x and y, while Eq. (10) assumes a similar behavior for x.

For Eq. (9), its null hypothesis is $H_0 : \partial_1 = \partial_2 = \ldots = \partial_q = 0$.

For Eq. (10), its null hypothesis is $H_0 : \delta_1 = \delta_2 = \ldots = \delta_s = 0$.

To test this hypothesis, we use the F test. It follows an F distribution with degrees of freedom q and $(n - k)$. Here, n is the sample capacity and q is equal to the number of lag terms x, which represents the number of parameters to be estimated in the constrained regression equation. And k is the number of parameters to be estimated in the unconstrained regression. If the F value calculated at the selected significance level ∂ exceeds the critical value F_∂, the null hypothesis is rejected, such that the lag x term belongs to this regression, indicating that x is the cause of y.

In summary, the proposed sentiment analysis module classifies comments on stock into bearish and bullish. The sentiment index is then calculated based on Eq. (8). The calculated sentiment index will be integrated with the historical data of stock prices the input for the later stock market prediction module.

3.3.2 Extracting trend item from stock closing price sequence

In the proposed work, we are facing great challenges caused by the non-stationary stock time series. Specifically, in time-series analysis, one of the most important assumptions is that the sequence is stationary. That is to say, there is a consistent structural change relationship between the sequence values of each period and that of the previous periods, so that a model can be built to analyze and predict the sequence values.

To limit the impact of the non-stationary stock sequence, we propose a stock closing price prediction model which extracts trend items from non-stationary sequences for prediction. The EMD method can theoretically be applied to the decomposition of any type of signals; thus, it has very obvious advantages in dealing with non-stationary and nonlinear data. In particular, the original stock closing price sequence is fed into the EMD module as inputs. Consequently, several IMFs of the stock closing price will be obtained as outputs. The first IMF c_1 contains the component with the smallest time scale (i.e., the highest frequency) in the original signal. As the IMF order increases, the corresponding frequency component gradually decreases, and the residue r_n contains the component with the lowest frequency. The convergence criterion of the EMD decomposition is that the decomposition residual is a monotonic function, and its period is greater
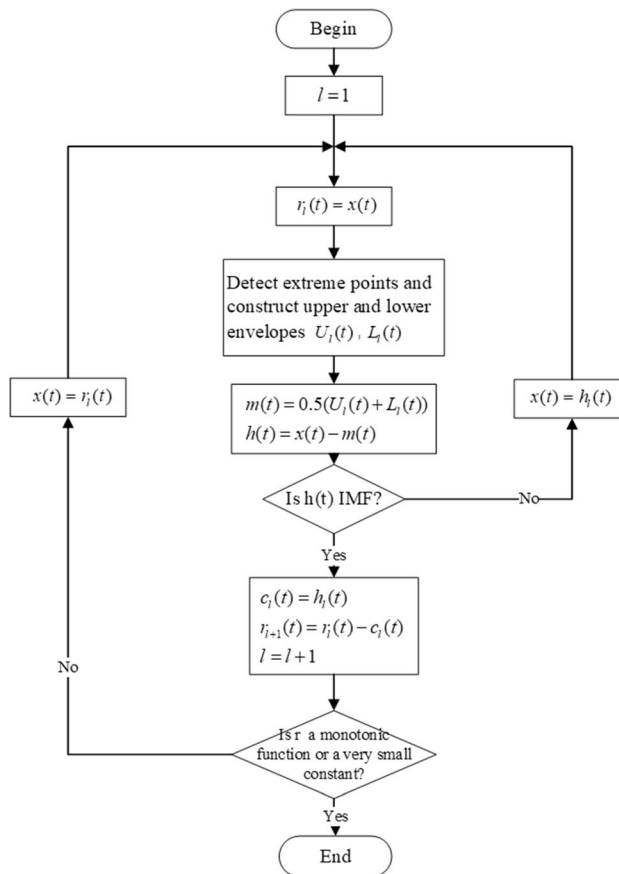

Fig. 3 The flowchart of EMD

than the record length of the signal. In other words, r_n is the trend term of the signals. The trend term extracted by Medill will later be fed into the LSTM prediction model proposed as inputs. The details of the EMD are shown in Fig. 3.

3.3.3 Revised LSTM model with attention mechanism

To better capture the nonlinear relationship between different influential factors and the output stock closing price, we adopt LSTM model for prediction. Furthermore, to enable the model to focus on information that is more critical to the current task objective, we propose to integrate an attention mechanism in the LSTM model. In this section, we will discuss the revised LSTM model in details. Specifically, as shown in Fig. 4, the structure of the model proposed in this work is divided into four layers: input layer, LSTM layer, attention layer and output layer.

– *Input layer* The sentiment index obtained from the emotion index model and the historical data daily are selected as the input eigen matrix. Each eigen matrix is considered as a basic unit and provided to the next level in order.
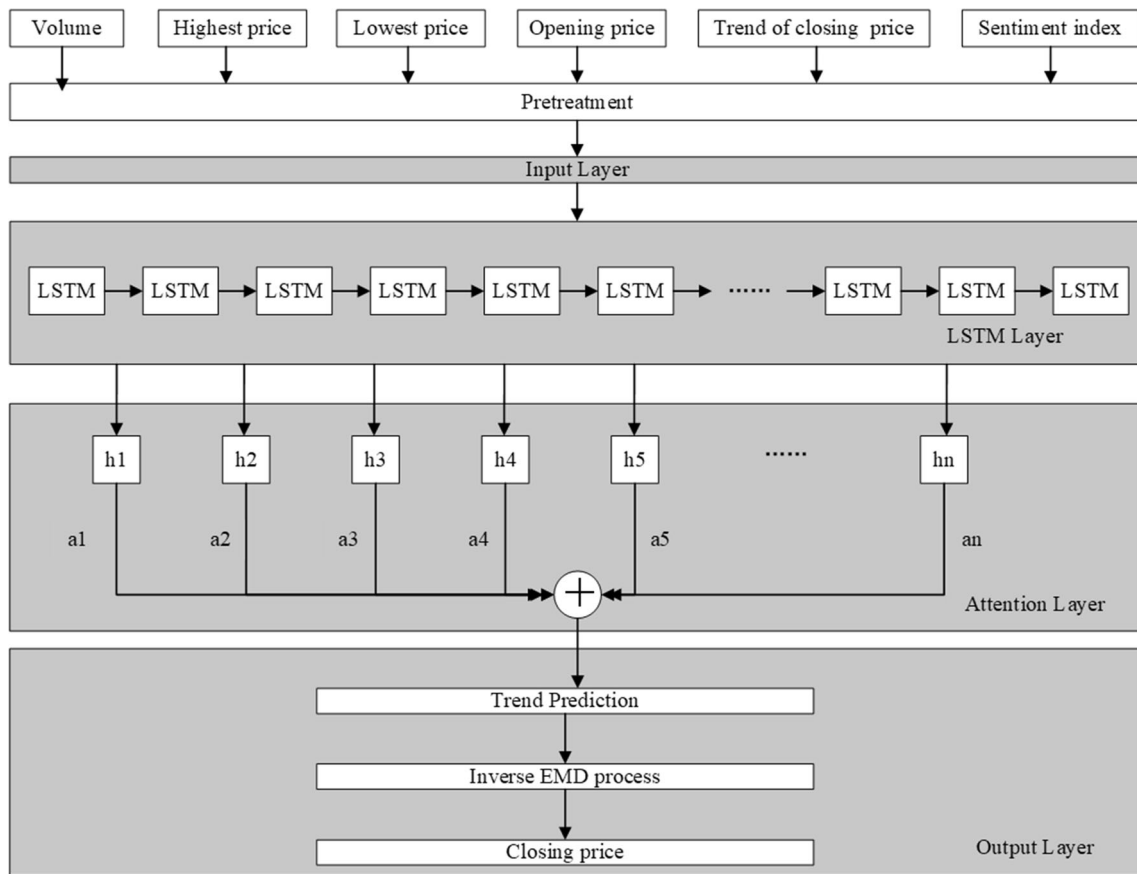
Fig. 4 The structure of EMDAM_LSTM model

- *LSTM layer* In this paper, the neural network layer where the LSTM unit is located is called the LSTM layer, which is one of the main layers of the model. There are multiple LSTM cells in the LSTM layer, and the number of LSTM cells is equal to the number of time steps. Since we select the historical data of the first 30 days to predict the closing price trend item on the 31st day, the number of LSTM cells is 30. The vector enters the LSTM unit after passing through the input layer. The output of the LSTM at the previous point in time will be used as an input to the LSTM at the next point in time. The number of neurons in the hidden layer is 512.

- *Attention layer* The main function of this layer is to allocate the probability for the information of each time step and to combine the information obtained by each time step in the LSTM layer.

The purpose of the attention layer is to enable the model to automatically identify which components in the input have a greater impact on the classification result, and assign greater weights to them, so that the model can pay more attention to these components. The weight distribution in

the attention mechanism is implemented by the matching module. The degree of matching determines the value of the weight. In the process of back propagation, a match module that calculates the input and output matching is trained. The match module calculates the matching degree of the current input and output and then calculates the matching degrees between the current output and each of the inputs. Since the calculations here are not normalized, we use softmax to ensure that the sum of the output weights is 1. After getting the output and the weight of each input, its weighted vector sum can be calculated as the next input.

$$M = \tanh(H) \tag{11}$$

$$a = \mathrm{softmax}\left(w^T M\right) \tag{12}$$

$$\gamma = Ha^T \tag{13}$$

where H is defined as the output sum of t time steps in LSTM layer; a is the output of the match module, that is, the weight, or the assigned attention; w is the parameter matrix of the network; and γ is the weighted sum of the individual components of H, as well as the input of LSTM at the next moment.

– *Output layer* This layer uses the output of the attention layer to predict the closing price trend term through the activation function "relu."

The matrix obtained through the attention mechanism has the same size as the input. The difference is that the attention mechanism adds "special attention" to some vectors by weighting the attention weights, thereby improving the performance of the model.

Finally, we combine the trend term prediction with the IMFs obtained by EMD to retrieve the prediction of the closing price.

3.4 Summary

In summary, as shown in Fig. 5, the model proposed in this paper is mainly divided into three key parts. First, we calculate a sentiment index through the sentiment index model and make it one of the features. Second, the trend term is extracted by EMD from the closing prices to be

predicted. Finally, the historical data and the sentiment index are taken together as features to predict the trend item through the LSTM model that incorporates the attention mechanism. The predicted trend item will then be used to restore the stock closing price. The availability and comparability of S_EMDAM_LSTM are discussed in Sect. 4.

4 Experiment results

In order to validate the effectiveness of the proposed scheme, we conduct experiments on AAPL (stock of Apple) and discuss the results in this section. The experiments are performed on a computer with Intel i5 2.7 GHz, 8 GBs of RAM, windows 10 system. All the comparison methods are implemented in Python programming language.
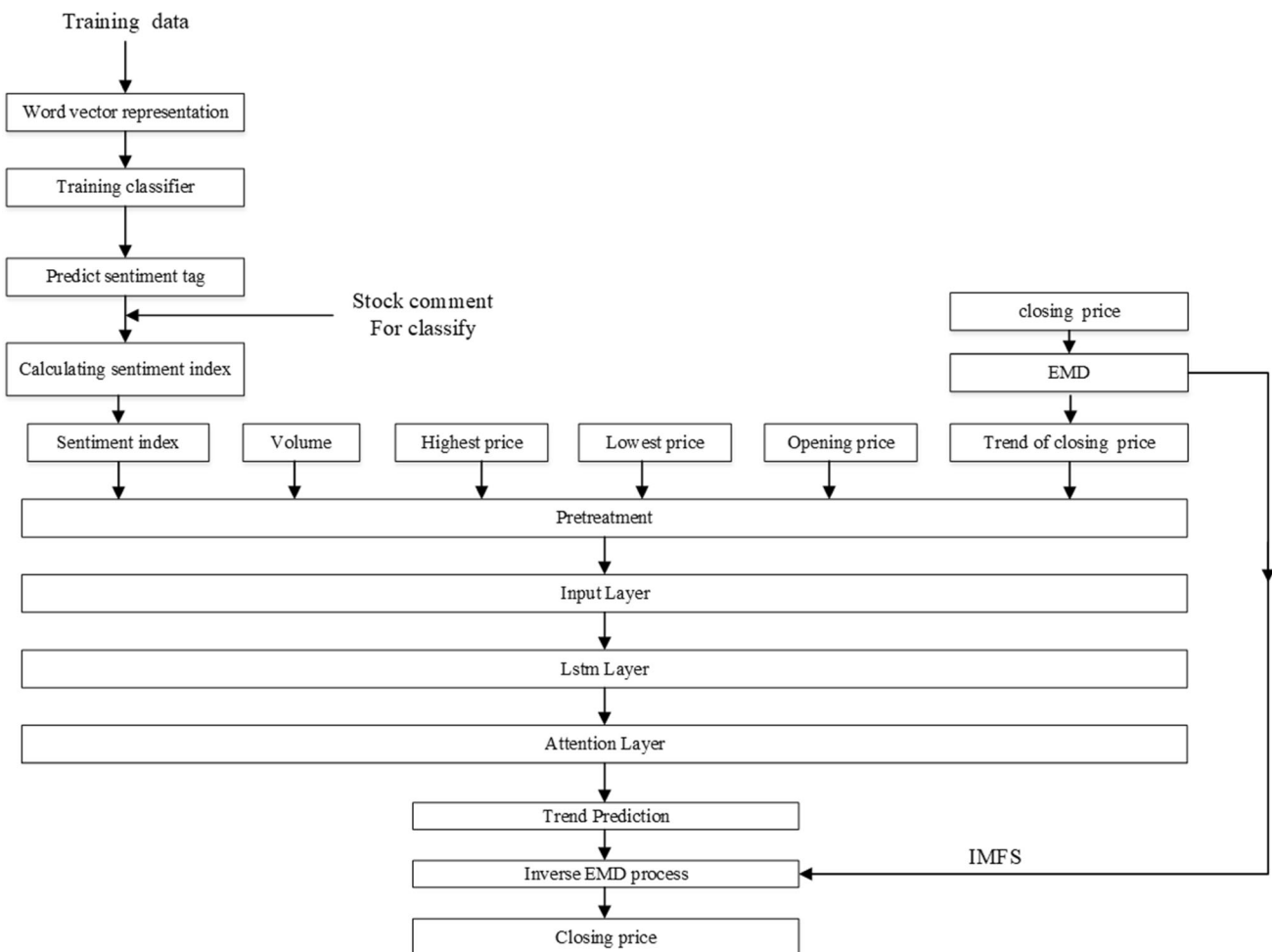

Fig. 5 The structure of S_EMDAM_LSTM

4.1 Dataset

As given in Table 1, the experimental dataset is divided into two parts: the stock comment dataset which is used to obtain the sentiment index and the historical data of AAPL.

The stock comment dataset includes comments for training the model and a final comment to be classified to calculate the sentiment index. First, the comments used to train the model come from comments made by stock-holders on stocktwits (https://stocktwits.com/). Although stocktwits is not the most recognized social platform for investors in the financial sector, investors can choose to give the comment mark a bearish or bullish tag. Therefore, a large number of highly accurate comments can be obtained, which is suitable for training text classification models of CNN. We crawl 96,903 comments on the site for training. Some examples are given in Table 2. Although the comments on the Web site have little noise, they may not represent the views of the broad range of investors. Therefore, the comments on Apple stocks used to finalize and calculate sentiment indices come from Yahoo Finance, the Web's #1 Finance site, which guarantees the accuracy of the data from the data source. Besides, to calculate the sentiment index, we aim to ensure the representativeness of the data by choosing the comments that receive the more number of likes. At the same time, we also choose the comments that receive roughly similar number of likes to avoid sample bias (e.g., two selected comments with very different number of likes). As a result, 80 comments are selected each day to calculate the sentiment index. Considering the number of likes (i.e., roughly 11 likes per comment), each comment reflects the sentimental tendency of 11 people. That is, 880 users' sentiment is considered every day. Last but not least, combined with the time dimension, the total number of samples used to calculate

the sentiment index is 96,903, which on the other hand guarantees the accuracy and representativeness of the data. Experiment results have also verified the effectiveness of the data.

In terms of historical data, we choose five features, including the opening prices, the highest price, the lowest price, the closing price and the trading volume of AAPL for all trading days between March 4, 2013, and February 28, 2018, which is the same to that of stock market comment we have crawled.

Experiments are conducted to evaluate the performance of the proposed scheme. Detailed experiment results and analysis are presented in the following sections.

4.2 Performance evaluation metric

In this paper, our evaluation indicators are divided into two categories. One is the basic evaluation index of the regression model. The other is the time offset indicator defined by us.

Among them, the basic evaluation indicators of the regression model include mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and R-square (R^2). Their calculation is shown as follows.

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |(y_i - \hat{y}_i)| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \tag{15}$$

$$MAPE = \frac{100}{m} \sum_{i=1}^{m} |\frac{(y_i - \hat{y}_i)}{y_i}| \tag{16}$$

Table 1 The description of dataset

Dataset		Size	Function
Comments on stock market	Comments from stocktwits	96,903 comments	Training sentiment classifier
	Comments from Yahoo Finance	80 comments a day for 1219 days	Calculating sentiment index
Historical data of stock market		Five features for 1219 days	Predicting stock prices

Table 2 Examples of comments

Bearish	Bullish
Crazy day so far!	Will make a new ATH this week. Watch it!
Drop it like its hot.	I think the market is going to rip faces next week. I feel a wave of optimism coming
Everything on my watchlist is still waaaaay too overvalued, I'll keep my cash thank you!	Time to drop has gone...big boys positioning themselves to create the next bull run. any dip bears create will be quickly bought up

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{m}(y_i - \hat{y}_i)^2\right)/m}{\left(\sum_{i=1}^{m}(\overline{y}_i - \hat{y}_i)^2\right)/m} \qquad (17)$$

Here, m is the total number of samples; y_i and $\hat{y}_i$ represent the actual and predicted value of the test set, respectively; and $\overline{y}_i$ represents the mean of real values of the test set. The normal range of R^2 is [0,1]. The closer this value to 1, the stronger the ability of the equation to interpret y, and the better the model fits the data.

In addition, we define the time offset indicator as t. For the predicted closing price, the stock price rise and fall is classified by comparing the stock prices of the two adjacent days. The highest bullish bearish classification accuracy rate is delayed by t days. The closer t is to 0, the smaller the model delay.

4.3 Model verification

This section gives the results of the Granger causality test to illustrate the feasibility of adopting the sentiment index for the stock closing price prediction.

The premise of the Granger causality test is the stability of the time series. Therefore, it is necessary to first augment the time series of this experiment to the Dick Fowler test (ADF test). The test shows that the stock closing price is not a stationary sequence, but the first-order differential sequence of the stock closing price is a stationary sequence. Because the first-order differential closing price sequence and the time series of sentiment index are not the same order, the subsequent Granger causality test cannot be performed. For this case, the experiment is processed as follows:

The first-order difference is made in both time series of closing price and sentiment index. The differential sequences are recorded as DC and DBI, respectively. The formula is as follows:

$$D(P) = P_t - P_{t-1}, P = DC, DBI \qquad (18)$$

It is verified that DC and DBI are consistent with the Granger causality test. Table 3 shows the Granger test results for different lag days.

Table 3(a) shows that when the lag days take different values, the probability of accepting the null hypothesis is < 0.05, so the null hypothesis can be rejected in the 95% confidence space. It is concluded that the change in DC is the reason for the change in DBI. The conclusion indicates that the sentiment index calculated by our sentiment index model can directly reflect the user's perception of the closing price and prove the accuracy of the sentiment index model.

Table 3(b) shows that when the lag days take different values, the probability of accepting the null hypothesis is also < 0.05, so the null hypothesis can be rejected in the 95% confidence space. It is concluded that the change in DBI is the reason for the change in DC. The conclusion indicates that the change in the sentiment index is a reason for the change in the closing price, which proves the feasibility of using the sentiment index to predict stock closing prices.

4.4 Experimental results and performance comparison

In this section, the major contributions proposed in this paper, including sentiment index, EMD and LSTM improved with attention mechanism, will be tested and verified, respectively. Specifically, we choose LSTM as our baseline and have conducted three experiments as follows.

In the first experiment, we validate the effectiveness of the sentiment index by comparing it to the LSTM model without sentiment index. The LSTM model considering the sentiment index is referred to as S_LSTM. The results are shown in Fig. 6. In Fig. 6, the x-axis and the y-axis

Table 3 The Granger test results for different lag days

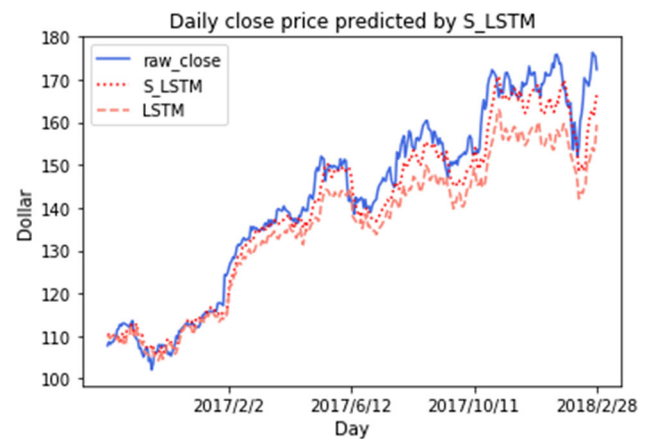| Null hypothesis	Lag days	Prob
(a)		
DC does not Granger cause DBI	2	0.0120
	3	0.0006
	4	0.0024
	5	0.0040
(b)		
DBI does not Granger cause DC	2	0.0342
	3	0.0410
	4	0.0462
	5	0.0160

Fig. 6 Daily close price predicted by S_LSTM

represent the date the closing price of the AAPL, respectively. The line labeled "rawclose" represents the real stock closing price, the line labeled "LSTM" is the predicted closing price of the baseline, and the line labeled "LSTM with sentiment index" is the predicted closing price of the LSTM model with sentiment index. As shown in Fig. 6, compared to the baseline algorithm, S_LSTM is closer to actual value, indicating that the introduction of sentiment indicators improves the prediction accuracy.

This is because the baseline algorithm assumes that capital market participants are all computers, all without emotion, super-rational, and the behavior is completely in accordance with the principle of interest. In fact, this is not the case. Not every market participant can act in a rational way according to the theoretical model. Human irrational behavior plays an important role in the economic system. For example, people are often prone to errors in investment judgments and decision making, and when such mistakes occur, they are often very sad/frustrated. Therefore, in the process of investing, investors often show an indecisive personality trait in order to avoid the appearance of regret. When investors decide whether to sell a stock, they are often affected by the emotional cost of buying when the cost is higher or lower than the current price. Because of fear of regret, they try to avoid regrets.

Behavioral finance also provides support for this view. It studies and explains the phenomenon of stock market changes based on human psychological characteristics and behavioral characteristics. It historically acknowledges that stock market changes are not purely objective in many cases, but related to the psychological and behavioral characteristics of participants. The stock market is largely a reflection of human nature, and many phenomena in the stock market do not conform to the principles of science and established logic.

In the second experiment, we further validate the effectiveness of the attention mechanism The LSTM model

that considers the sentiment index and attention mechanism is denoted as S_AM_LSTM. The results are shown in Fig. 7. In Fig. 7, the x-axis, the y-axis, the line labeled "rawclose" and "LSTM" are same to that of Fig. 6, respectively. And the line labeled "prediction" is the predicted closing price of the LSTM model with sentiment index attention mechanism. Figure 7 shows that the introduction of attention mechanism improves the prediction performance.

We predicted the 31st stock closing price through historical data and sentiment index for the past 30 days. However, the sentiment index and stock price history data have different degrees of influence on the results to be predicted on different dates. The sentiment index, opening price, highest price, lowest price and transaction volume of the same date also have different effects on the predicted results. Errors are introduced if they are treated equally to give the same weight. The attention mechanism solves this problem very well. Based on the LSTM model improved with attention mechanism, through training and learning, different weights can be reasonably assigned according to the degree of influence. Through the attention mechanism, this paper selectively screens a small amount of important information and focuses on it, while ignoring most of the unimportant information.

In the third experiment, we further examine the impact of the proposed combination of EMD with LSTM. The LSTM model that takes into account the attention mechanism, the sentiment index and the EMD is denoted as S_EMDAM_LSTM. The results are shown in Fig. 7. In Fig. 7, the x-axis, the y-axis, the line labeled "rawclose" and "LSTM" are same to that of Fig. 6, respectively. And the line labeled "prediction" is the predicted closing price of the LSTM model with sentiment index, attention mechanism and EMD. Figure 8 shows that the introduction of EMD improves the model performance.
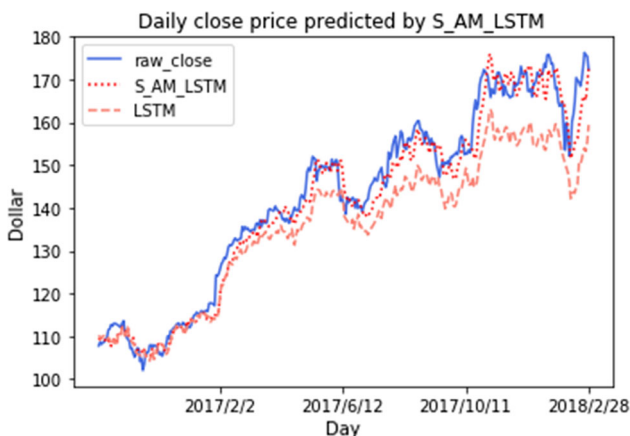

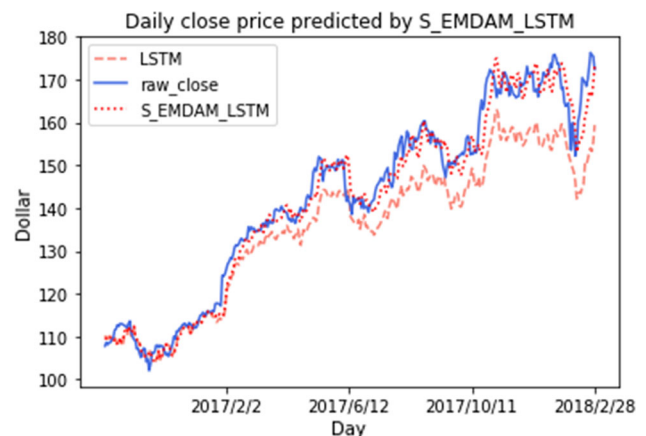Fig. 7 Daily close price predicted by S_AM_LSTM

Fig. 8 Daily close price predicted by S_EMDAM_LSTM
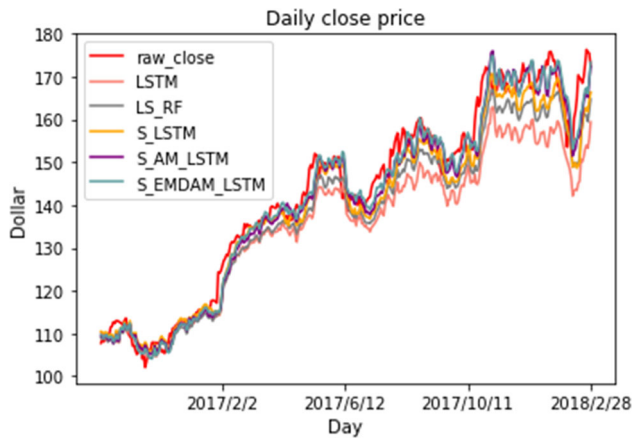
Fig. 9 The prediction results under different models

Simple sequences are more predictable than complex sequences, and complex variability factors are eliminated by EMD to obtain more predictive temporal variation trends, i.e., those that change over time.

In order to prove the validity of the proposed scheme, we also compared it with the LS_RF model in the literature [42]. Figure 9 shows the prediction results under each model. In Fig. 9, the x-axis and the y-axis are same to that of Fig. 6.

As can be seen from the figure, LSTM can be used to predict the stock market, and the S_EMDAM_LSTM model proposed in this paper is the closest to the actual value of the closing price. The sentiment index, attention mechanism and EMD that we incorporate in LSTM also can improve LSTM.

Table 4 gives the detailed results of the evaluation indicators for each model where t is the time offset and ACC is the prediction accuracy under the corresponding time offset. Figure 10 shows more intuitive results. Figure 10 and Table 4 show the quantitative impact of sentiment orientation on stock closing price. The introduction of sentiment orientation has improved the evaluation indicators. As shown in Fig. 10, our model has achieved better results, for both the regression model indicator, including MAPE, MAE, RMSE and R^2, and the time delay indicator. The time delay of stock closing price predicted by 30-day

historical data has been shortened from 9 to 2 days, which has been greatly improved.

Figure 11 and Table 5 show the accuracy and time offset comparison between our proposed model and the model in reference [42]. Compared with the LS_RF model, our model has increased the correlation coefficient by 5.74% and the accuracy of the rise and fall classification by 11.01%.

The improvement in the results comes from three aspects: (1) based on behavioral finance, considering the influence of emotional factors on stocks, (2) based on EMD extraction trend items, predicting more predictable sequences and (3) through improved LSTM with attention mechanism, paying more attention to more important information in large amounts of data.

Firstly, in our model, human factors are not excluded as hypotheses, so behavioral analysis is included in theoretical analysis. Our model considers that the stock price is not determined solely by the intrinsic value of the firm, but is also largely influenced by the investor's subjective behavior. That is, the investor's psychology and behavior have a significant impact on the price decision and its changes in the securities market. In our S_LSTM model, the essence of bullish bearish analysis of stock reviews is sentiment analysis, and the introduction of sentiment index introduces the investor's overall emotional tendency, taking into more account factors that affect the stock price.

In addition, the performance improvement is also because of the addition of the attention mechanism. Let the model automatically recognize the most influential components in the input sequence and assign higher weights to them, so that the model increases the attention to the component. Through the attention mechanism, we can get the target area that needs to be focused on, which is the focus of attention, and then invest more attention resources in this area to get more details of the target to focus on. Ignoring other useless information, the means is able to quickly extract high-value information from a large amount of information using limited attention resources. Through the attention mechanism, this paper selectively screens only a small amount of important information and focuses on it, while ignoring most of the unimportant information. The process of focusing is reflected in the calculation of the

Table 4 The detailed results of the evaluation indicators for each model

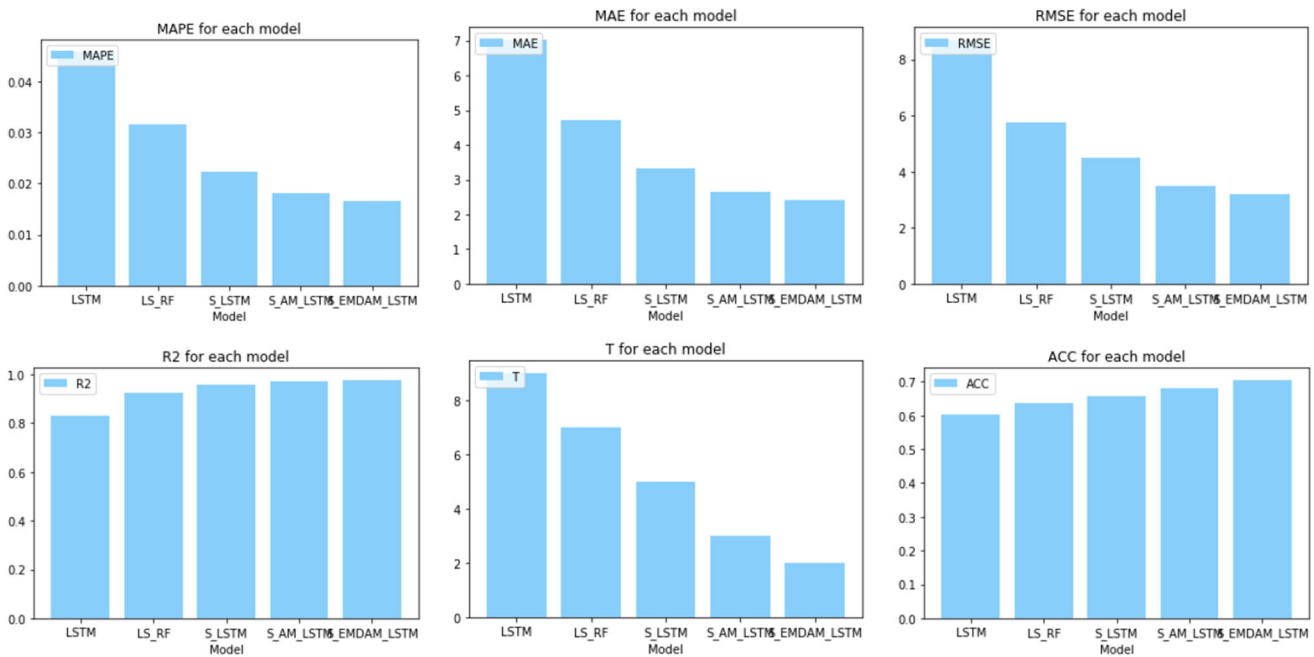|	MAPE (%)	MAE	RMSE	R2	t	ACC
LSTM	4.58	7.031646	8.712122	0.832031	9	0.6012
LS_RF	3.15	4.713792	5.756513	0.926667	7	0.6354
S_LSTM	2.23	3.327806	4.482615	0.955532	5	0.6573
S_AM_LSTM	1.82	2.64963	3.475939	0.973262	3	0.6808
S_EMDAM_LSTM	1.65	2.396121	3.196534	0.977388	2	0.7056

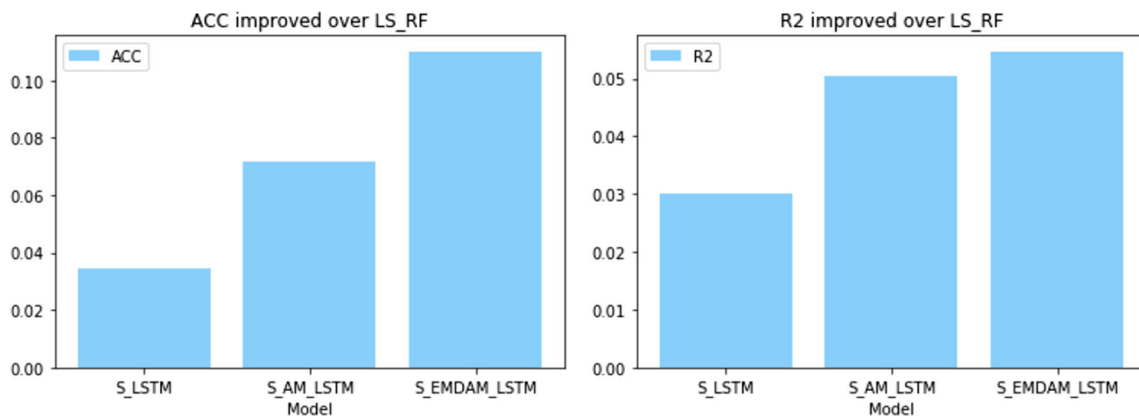Fig. 10 The accuracy and time offset comparison between different models

Fig. 11 The increase in evaluation indicators

Table 5 The increase in evaluation indicators

	S_LSTM (%)	S_AM_LSTM (%)	S_EMDAM_LSTM (%)
R^2	3.02	5.03	5.47
ACC	3.45	7.15	11.01

weighted coefficients, where the weight represents the importance of information.

Moreover, the addition of EMD to decompose complex sequences into stationary sequences also improves the model predictability. Through the EMD, the trend items are extracted, and the random factors determined by the fuzzy module are removed from the data, so that the patterns can be better predicted.

As a summary, with LSTM combined with sentiment index, attention mechanism and EMD, the proposed scheme consistently achieves the highest accuracy, the lowest time offset and the closest predictive value when predicting the stock market. Although there have been many studies focusing on the rise and fall of stocks, there are few studies that focus on the specific price and time-liness of stocks. Timely and accurate prediction of stock prices is critical to investor choice and national economic stability. More accurately and timely stock price can be predicted, more timely and reasonable regulation and healthy guidance of the stock market can be provided, which will make sense for backing sustainable development of the economy solidly.

5 Conclusion

In this work, a novel LSTM-based model is proposed for stock market prediction. Expressly, for one thing, sentiment index is used to take the investor's emotional tendency into consideration. For another, the LSTM-based model is improved by EMD, which decomposes the complex stock pricing sequence into simple and more predictive sequences, and attention mechanism, which help the model focus on the most contributing information of the current task target.

According to experiments conducted on the dataset of AAPL, the performance of the proposed scheme has been verified. The experimental results show that the proposed scheme outperforms the comparison schemes consistently in three main aspects, including closer predicted closing price, higher rise and fall classification accuracy and lower time offset.

Therefore, the proposed scheme in this paper shows great potential to benefit the country by providing direction for government on rationalizing and guiding the stock market and to provide profits for individuals by guiding investments.

Last but not least, in this work, we adopt CNN as our based model of sentiment index. In the further work, CNN in the proposed scheme can be replaced by other base learners. For example, using some semi-supervised models for emotional classification and sentiment index calculations will reduce the requirements for stock review datasets to some extent. Further studies will be performed in the future work.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Fama EF (1998) Market efficiency, long-term returns, and behavioral finance. J Financ Econ 49(3):283–306
2. Wang B, Huang H, Wang X (2012) A novel text mining approach to financial time series forecasting. Neuro Comput 83(6):136–45
3. De Gooijer JG, Hyndman RJ (2005) 25 years of IIF time series forecasting: a selective review. Soc Sci Electron Publ 22(3):443–473
4. Neild S (2003) A review of time-frequency methods for structural vibration analysis. Eng Struct 25(6):713–728
5. Krunz MM, Makowski AM (2002) Modeling video traffic using M/G//splinfin/ input processes: a compromise between Markovian and LRD models. IEEE J Sel Areas Commun 16(5):733–748
6. Farina L, Rinaldi S (2000) Positive linear systems. Theory and applications. J Vet Med Sci 63(9):945–8
7. Contreras J, Espinola R, Nogales FJ et al (2002) ARIMA models to predict next-day electricity prices. IEEE Power Eng Rev 22(9):57–57
8. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
9. Judith EDPD, Deleo JM (2001) Artificial neural networks. Cancer 91(S8):1615–1635
10. Prasaddas S, Padhy S (2012) Support vector machines for prediction of futures prices in Indian stock market. Int J Comput Appl 41(3):22–6
11. Cavalcante RC, Brasileiro RC, Souza VLF, Nobrega JP, Oliveira ALI (2016) Computational intelligence and financial markets: a survey and future directions. Expert Syst Appl 55:194–211
12. Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of internet stock message boards. J Finance 59(3):1259–1294
13. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of returns. J Finance 61:1645e1680
14. Dragomiretskiy K, Zosso D (2014) Variational mode decomposition. IEEE Trans Signal Process 62(3):531–544
15. Daubechies I, Lu J, Wu HT (2011) Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool. Appl Comput Harmon Anal 30(2):243–261
16. Tang H, Chiu KC, Lei X (2003) In: Proceedings of 3rd international workshop on computational intelligence in economics and finance (CIEF'2003), North Carolina, USA, September 26–30, pp 1112–1119
17. Duan T (2016) Auto regressive dynamic Bayesian network and its application in stock market inference. In: IFIP international conference on artificial intelligence applications and innovations. Springer, Berlin
18. Hinton G, Deng L, Yu D et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97
19. Zhang DZD, Song HSH, Chen PCP (2008) Stock market forecasting model based on a hybrid ARMA and support vector machines. In: International conference on management science and engineering. IEEE
20. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
21. Xingjian SH, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems 2015, pp. 802–810
22. Ma X, Tao Z, Wang Y et al (2015) Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp Res Part C Emerg Technol 54:187–197
23. Liu H, Mi X, Li Y (2018) Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM. Energy Convers Manag 159:54–64
24. Ding X, Zhang Y, Liu T, Duan J (eds) (2015) Deep learning for event-driven stock prediction. In: International conference on artificial intelligence
25. Fischer T, Krauss C (2017) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res. S0377221717310652
26. Mao Y, Wang B, Wei W, Liu B (2012) Correlating S&P 500 stocks with twitter data. In: Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research, New York, vol 12e16, pp 69–72
27. Mittal A, Goel A (2009) Stock prediction using twitter sentiment analysis. https://pdfs.semanticscholar.org/4ecc/55e1c3ff1cee41f21e5b0a3b22c58d04c9d6.pdf. Accessed 9 May 2016

28. Kaminski J, Gloor PA (2014) Nowcasting the bitcoin market with twitter signals. Accessed 9 May 2016

29. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. J. Finance 61(4):1645–1680

30. Baker M, Wurgler J (2007) Investor sentiment in the stock market. J Econ Perspect 21:129–152

31. Connor B (2010) In: Fourth international AAAI conference on weblogs and social media. Aaai Publications

32. Guo K, Sun Y, Qian X (2017) Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. Physica A Stat Mech Appl 469(C):390–396

33. Zhou Z, Zhao J, Xu K, (2016) can online emotions predict the stock market in China?. In: Cellary W, Mokbel M, Wang J, Wang H, Zhou R, Zhang Y (eds) Web information systems engineering—WISE. WISE 2016. Lecture Notes in Computer Science, vol 10041. Springer, Cham

34. Hu Z, Liu W, Bian J, Liu X, Liu T-Y (2018) Listening to chaotic whispers: a deep learning framework for news-oriented stock trend prediction. In: Proceedings of the eleventh ACM international conference on web search and data mining, pp 261–269. ACM

35. Kim Y (2014) Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp 1746–1751

36. Huang NE, Shen Z, Long SR et al (1971) The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis. Proc A 1998(454):903–995

37. Sharma R, Pachori R, Acharya U (2015) Application of entropy measures on intrinsic mode functions for the automated identification of focal electroencephalogram signals. Entropy 17(2):669–691

38. Wang S, Zhang N, Wu L et al (2016) Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method. Renew Energy 94:629–636

39. Ben Ali J, Fnaiech N, Saidi L et al (2015) Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. Appl Acoust 89:16–27

40. Antweiler W, Frank MZ (2004) Is all that talk just noise? The news content of internet stock message boards. J Finance 59:1259–1294

41. Ding M, Chen Y, Bressler SL (2006) Granger causality: basic theory and application to neuroscience. Quant Biol 437:826–831

42. Sharma N, Juneja A (2017) Combining of random forest estimates using LSboost for stock market index prediction. In: 2nd international conference for convergence in technology (I2CT), Mumbai, pp 1199–1202. https://doi.org/10.1109/I2CT.2017.8226316