Arabian Journal for Science and Engineering (2023) 48:9941–9962
https://doi.org/10.1007/s13369-022-07444-7

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Analysis of Stock Market Public Opinion Based on Web Crawler and Deep Learning Technologies Including 1DCNN and LSTM

Jizheng Yi[1,2] · Junsong Chen[1,2] · Mengna Zhou[1,2] · Chao Hou[1,2] · Aibin Chen[2] · Guoxiong Zhou[2]

Received: 7 April 2022 / Accepted: 30 October 2022 / Published online: 15 November 2022

Abstract
As the center of the financial market, the stock market is popular with the public attention of investors. It is of great significance for investors that an effective analytic method of stock public opinion is proposed. As the main communication platform, the forum not only provides the investors with investment information but also comments related to the stock market. In view of the defects of text emotions and investment problems, this paper proposes a framework based on web crawler and deep learning technologies including one-dimensional convolutional neural networks (1DCNN) and long short-term memory (LSTM), to evaluate the stock market volatility. Among them, the extracted features include not only the stock price but also the text information. Firstly, we develop the crawler technology to grab large-scale text data from the internet and they are manually labeled their emotions by analyzing the relevant financial knowledge. Secondly, as the character-level text classification method, the 1DCNN is designed for text sentiment classification to detect the reliability of text annotation. Finally, considering the time sequence of price and the continuity of post influence, the emotional and technical features are combined to estimate the fluctuation of the stock market in different industries by the LSTM model. We test four evaluation indexes, the classification accuracy of the model is 74.38%, the accuracy rate is 76.83%, the recall rate is 70%, and the F1 value is 72.8%. The results show that the combination of characteristics of internet public opinion more effectively evaluates the changes in the stock market.

Keywords Stock market · Public opinion analysis · Web crawler · One-dimensional convolutional neural networks (1DCNN) · Long short-term memory (LSTM)

1 Introduction

With the rapid development of the market economy, the financial market at the core of the market economy has also achieved a great breakthrough. On February 23, 2019, the Chinese President emphasized that "preventing and defusing financial risks, especially systemic financial risks, is the fundamental task of financial work." It can be seen from this that financial risk is one of the cores of the development strategy all the time [1]. In fact, other countries in the world also attach importance to the development of the financial market. In order to better prevent risks, we need to adopt better methods to analyze them. Stock markets have become increasingly complex and changeable. The existing basic theories no longer apply to the changes in the stock market. The stock market is of great difficulty to predict based on the theory of random walks. Nevertheless, the research on stocks is still a cynosure [2–5]. Behavioral finance believed that stores are influenced by the words and actions of others to a large degree when they make decisions [1].

In the context of the development of information technology at full speed, the Internet has gradually penetrated and exerted a great influence on people's daily life. Netizens can quickly obtain information and discuss their views on the Internet platforms, such as Weibo, WeChat and forum. These social media have become the main platforms to express people's emotions [6]. Various opinions that express people's thoughts are gradually generated. As time goes on, the

Junsong Chen and Mengna Zhou have contributed equally to this work.

✉ Jizheng Yi
kingkong148@163.com

[1] College of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha 410004, China

[2] Institute of Artificial Intelligence Application, Central South University of Forestry and Technology, Changsha 410004, China

generation of this information makes us gradually enter the era of big data, which inspires us to study various problems [7–9]. Apala et al. [10] predicted box office through the data from Twitter, YouTube and Internet Movie Database (IMDb). Golbeck et al. [11] adopted public information of the users on Facebook to predict their personalities. At the same time, the Internet has become the main platform for investors to communicate in the stock market. They could obtain many pieces of information about the stock market, macroeconomic indexes, expert comments and analyses, etc. The posts, post volume attention and other information in social media can all serve as the research basis for the stock market [12]. Users on the Internet express their true thoughts, attitudes, emotions, and other opinions on the network events through open and casual information communication platforms, thereby forming network public opinion. It can reflect the development trend of public opinions and help managers better understand public opinions [13]. Various social media gather the ideas of numerous people and lots of stock information can be mined from texts related to stock to evaluate the development trend of stocks [14–19]. In this paper, we obtain the text information from web crawler in East money. East money is an authoritative financial website with a large data flow and up-to-date information, which is also the website with the most visitors among all the financial websites. There are many pages in the posts on the forum. If you download information manually, it will be not only inefficient but also a heavy workload. To solve this problem, web crawler technology which grabs the relevant web page information directionally is generated. Web crawlers, also known as web spiders, are programs or scripts that automatically grab the information on the World Wide Web based on certain rules, and effectively access relevant web pages and links to obtain the required information according to the given capture target. In light of the system structure and implementation technology, web crawlers are generally divided into general crawlers, focused crawlers, incremental crawlers and deep web crawlers [20–22]. However, in practical applications, most crawlers are implemented in combination [23, 24].

The text classification is the most crucial for analyzing stock movement in this paper. In the era of big data, the traditional methods of text sentiment analysis mainly include artificial dictionary construction and machine learning. But the two methods not only cost lots of manpower but also have low efficiency and quality. Therefore, deep learning, an important research field in machine learning, is utilized for text analysis. Deep learning improves the accuracy of text classification by constructing a network model simulating the human brain and nervous system to analyze the texts and automatically optimize model parameters. Deep learning has multi-layer structures. The nonlinear mapping between these structures can make it better deal with complex functions. A large number of text data will cause a burden on the learning process. However, deep learning obtains the important variables of input data through a layer-by-layer learning algorithm to avoid the phenomenon of overfitting. At present, convolutional neural networks (CNN) and recurrent neural networks (RNN) mainly are adopted for text classification [25, 26]. Because the classification accuracy not only depends on the model but also has many uncertain factors. In this paper, we first compare the commonly adopted models in text classification and use the best performing model 1DCNN to implement sentiment classification in text experiments. Secondly, public opinion analysis, combines the text and data characteristics of the stock market to evaluate the stock trend. During this process, we choose long short-term memory (LSTM) network to complete the public opinion analysis experiment.

Nowadays, deep learning technology has gradually entered various fields including the financial world. We adopt 1DCNN to classify the text and analyze the sentiment of the text with the stock market, which is very significant research. The research not only expands the application field of deep learning but also brings good news to the financial market. This paper firstly analyzes the sentiment expressed by them through the title of posts in the Guba, and then analyzes the trend of the stock market combining text information with the historical data of the stock.

To sum up, the main contributions of this paper are listed as follows:

- In this paper, web crawler technology playing the role of the information transmission channel is developed to obtain text data, in which the most important part is to access the web server (including user authorization and file download) and parse the required HTML files. Here the Python language is adopted to implement them with requests and Beautiful Soup libraries.
- The commonly manual dictionary construction method is not adopted in this paper to process the text, but character embedding is utilized to realize the text classification. This method recognizes the emoticons in the text without considering the semantic and grammatical structure of the language.
- Combining two features, this paper proposes a framework to enhance the assessment of stock market movement. The two features are the emotional feature of text and the stock price feature of stock market trading. Among them, the sentiment tendency of large-scale text data should be labeled manually before the feature extraction.
- In order to obtain more comprehensive and valuable information to evaluate the stock trend, the characteristic information is extracted to analyze the trend of the stock and real multi-classification (the stocks' rise and fall).

The remainder of the paper is structured as follows: Sect. 2 reviews some related works. In Sect. 3, we introduce the details of the proposed method. Section 4 describes and analyses the experimental results. In Sect. 5, we summarize the paper and outline further work in future.

2 Related Work

The most important point to analyze the trend of stock price by network public opinion is how to realize sentiment analysis on the obtained text. In this section, we would like to introduce the related work from the following two aspects.

2.1 Deep Learning in Text Sentiment Classification

Since the rapid development of deep learning technology in 2012, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have gradually been widely used in the field of Natural Language Processing (NLP). It makes text classification easier and the accuracy continuously improved. CNN is a feedforward neural network that adopts convolution calculations instead of general matrix multiplication. CNN reduces the connection between network layers by sharing weights. To reduce the number of parameters between network layers and avoid the overfitting risk, CNN also performs convolution operations on text information to extract their local features. The subsampling, also known as the pooling layer, contains max-pooling and mean pooling. The convolution and subsampling can reduce the complexity and parameters of the model. The most important part of CNN for text classification is the convolution layer. It can solve the problem of sequence information loss between words caused by traditional classification methods. After passing the convolution layer, n words can be combined to improve classification accuracy. The first use of CNN for text classification was proposed by Kim in 2004 [27]. In his experiment, the single-layer CNN was adopted to model the text, the preprocessed text word vector was taken as input, and then CNN was adopted to realize sentence-level fiction. Sun et al. [28] proposed a CNN Weibo sentiment analysis method that combined posts and comments with a new convolutional auto-encoder. It can extract contextual emotional information from Weibo conversations. Liao et al. [29] designed a simple CNN model and analyzed sentiment in the Twitter database to predict user satisfaction with products and specific environments, or the damage after a disaster. The method has higher accuracy in sentiment classification than traditional support vector machines (SVM) and Naive Bayes. Dos Santos et al. [30] proposed a new deep convolutional neural network that adopted information from character to sentence level for sentiment analysis of movie reviews and Twitter messages. CNN model is mainly used for short text, and the RNN model is generally used for long text. RNN is a kind of recurrent neural network that focuses on structure level and has memory function. By self-feedback neurons, they process sequences of arbitrary length. RNN preprocesses text of different lengths: truncating text for long data and filling text for short data. RNN takes each word as a time node, and the word vector as the input feature of the text. It usually combines sentences forward and backward to construct a bidirectional feature. The classification model of RNN is very flexible and has various structures. Zhang et al. [31] proposed a sentiment method based on an RNN. This method utilizes distributed word representation technology to construct a vector for words in sentences, and then uses RNN to train fixed latitude vectors for sentences of different lengths. In this way, the resulting sentence vector can contain word semantic and sequence features. Abdi et al. [32] proposed a method based on deep learning to classify the users' opinions by comments. The method uses the advantages of RNN composed of Long Short-Term Memory (LSTM) and sequence processing to overcome the disadvantages of sequence and information loss in traditional methods. Yan et al. [33] designed the model of encoder and decoder structure by using LSTM, and solved the problems in time series prediction that multiple input features have different influences on the target sequence and the data before and after the sequence have strong time correlation by assigning the weights of different input features and time points. The experimental results show that the unified feature set learning method significantly obtains better performance than the method of learning from a feature subset. This paper selects 1DCNN to achieve the classification of the text sentiment through the experiment comparison on related models.

2.2 Analysis of Online Public Opinion

Social networks are full of people with different identities and educational backgrounds. For professionals, their opinions are often followed by others and they may become leaders in the investment market. However, for ordinary people, their lack of professional knowledge makes them unable to obtain accurate information, leading to blind following. With the advent of machine learning, these problems have also been solved to some extent.

Schumaker et al. [34] used financial news as a text extraction source, and support vector machines (SVM) as a classification model to predict the stock market. They studied the role of financial news in three different text feature representations: word bags, noun phrases, and named entities. The results showed that noun phrases performed better on stock prediction than word bags, and the prediction by SVM is better than that by linear regression. Some methods to analyze the stock market are by using the emotions conveyed by the text [35–37]. Bollen et al. [38] proposed a

self-organizing fuzzy neural network to study the influence of Twitter sentiment value on the stock market price prediction. It was found that certain specific sentiments can improve the prediction accuracy in the stock price prediction. The sentiment states of this experiment are divided into Clam, Alter, Sure, Vital, Kind, and Happy, among which Clam can be used to predict Dow Jones Index. Patel et al. [39] proposed a Support Vector Regression (SVR), Artificial Neural Network (ANN), and Random Forest fusion method to predict the CNX Nifty and S&P index of the Indian stock market.

In recent years, deep learning belonging to machine learning is also attracting more and more attention in the financial market, mainly including CNN and RNN. Ding et al. [40] proposed a deep learning method based on event-driven stock market prediction, in which news events were selected by text and expressed by dense vectors before inputting the model. Then, the deep convolutional neural network was used to model the long-term and short-term effects of stock price movements. Li et al. [41] proposed a PCC-BLS framework based on the Pearson correlation coefficient (PCC) and generalized learning system (BLS), which was compared with 10 machine learning algorithms to obtain the best performance and the highest model fitting capability. Singh et al. [25] utilized a method of combing $(2D)^2$PCA with deep learning to evaluate the stock. It was found that this method improved the evaluation accuracy by 4.8% compared with the Radial Basis Function Neural Network (RBFNN). Vargas et al. [26] focused on the structure of CNN and RNN to predict the standard intraday direction. Results showed that CNN can better capture semantic information in text, while RNN can better capture context information and model complex time characteristics for stock market forecasting. Because the trend of the stock market will ultimately be attributed to the discussion of time series, this paper chooses LSTM, a variant of RNN that is used to do time series investigation, analyze the stock market public opinion, evaluate its tendency and provide people with auxiliary investment reference.

3 Proposed Methodology

The overall frame structure is shown in Fig. 1. This paper analyzes and studies the stock market by feature fusion, including financial text and stock data features. Based on 1DCNN, firstly, the marked financial texts are considered to be the input of the classification model to realize the sentiment classification. Secondly, the text feature level fusion is realized according to its classification results, that is, the transformation from text data to numerical data (text sentiment value). Thirdly, the text sentiment value and the stock data are combined to get the time series data as the input of the public opinion analysis model. During this process, this paper adopts different durations and different dimension data for comparison. Finally, based on the analysis results, the stock trend is evaluated and some recommendable references are given.

3.1 The Algorithm Flow and Description

The flowchart is shown in Fig. 2. This paper mainly includes two types of features: one is the text features, and the other is stock historical data features. It is necessary to convert text into a vector as input because the text belongs to unstructured data. Text features and the technical features of stock data are the input to the model of this paper.

3.2 Web Crawler

IN this paper, As the starting point of this paper is to effectively evaluate the changes in the stock market in combination with the characteristics of online public opinion, which involves obtaining a large amount of information, such as post titles, post comments, and publishing time. As a data capture tool, the web crawler can quickly capture the specified information that needs to be obtained, so this paper uses web crawler technology to obtain text data related to the stock market. The frame of the web crawler in this paper is shown in Fig. 3. Its basic workflow is as follows:

1. Firstly, one or more web link Uniform Resource Locators (URLs) given in advance are taken as the initial page. In addition to text messages, web pages also contain hyperlinks, through which the web crawler system can get specific web pages. The web link given in this paper is determined based on the stock code, and the stock website is found according to the code.
2. Then, according to a certain network analysis algorithm, the links unrelated to the topic are filtered out, and the effective links are stored in the URL queue.
3. Finally, take the URL from the queue and download the corresponding web page. The corresponding time span in this paper is also set to crawl the information of a stock within a certain period. In this paper, we use the modules of //div [class = " articleh"] and //div [class = " articleh normal_post"] to get the number of the titles, the number of comments, the content of the title, the author and the time of posts. When accessing the content, the URL where the title is located is also stored in the Excel table, and an ID is set for each title.
4. In this paper, we need to grab the text data of five stocks. Thus, when finishing the data of the specified web page and the corresponding time span, we will analyze the next URL, put it into the queue to be grabbed, and enter the next cycle. Repeat the previous work until the conditions are met.

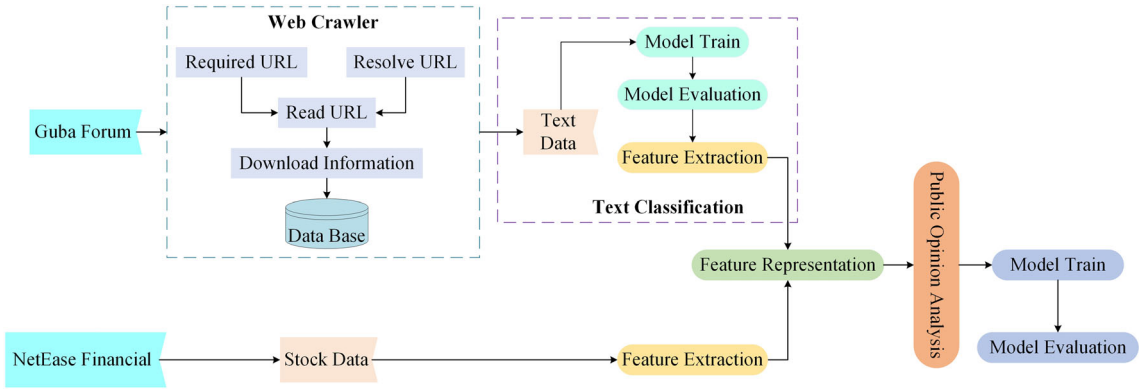

Fig. 1 Illustration of the basic frame structure. The data mainly include two parts: financial text and stock price data

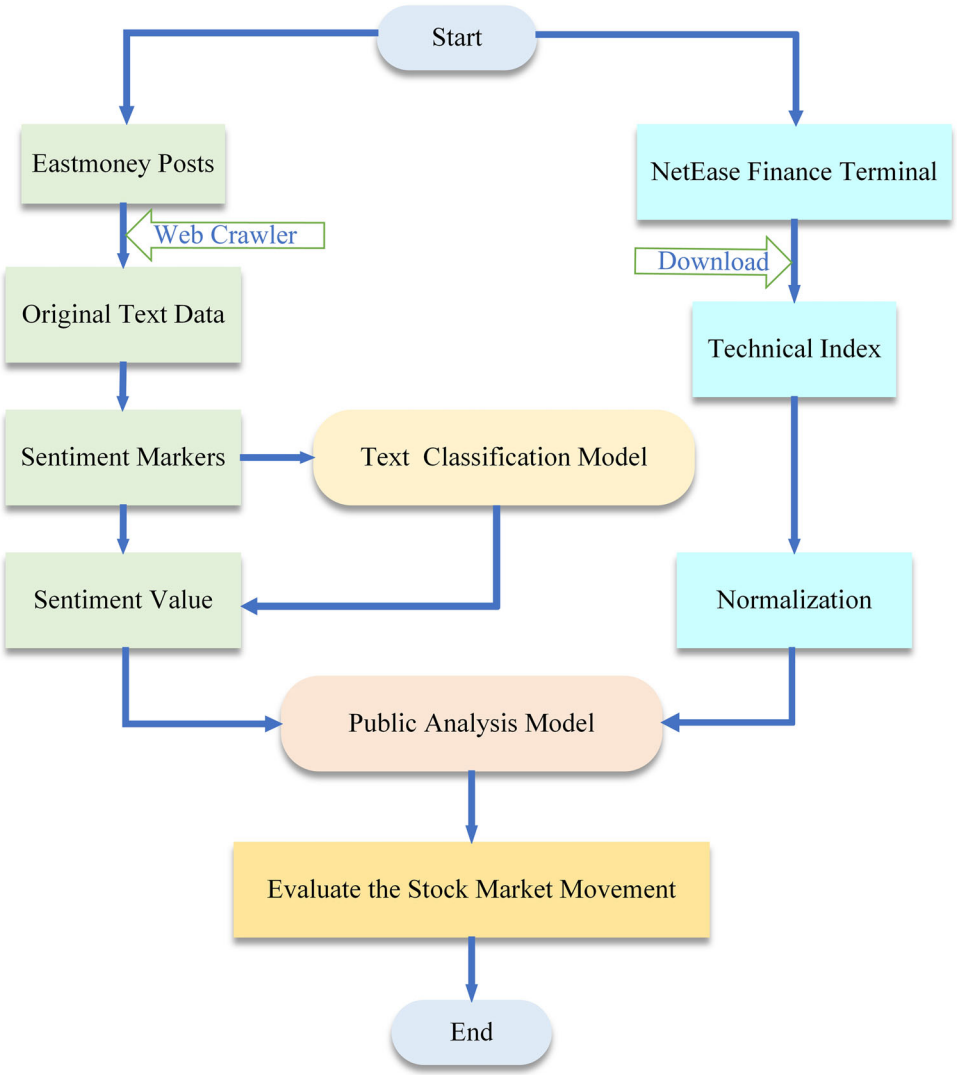

Fig. 2 The flow diagram. The whole process includes two parts: financial text classification and stock market public opinion analysis

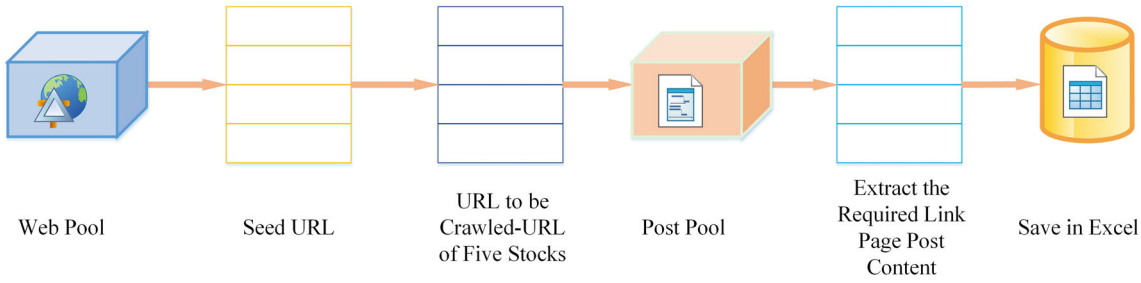

Fig. 3 Web crawler frame diagram

3.3 Data Collection

In this paper, we choose the time period from 2nd Jan 2019 to 29th Mar 2019 to carry out the experiments. The period includes 290 trading days in total. Five representative stocks are selected and each stock is allotted 58 trading days. The data we need include two parts: the stock-related trading data and the title from the stock forum, in which the stock data are obtained from the public data set and the text data are collected in the form of a web crawler.

3.3.1 Text Data

Figure 4 shows the flowchart of obtaining text data in our experiment. The text data can be obtained from various social media. In foreign countries, they mainly crawl Twitter [34]. The extracted content generally includes the time of posts, the number of followers and the content of tweets. However, in China, we mainly crawl the Guba forum or Sina Weibo. The information is generally the reading, comment number, post title, post comments, and release time. What's more, there are news data including ordinary news and financial news, such as Yahoo, the financial times and some domestic leading news sites [42–45]. In this paper, text data are obtained from Guba forum in the East money by the crawler. East money is a Chinese authoritative website among financial websites, with a large data flow and up-to-date information. There are many texts related to stock markets on this website. The text data information includes posts, post volume, attention, post time, user name and comments.

3.3.2 Stock Data

Stock data are generally obtained from stock exchanges. There are four stock exchanges in China including the Hong Kong Stock Exchange, the Taiwan Stock Exchange, the Shanghai Stock Exchange and the Shenzhen Stock Exchange. In addition, financial data can also be obtained by the Tushare package in Python. In this paper, we obtain stock data from the open dataset of the NetEase financial website for analysis.

Table 1 Basic features and the formulas

Feature	Formulas
Opening Price (O_t)	The first price trades during at t
Closing Price (C_t)	The final price at trades during t
Highest Price (H_t)	The highest price at trades during t
Lowest Price (L_t)	The lowest price at trades during t
Volume (V_t)	The numbers of the share during t
Price change	$C_t - C_{t-1}$
Volume change	$V_t - V_{t-1}$
Differential sequence	$r = \frac{C_t - C_{t-1}}{C_{t-1}}$

The stock data comes from five different industries: the real estate industry, the high-end equipment manufacturing industry, the smart industry, the new energy industry and the banking industry. The five stocks are representative stocks in the industry. They are Vanke A, Aerospace Science and Technology, HengBao Shares, Woer Heat-Shrinkable Material, and Industrial and Commercial Bank of China (ICBC), respectively. Due to the long names of Aerospace Science and Technology and Woer Heat-Shrinkable Material, we will abbreviate them to AST and WHSM, respectively, in subsequent writing. The stock data generally include the opening price, closing price, lowest price, highest price, and volume. The stock data comes from the stock trading records of the stock exchange market, and the price and trading volume of each transaction constitutes the basis of the stock data. There is no fixed time interval between each transaction. There may be only a few low-frequency stock transactions in an hour, while there may be dozens of high-frequency stock transactions in a second. To record these data, it is usually recorded at fixed time intervals in the field of securities. Table 1 is the description and formula of the basic characteristics of stock data. We obtain all the characteristic data in the table to form the experimental data set. t represents a given trading day, and t-1 is the nearest trading day before t. The basic features are the technical features mentioned above in Sect. 3.1.

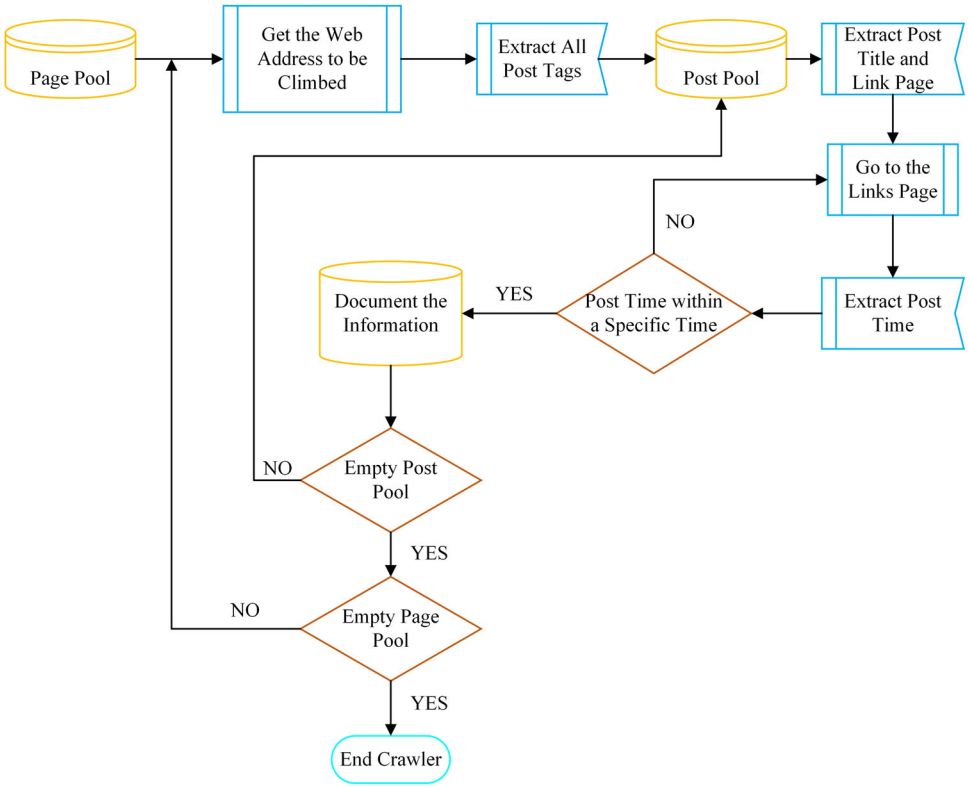

Fig. 4 The process of obtaining text data

3.4 Text Classification Model Based on 1DCNN

The types of text sentiment can be learned from the related literature. Text data are mainly divided into three or five categories. Three categories refer to bullish bearish and neutral [46, 47]. Five categories are strongly bullish, bullish, neutral, bearish, and strongly bearish [48]. Although the five types of sentiments are rare in the past arch, they clearly distinguish the emotional intensity. Hence, in this paper, to better understand the post sentiments, the post content is divided into five sentiments. Neutral refers to ambiguity (that is, bullish or bearish cannot be clearly expressed.) and some noise posts. To realize better calculation, STB represents strongly bullish, B represents bullish, H represents neutral, D represents bearish and STD represents strongly bearish in this paper.

Figure 5 gives the main model structure for text classification. The text of this paper is from the post title of the Guba forum of the authoritative website in China. Compared with English, its semantic grammar is more complex. In this paper, in order to pursue a network with low computing cost and superior classification performance, 1DCNN is adopted to realize text classification with embedded characters. Language, no matter what it is, is made up of characters. The characters used in this paper mainly contain 26 letters, 10 numbers, various other tag numbers, etc. After inputting the text, the model first constructs a vocabulary to form a character-level representation, and then uses one-hot encoding to quantify the characters. Text classification based on character embedding not only does not need to consider the single meaning of words (grammatical semantic information, but also can recognize the emoticons in the posts.

The model mainly includes an embedding layer, convolution, pooling, fully connected layer, etc. The sentence length is set to 100, and the word vector dimension is 50. That is, the embedding layer dimension is 100×50. The convolution kernel is 3, and the learning rate is set as 0.001. The main component of the model is the temporal convolutional module [49], which computes a 1D convolution in the text classification of this paper. Assume we have an input function $g(x) \in [1, m] \to \Re$ and a kernel function $f(x) \in [1, n] \to \Re$. The convolution function $Q(y) \in [1, \lfloor (m-n)/d \rfloor + 1] \to \Re$ determined by $g(x)$ and $f(x)$ with stride d is shown in (1).

$$Q(y)=(f * g) = \sum_{x=1}^{n} f(x) \cdot g(y \cdot d - x + c) \tag{1}$$

where $c = n - d + 1$ denotes an offset constant. $*$ represents the convolution operation. A set of kernel functions $f_{ij}(x)[(i = 1, 2, ..., l), (j = 1, 2, ..., k)]$ that we call weights are utilized to parameterize the module. l and k are the feature sizes of the input and output, respectively. g_i and

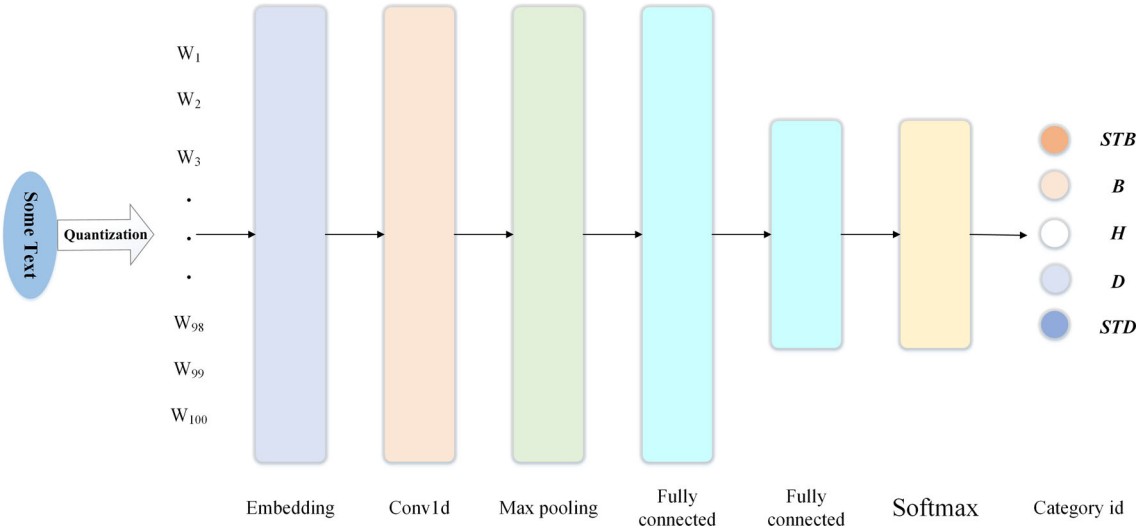

Fig. 5 Illustration of the model structure for text classification. The embedding, conv1d, max-pooling, fully connected and softmax are contained in the model. $\{W_i\}$ represents a vector with length of i for post

Q_j are the input and output features. The outputs of the module $Q_j(x)$ are gotten by a sum over i of the convolution between $g_i(x)$ and $f_{ij}(x)$.

To reduce the parameter number and avoid model overfitting, the temporal max-pooling is added behind the convolution layer. The 1D version of the max-pooling module in computer vision is used in the paper [49, 50]. Given an input function $g(x) \in [1, m] \rightarrow \Re$, the temporal max-pooling function $Q(y) \in [1, \lfloor (m-n)/d \rfloor + 1] \rightarrow \Re$ of $g(x)$ is defined as:

$$Q(y) = \max_{x=1}^{n} g(y \cdot d - x + c) \tag{2}$$

where $c = n - d + 1$ is an offset constant.

The cross entropy is adopted to calculate the loss in the classification model, and its formula is as follows.

$$L = \frac{1}{N} \sum_{i} L_i = \frac{1}{N} \sum_{i} - \sum_{c=1}^{M} y_{ic} \log(p_{ic}) \tag{3}$$

where M is the number of species. y_{ic} stands for the labels. If the class c is the same as the sample i, y_{ic}=1. Otherwise, it is 0. p_{ic} represents the prediction probability that the observation sample i belongs to the text category c.

Adaptive moment estimation (Adam) [51] is considered as the optimizer for this paper. The advantage of Adam is that after bias correction, each iterative learning rate has a certain range. It makes the parameters more stable. In our experiment, a minibatch of size is 64. We also insert 1 dropout module in between the 2 fully-connected layers for regularization. They have the dropout probability of 0.5.

Finally, softmax is used to calculate the feature of text classification. The calculation formula is:

$$X = \text{softmax} = e^{z_i} / \sum_{k=0}^{4} e^{z_k} \tag{4}$$

where z is the output of the fully connected layer. $i = 0$, 1, 2, 3, 4 stands for five categories (STB, B, H, D, STD) in this paper and X represents the probability of the text classification output.

Since the data selected in this paper is a quarterly data, the text data of two months is taken as the training set, and the data of a month is used as the test set. During the training process, the training parameters of 1DCNN are backpropagated to reduce the error between the predicted value and the real value.

3.5 Public Opinion Analysis Model

The main structure of the public opinion analysis model is shown in Fig. 6. RNN is frequently used in time series analysis and other tasks [52, 53]. RNN is considered to be recurrent because the current sequence depends on the previous sequence and they are related to each other. In our dataset, we need a continuous time series as the input feature. However, RNN will not tackle the long-term dependence due to gradient problems. LSTM and Gated Recurrent Units (GRU) are proposed to solve the issue by employing gate control mechanism. As variants of RNN, they can make up for the loss caused by gradient disappearance to a large extent. In our model, we choose the LSTM because it has better performance in processing time series.

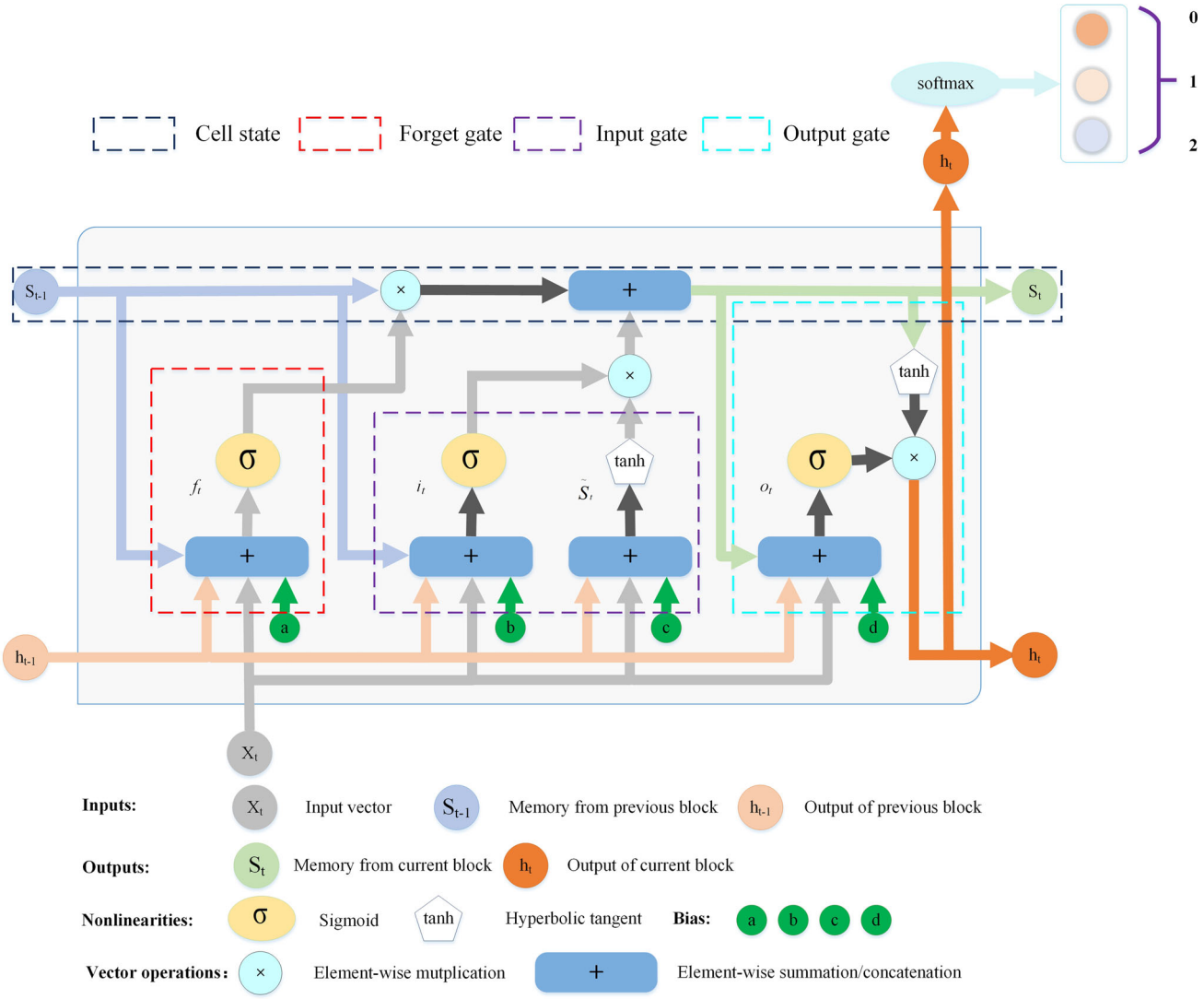

Fig. 6 The main structure of public opinion analysis model. The type of the public opinion data belongs to time series. S_{t-1} stands for the previous time series information, and h_{t-1} represents the output of previous time series value. S_t is the current time series information we need to process, and h_t means the current result of the time series. **a**, **b**, **c,** and **d** are the offsets in different processes. The blue dotted box is cell state and the red dotted box indicates the forget gate. The purple represents the input gate, which contains old and new memories. Sky Blue stands for the output gate, which is utilized to output the results of public opinion analysis. y_t is the classification results through the softmax layer. 0, 1, and 3 are the public opinion categories

At time t, the calculated process of the LSTM is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

$$\tilde{S}_t = \tanh\left(W_S \cdot \left[h_{t-1}, x_t\right] + b_S\right) \tag{8}$$

$$S_t = f_t \cdot S_{t-1} + i_t \cdot \tilde{S}_t \tag{9}$$

$$h_t = o_t \cdot \tanh(S_t) \tag{10}$$

where f_t represents the forget gate. It determines the state of the unit at the previous moment S_{t-1}. i_t stands for the input gate and o_t is the output gate (as shown in Fig. 6). $\tilde{S}_t$ indicates new memory cell and S_t is the current public analysis information. W and b represent the weights and offsets in different operations, respectively.

Among the model of public opinion analysis, softmax is used to calculate the feature of public opinion. The calculation formula is:

Table 2 Setting of experimental parameters of the proposed method

CNN		RNN	
Parameter	Value	Parameter	Value
Number of epochs	50	Number of epochs	30
Batch	64	Batch	128
Learning rate	1e-3	Learning rate	1e-3
Optimizer	Adam	Optimizer	Adam
Loss	Cross Entropy	Loss	Cross Entropy
Output of the first fully connected layer	128	Number of hidden layers	2

$$Y = \text{softmax} = e^{z_j} / \sum_{k=0}^{2} e^{z_k} \tag{11}$$

where z is the output of the fully connected layer. $j = 0, 1, 2$ stands for down or flat or up and Y represents the probability of public opinion analysis output.

The loss function used in the public opinion analysis model is shown in (3). The optimizer is Adam, which is consistent with the text classification model. In the experiment, the ratio coefficient of train set and validation set is 0.85. The number of hidden nodes set at the beginning of this paper is 128. The range of learning rate is 0.001 ~ 0.000001, and the value of batch size is between 4 and 100.

4 Experiment and Results Analysis

IN this section, we discuss and compare several relevant methods. The time of experimental data (forum posts and historical stock data) is from 2nd Jan 2019 to 29th Mar 2019, the research content of this paper is divided into stock text classification and public opinion analysis, so we adopt CNN combined with 1DCNN to classify financial text, and then adopt RNN combined with LSTM to complete the public opinion analysis experiment, we list the parameters of the experiment in Table 2. The CNN and RNN adopt the same loss function and optimizer as in the literature [54, 55, 59]. In addition, the number of fully connected output channels in the first layer of the 1DCNN of the CNN model is the same as that of the literature [54], and the number of hidden layers of the LSTM in the RNN is the same as that of the literature [59]. In addition, parameters such as learning rate and period need to be adjusted to select the appropriate parameter values. Therefore, we adopt a 60% subset of the dataset as a training set to learn the sample fitting parameters, a 20% subset as a test set to evaluate the model performance and a 20% subset as verification set to find the appropriate parameter set. All the experiments were performed on a regular workstation (CPU: Intel(R) Core (TM) CPU i7-8700 CPU @ 3.20 GHz; GPU: GTX1070Ti; RAM: 32.0 GB).

4.1 Data Preprocessing

4.1.1 Text Data

The text needs to be preprocessed before classifying using networks, mainly including emotional marking and word vector conversion. As shown in Table 3, these are some examples of sentiment marking text from Guba. To distinguish sentiment better from the visual point of view, the core words are marked in the table. In our experiments, firstly, the text is manually marked, and then, the character embedding method is used to detect the accuracy of classification. The total number of processed data is 9066 (Among them, STB is 482, B is 1830, H is 4348, D is 2012 and STD is 394). The period of each stock is from January 2, 2019 to March 29, 2019 (290 trading days in total).

After the results of sentiment classification in this paper, we need to calculate the sentiment of each emotion. The sentiment feature is as follows:

$$P_j = \frac{P_t}{P_{\text{STB}} + P_{\text{B}} + P_{\text{H}} + P_{\text{D}} + P_{\text{STD}}} \tag{12}$$

where t is the types of sentiment. P_j represents the value of sentiment t on the day j. P_t, P_{STB}, P_{B}, P_{H}, P_{D}, and P_{STD} indicate the number of each sentiment on the day j, respectively.

The overall day sentiment is as follows:

$$P = \text{Max}\left(P_j\right) \times w_i \tag{13}$$

in which w_i denotes the weights of STB, B, H, STD, D, respectively, + 2, + 1, 0, -1, -2. The value of the weight is assigned by the reference [33]. The neutral sentiment is generally not included in the calculation of sentiment value. After calculating the mood value, it needs to be normalized by (14).

4.1.2 Stock Data

1) Normalization

Because stock data changes greatly, the trend of stock data varies across industries. To compare with different stocks, the stocks need to be generally standardized after acquisition. The commonly used normalization methods include minimum–maximum normalization, decimal scaling normalization and Z-score normalization. In this paper, we adopt the most widely used minimum–maximum normalization,

Table 3 Text examples from the Guba forum

Document name	Stock type	Text categories	Text
2019-03-18	Vanke A	**STB**	Weiwu is **full of warehouses** today, and the continuous trading mode is about to start crazily
2019-03-18	Vanke A	**STB**	It seems that Vanke will **go up sharply** as the big order keeps advancing
2019-03-22	HengBao Shares	**H**	Hengbao's funds revealed on March 22nd
2019-03-28	ICBC	**B**	Tomorrow I'm going to **add** warehouses [laughter] [laughter]
2019-02-28	HengBao Shares	**D**	Comrades **left** the market
2019-02-11	WHSM	**H**	Woer Heat-Shrinkable Material disclosed the fund on February 11
2019-01-31	WHSM	**STD**	Woer Heat-Shrinkable Material **fell 5%** on January 31
2019-01-25	ICBC	**D**	**Down** on Monday!
2019-01-07	AST	**STD**	I've **cleared the warehouse**. There's no time to watch a dealer show when there's opportunity out there
2019-01-03	AST	**B**	There is also the possibility of **rising**

In order to facilitate reading and avoid format problems, the Chinese text samples in the form have been translated into English ones. But in the experiments, we still select Chinese text samples as the experimental materials

and its transformation function is as follows:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{14}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the sample data, respectively.

2) Stock data classification

In stock trend evaluation, it is of more practical significance to estimate whether the stock trend will rise or fall sharply than only to evaluate the rise and fall. Therefore, we set up three categories of stock data in this paper. The specific classification standard is shown in (15).

$$\text{class} = \begin{cases} 0, & r < -1\% \\ 1, & -1\% \le r < 1\% \\ 2, & r \ge 1\% \end{cases} \tag{15}$$

where 0 stands for a sharp drop, 2 is a sharp rise and 1 represents a gentle trend (r is the differential sequence as mentioned in Table 1). The rise and fall ranges are divided according to the stock data interval. It can be seen from the Fig. 7 that the price range of five stocks is generally distributed in – 5 and 5%, and the description of rising and fall in the text data is generally between – 5 and 5%. Hence, to balance the sample, this paper divides the stock data into three categories, with − 1 to 1% as the boundary. Figure 8 shows the sample sizes after classification. It is basically balanced, and only the number of samples with category 0 is less.

4.1.3 Public Opinion Data

Public opinion data include text and stock data. Through different processing methods, the data are divided into five-dimensions: 3, 4, 5, 6, and 7. The time span is divided into 3 days, 5 days, 7 days, 10 days, and 15 days. They are considered as the input to the public opinion model.

The 3D data is mainly realized by the weight assignment of text and transaction data, and the formula involved is:

$$Data_{3\text{ - dimension}} = \left(text/2,\ (T_1 + T_2)/2,\ r\right) \tag{16}$$

where $T_1 = |O_t - C_t|$. $T_2 = |H_t - L_t|$. O_t and C_t are the opening and closing prices, mentioned at time t in Table 1. H_t and L_t are the highest price and the lowest price at time t. T_{1i} and T_{2i} represent the difference of the corresponding stock prices at the time i, respectively. $text$ is the sentiment value of text and r_i is the difference sequence at the time i mentioned in Table 1.

The 4D data increase the stock trading volume based on the three-dimensional one. The calculation formula is defined as.

$$Data_{4\text{ - dimension}} = \left(text/2,\ (T_1 + T_2)/4,\ V/4, r\right) \tag{17}$$

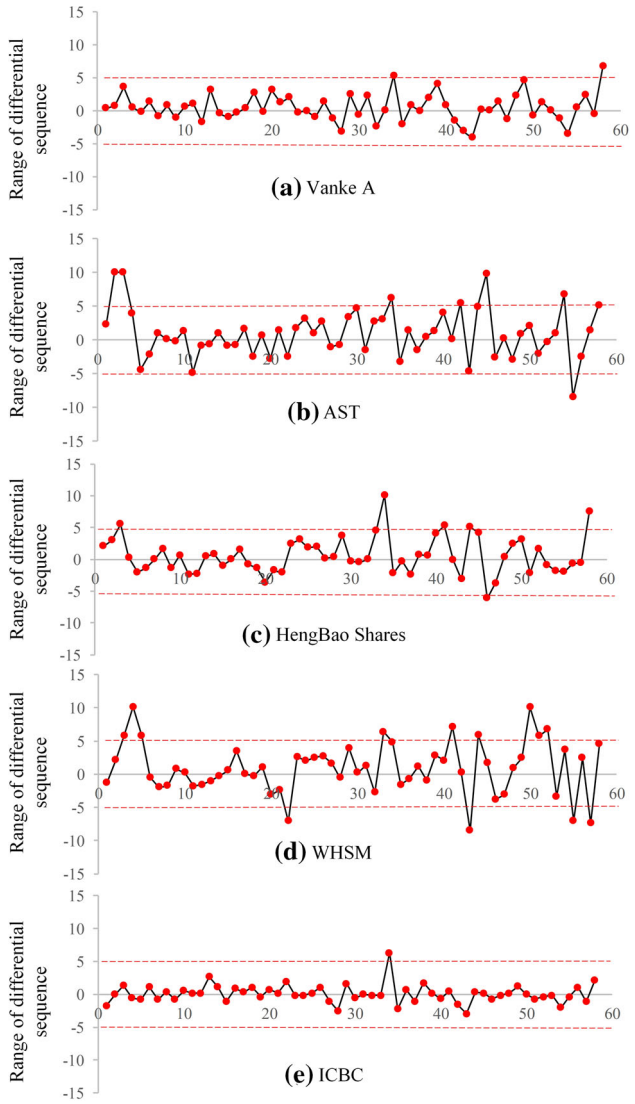

Fig. 7 Differential sequence distribution. **a–e** are the different sequence points of the five stocks, respectively. From the vertical axis of the figure, it is found that the variation range is mainly between − 5 and 5%

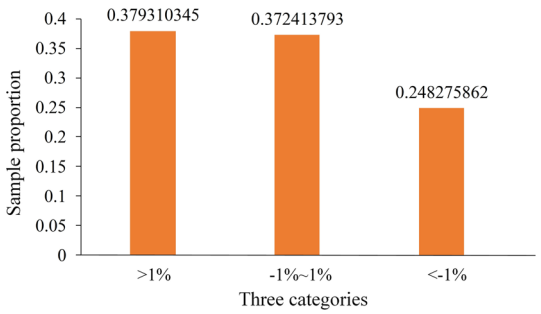

Fig. 8 Sample proportion of stock data with three classifications

Table 4 The proportion of correct samples

Text Categories	T_{correct} (%)
STB	69.11
B	60.68
H	83.97
D	73.07
STD	63.00
Average T_{correct}	70.00

where V_i stands for the stock volume at time i mentioned in Table 1.

The five-dimensional data consist of stock opening price, closing price, highest price, lowest price and differential sequence. The 6D data increase text data. The 7D data add the trading volume of stocks based on the six-dimensions.

4.2 Text Classification

4.2.1 Results and Analysis of Confusion Matrix

IN order to observe the number of misjudged categories in text classification, the confusion matrix is considered as a standard format for accuracy evaluation. Figure 9 shows the confusion matrix result of text classification, which reflects the accuracy of classification from different aspects.

The correct decision rate of each class is defined by the following formula:

$$T_{\text{correct}} = \frac{n_{\text{correct}}}{n_{\text{all}}} \times 100\% \tag{18}$$

where n_{correct} represents the number of correct categories, and n_{all} stands for all sample.

Table 4 shows the results of correct samples in text classification according to (18). The higher the proportion of the text category, the less likely it is to be misclassified into other categories. From Table 4, the proportion of H is the highest, and its T_{correct} is 83.97%. Simultaneously, it is seen from Fig. 9 that the number of correct samples of H is 812, accounting for the largest proportion of the whole correct categories. However, the average classification proportion of the other four categories is relatively low. The main reason is that the text data itself is unbalanced, the neutral sample data accounts for the largest proportion, and the number of other types of samples is less.

4.2.2 Comparison of Experimental Results of Different Methods

Table 5 shows the results of text classification via different methodologies. Since the text category in this paper is

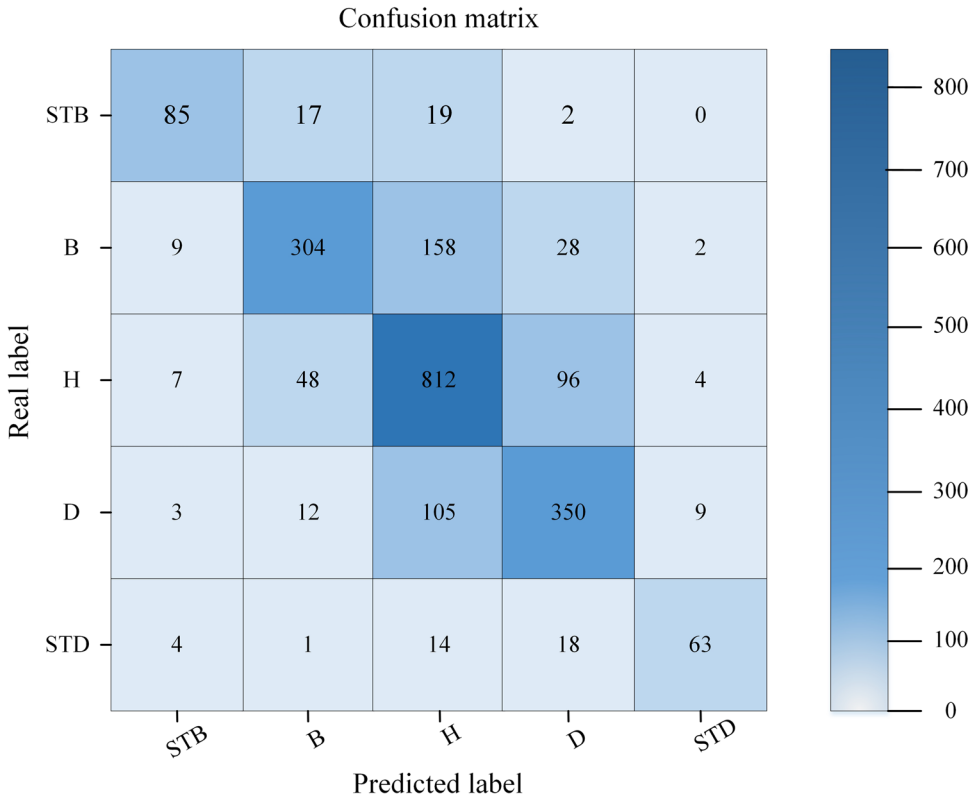

Fig. 9 Confusion matrix of the proposed method. As the number of correct categories increases, the gradient color in the image darkens. Diagonals are the number of correct classifications for each category

Table 5 Text classification accuracy of different models

Model	Accuracy	Precision	Recall	F1
RNN [56]	0.6926	0.664	0.634	0.646
LSTM + Attention [57]	0.6912	0.698	0.6420	0.6664
GRU + Attention [58]	0.6972	0.7265	0.6382	0.6743
CNN + Word2vec	0.6198	0.6520	0.5581	0.5942
CREST NSE [59]	0.7023	0.7345	0.6512	0.6903
CNN (This paper)	**0.7438**	**0.7683**	**0.7000**	**0.7280**

Bold values highlight the experimental effect of the proposed method

manually labeled based on the relevant financial information, text classification models are commonly used to detect the accuracy, to ensure the accuracy of the classification and the follow-up experiments.

As shown in Table 5, this paper conducts comparative experiments according to five methods: Literature [56], Literature [57], Literature [58], Literature [59] and 1DCNN + Word2vec. Among the five methods, documents [56] and [59] use the original recurrent neural network for classification, and documents [57] and [58] are all text classification realized by using the variation of recurrent neural network and adding attention mechanism (by enhancing the function of feature extraction, thus improving the model effect) in the feature extraction part, in which the variation improves the deficiency of the original recurrent neural network to some extent. 1DCNN + Word2vec is a word vector processing model which adds Word2vec to one-dimensional convolution. Through the comparison of four methods to verify the effectiveness of this method, the input data of each network model has undergone the same preprocessing. Compared with the literature [56], literature [57], literature [58], literature [59] and 1DCNN + Word2vec, the classification model in this paper has obvious advantages in four indexes. Among them, the average accuracy of the proposed method is relatively improved by 4%, the precision is relatively improved by 4%, the recall is relatively improved by 5% and the F1 is relatively improved by 4%. In view of the classification model of the same dataset, our proposed method has better performance and classification accuracy, which is more worthy of praise.

4.3 Experiment Results of Public Opinion Analysis

1) Comparison of experimental results of different data characteristics.

Table 6 shows the experimental parameter settings and experimental results of public opinion analysis data with different dimensions and different time spans. There are five groups of data with different dimensions (three-dimensional, four-dimensional, five-dimensional, six-dimensional, and seven-dimensional), and each group of data contains five characteristic data with different continuous times (3 days, 5 days, 7 days, 10 days, and 15 days). The experimental parameters include learning rate, number of hidden layers, number of hidden nodes and activation function, and the experimental accuracy of each case is tested.

It is seen from Table 6 that compared with the experimental results of five-dimensional data without text information, other experimental results with additional text information are relatively good. Therefore, in theory, increasing the text information (public opinion information) can improve the evaluation level and provide a certain basis for the evaluation of the stock market trend. That is, we can use the text information on the Internet as a reference index for subsequent decision-making. At the same time, we can also add a guarantee for our investment. Secondly, it can be seen that the time span of different dimensions has different characteristics. Among the three-dimensional data, the time span is 3 days (average accuracy rate is 52.76%), four-dimensional data are 15 days (55.15%), five-dimensional data are 3 days (48.82%), six-dimensional data are 15 days (average accuracy rate is 54.55%), and seven-dimensional data are 10 days (average accuracy rate is 57.33%) with the best experimental results.

2) Experimental Analysis of Stock Samples with Different Industry Representatives.

Table 7 shows the representative stock-related information of the intelligent industry, including text samples and stock-related technical indicators (O_t, C_t, H_t, L_t, and r represent opening price, closing price, highest price, lowest price, and differential sequence, respectively.). Considering the space, the stock data of the other four industries are shown in appendixes A–D, respectively. T and P represent the real and predicted result of the public opinion analysis in the Table, respectively. The text samples are the randomly selected 5-day data, five text samples per day. The prediction value is realized based on the text and stock price data. The reason why category 0 is wrongly predicted as category 2 on the second day of the table is that (1) the interference caused by strong bullish and bullish samples in the text data; (2) the unbalanced samples. The data selected in this paper are quarterly data, and the sample number of category 0 is relatively small. On the third day, category 1 was predicted as category 2 mainly because compared with category 2, the sample number of category 1 was relatively small, which led to bias in emotional tendency. In addition, the common grounds of prediction errors are: (1) time interval; (2) data dimension; (3) the characteristics of text and stock price data and their weights distribution.

The above experimental results show that we cannot just rely on a single text feature or stock feature to evaluate the effect of the experiment, but analyze the result by combining with it different characteristics and data existing way (time span and data dimension). Consequently, under the condition of sufficient sample data, whether it is feature information or data representation ways, the experimental results can be better only within the appropriate limits.

4.4 Analysis of Stock Market Public Opinion

In our experiments, it is found that the experimental results with text information are better than those with single stock market trading information. The public opinion analysis of the stock market refers to the use of text information on the Internet to explore the trend of the stock market, that is, whether the text can be used as an indicator of the stock market evaluation. By observing the users of fortune.com, we find that in the stock market, whether online or offline, forums like “Guba” have become a platform for traders to obtain information, forming a large-scale social network. In this network, users interact with each other by adding friends and forwarding comments. Among them, text information is the most important basis for transmitting information. It will resonate with the content of comments and make certain information more inclined to comment on the author. Hence, text information uses the impact of user behavior in social networks to a certain extent, thereby further mapping the stock market price changes. Theoretically speaking, adopting textual information as an opinion indicator for stock market evaluation. Across the board, the sentiment changes corresponding to the text affect the trend of the stock market to some extent and can also supply investors with some constructive recommendations (people can reasonably make use of the changes of network sentiment and stock market historical data to make decisions). For example, when the sentiment of text information on a certain day is optimistic about the future market, investors can buy an appropriate share of the corresponding stock under the premise of comprehensive consideration. Otherwise, if the sentiment is negative, the investors will sell it.

Table 6 Parameter setting and accuracy of different dimensions of public opinion analysis experiment

Data dimension	Time span	Learning rate	hidden layers	Hidden nodes	Activation function	Accuracy (%)
Three-dimensional	3	1×10^{-5}	1	60	Sigmoid	52.76
	5	1×10^{-5}	1	50	Sigmoid	51.85
	7	1×10^{-6}	1	20	Sigmoid	51.45
	10	1×10^{-5}	1	50	Sigmoid	48.00
	15	1×10^{-5}	1	50	Sigmoid	43.03
Four-dimensions	3	1×10^{-7}	1	128	Sigmoid	48.03
	5	1×10^{-6}	1	60	Sigmoid	54.81
	7	1×10^{-6}	1	30	Sigmoid	52.90
	10	1×10^{-5}	1	50	Sigmoid	49.33
	15	1×10^{-6}	1	80	Sigmoid	55.15
Five-dimensions	3	1×10^{-5}	1	60	Sigmoid	48.82
	5	1×10^{-4}	1	60	Sigmoid	42.96
	7	1×10^{-6}	1	60	Sigmoid	47.83
	10	1×10^{-6}	1	60	Sigmoid	46.67
	15	1×10^{-5}	1	80	Sigmoid	47.27
Six-dimensions	3	1×10^{-5}	1	40	Sigmoid	53.54
	5	1×10^{-6}	1	100	Sigmoid	51.85
	7	1×10^{-6}	1	40	Sigmoid	48.55
	10	1×10^{-5}	1	60	Sigmoid	47.33
	15	1×10^{-4}	1	60	Sigmoid	54.55
Seven-dimensions	3	1×10^{-5}	1	25	Sigmoid	49.61
	5	1×10^{-6}	1	60	Sigmoid	47.41
	7	1×10^{-6}	1	30	Sigmoid	52.17
	10	1×10^{-6}	1	60	Sigmoid	57.33
	15	1×10^{-5}	1	60	Sigmoid	52.73

5 Conclusions and feature work

The change of the stock market plays an important role in the trend of the national economy. With the popularity of artificial intelligence, the future study of the stock market has become a hot topic. Thus, the work in this paper is significant. The rise of the Internet has brought more and more attention to the stock market, giving investors the space to speak freely. According to these, this paper proposed a public opinion analysis framework of stock market, and compared the influence of different dimensions and different time spans on the results of the experiments. The experiment mainly includes two parts.

Firstly, for text data, we obtained the required public opinion data (corresponding to the text data of five leading stocks in different industries) through certain rules of crawler technology set in this paper, then cleaned the useless information in the text and manually labeled the text emotions, and finally applied the designed 1DCNN to classify text sentiment. In the data crawler, we adopted the regular expression Beautiful Soup in Python to get the data we really need. For text sentiment classification, we compared the experimental results of different models, as well as the results of different processing methods from the same model, and finally proposed a text sentiment classification model. The accuracy of our model is 74.38%, which is better than the other ones, and proves the effectiveness of the proposed method.

Secondly, this paper puts forward the public opinion analysis framework of the stock market. The above-mentioned text data is added to the stock market trading data. The sentiment value of text data mapping and trading data of the stock market are combined as the input features of the analysis model. In the public opinion analysis experiment, we compared the effects of input characteristics of different dimensions (they are composed of Chinese texts, stock price, and stock trading volume that have different weights. Among them, stock price includes opening price, closing price, highest price, and lowest price) and different time spans.

Table 7 The representative stock-related samples information of the intelligence industry

Stocks types	Text samples	Stock-related indicators					T	P
		O_t	C_t	H_t	L_t	r		
002104- HengBao Shares	STB-Hengbao shares rose 5% on March 11 STB-Hengbao shares rose rapidly on March 11 B-Buy 002104 B-The trading is limited today B-Now wash the plate, and immediately limit the trading	7.68	7.87	7.94	7.58	5.07	2	2
	STD-It will continue to drop more than 5 points tomorrow STB-Today, this turtle grandson finally soared D-It seems to have dropped to the limit STD-Clear the warehouse at 8.2 yuan D-Quickly sell it, it is going to limit down tomorrow	8.1	7.7	8.3	7.65	− 6.10	0	2
	B-The one card e-commerce innovation board is really good for Hengbao H-Catch it under seven yuan, please wait B-Financing to buy 002104 target price of 20 yuan D-Today, the constant God falls frighteningly D-This plunge is very good!	7.5	7.44	7.57	7.26	0.40	1	2
	B-Today's statement should be like this: there is no purchase in the end of the day, and the trading limit will be clearly set next Monday D-This cancer stock is undergoing chemotherapy every day. It's the only one in the world D-While there is no drop sharply, we should change industrial marijuana and seize the opportunity D-Mud is mud forever! Out! H-It closed around 7.60	7.82	7.75	7.85	7.55	− 0.90	1	1
	D-Today, the stock market is down, everyone is very excited. Is it Xiaosan who cut the leeks of Zhuang again? STD-When the price is 7.69, the warehouse was cleared completely. I would rather go empty than be covered D-Market will quickly down to 2800, Hengbao directly to about 6.7. Get ready D-It's starting to pull again. You can throw a little B-Strong Hengbao shares; There is room to rise	7.68	7.46	7.81	7.4	− 1.97	0	0

In order to facilitate reading and avoid format problems, the Chinese text samples are translated into English ones. The stock data of the other four industries are shown in the appendixes A–D, respectively

The methods proposed in this paper also have limitations. For the time series characteristics of stock price, the threshold we set in the process of finding trend characteristic points is affected by the receipt set, and manual setting will be more complicated. If you can automatically update the threshold, you can get twice the result with half the effort. As the text data used in this paper is the title of the forum, some titles may be incomplete or short, which are actually irrelevant to our research context. We calculate the stock sentiment intensity based on daily text. The incomplete or short text would affect the accuracy of the sentiment classification algorithm and the evaluation of stock market trends. Therefore, there is still some room for improvement in future. Adding targeted information sources (such as comments corresponding

to titles, multiple social media data or stocks of the same type) and adopting other feature fusion mechanisms and weight distribution methods will further improve the experimental results, enhance the authority and reliability of the proposed methods, and prepare for further research.

Acknowledgments This work was supported in part by the Hunan Provincial Natural Science Foundation of China (Grant No. 2022JJ31022) and the Undergraduate Education Reform Project of Hunan Province (Grant No. HNJG-2021-0532). The authors are grateful for this support.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendices

Appendix A

Stocks types	Text samples	Stock-related indicators					T	P
		O_t	C_t	H_t	L_t	r		
000002-Vanke A	STB-Is the spring of real estate stocks coming? STB- It seems that Vanke is going to go up a lot B-Vanke monthly line is very stable, continue to increase! H-Come on, Vanke. Keep up with Poly Real Estate D-28.67 sell half and revive first	28.06	29.37	29.42	27.76	4.63	2	2
	B-Vanke's stock price will double in one year. It's up to you to improve the economy B-It's time to celebrate the trading limit STB-Buy boldly B-0320 comments: continue to shock, we have added warehouses, and are ready to wait for the profit D-Determined to sell Vanke	29.15	29.53	29.76	28.65	1.30	2	2
	STB-Run into the market and the stock will rise tomorrow D-Unlimited draw back to lure more. In the afternoon, the stock diving, mainly watching! B-Hold D-Sold Vanke because it was too junk H-When is the dividend?	28.51	28.29	28.96	28.21	0.46	1	2
	B-Be optimistic about Vanke A B-Vanke trend stable, optimistic about Vanke! B-Performance continued to grow steadily H-Vanke's rise is often quiet B-Good news one by one	28.55	28.95	29.05	28.23	2.333	2	2
	STB-100 points skyrocketing! Can the strong market last? STB-Vanke rose again, rising 40 points B-Vanke is great STB-In March 29, Vanke A disk rose 5% B-Today is higher than before	28.8	30.72	30.75	28.75	6.67	2	2

In order to facilitate reading and avoid format problems, the Chinese text samples are translated into English

Appendix B

Stocks types	Text samples	Stock-related indicators					T	P
		O_t	C_t	H_t	L_t	r		
000901-AST	B-The space sector will rise by 1000 yuan in future STB-Week month line just full cross shareholding B-Super bull fork, keep the first-line shares for trading STB-Aerospace Science and Technology rose rapidly on March 12 B-In the momentum, the stock is ready to rise	13.95	15.18	15.2	13.7	9.84	2	2
	D-I just bought it on 14.8, and when I bought it, it fell down. Now, I feel very sad B-The gun rack can't fall down. There will be a good play tomorrow D-A fool can see that he is luring more people H-Aerospace Science and Technology rebounded rapidly on March 14 D-It stops in the afternoon	14.6	14.84	15.3	13.5	0.39	1	1
	B-The opening price will rise next week H-There's no problem with stocks. There's a problem with the makers H-When will it go up today? D-It's time to make up for the drop H-Waste of good market	14.99	14.42	15.05	14.31	− 2.83	0	0
	B-It's really weird that the price limit will be raised after 15 yuan B-You go up slowly STB-Full warehouse D-It's a good choice to buy 316 STB-Aerospace Science and Technology rose rapidly on March 25	14.7	15.7	16.17	14.55	6.80	2	2
	D-I don't think so good and I cut it! I'll come back in when it pulls back 20% or so B-Buy one at 14.00 STD-It's frightening to fall like this. I don't know where the bottom is D-Drop to 60 days' line support! D-Fall did not resist, rise is all pressure, and the stock broke the 20 support line. It is bearish in the short term!	14.4	14.04	14.48	13.69	− 2.36	0	1

In order to facilitate reading and avoid format problems, the Chinese text samples are translated into English

Appendix C

Stocks types	Text samples	Stock-related indicators					T	P
		O_t	C_t	H_t	L_t	r		
002130-WHSM	STB-It can play 250 million changed hands, and the turnover of 6, such a bull market model, I am afraid you do not limit trading? B-Plate tomorrow! STB-The situation is very good, full orders are all fast to buy STB-Nuclear power stocks are mostly held by retail investors, so they will rise sharply B-Don't sell it if you take it. It will be more than 15 yuan this year	4.8	4.75	4.87	4.59	1.71	2	2
	STD-In this way, clearance at the price of 4.8 is still the right choice D-Too weak, junk stock STD-Let's run quickly, and the stock will limit down in the afternoon STD-Clear the warehouse! D-The stock is rubbish	4.58	4.43	4.61	4.35	− 3.06	0	0
	B-Congratulations on Wall nuclear the rising by the daily limit B-I bought it when I fell two points B-Congratulations on Wall nuclear trading STB-The stock bar is clamoring for us to raise funds into our stock market fully. Today, we are making a profit of 300 W, and we can't sell it out in less than 12 yuan STB-Nuclear explosion in the afternoon	5.18	5.69	5.86	5.1	6.75	2	2
	STD-Three outlooks are destroyed by a huge loss D-If there is a visible small drop in the end of the day, you will look forward to tomorrow B-I'm sure it's going to take a pull today. Hold on STD-Wall nuclear March 26 intraday decline of 5% STD-It closed down more than eight points today	5.75	5.3	5.76	5.26	− 7.02	0	0
	STD-Three outlooks are destroyed by a huge loss D-Tomorrow, we will continue to drop the limit STD-Wall nuclear March 28 intraday decline of 5% B-I'll add funds first STD-Run, run fast	5.31	5.03	5.46	5	− 7.37	0	2

In order to facilitate reading and avoid format problems, the Chinese text samples are translated into English

Appendix D

Stocks types	Text samples	Stock-related indicators					T	P
		O_t	C_t	H_t	L_t	R		
601398-ICBC	B-I'll sell out of Maotai to buy ICBC, tomorrow D-The bank's worst stock B-Buy ICBC at 3 pm! STB-Today, ICBC soared! B-This week is the end of the rally, and bank stocks are bound to rise	5.57	5.65	5.65	5.56	1.25	2	2
	B-Buy some D-Let's have a try with light warehouse B-We can take off after 65 D-Cut all at 5.64, and exchange shares of China Merchants Bank H-ICBC follows the market trend	5.63	5.61	5.65	5.59	− 0.71	1	1
	D-ICBC's price has dropped to 2500 points D-ICBC's trading volume fell STD-Banks can't handle the big drop B-Start buying! STD-The over falling stocks banks began to be paid attention	5.53	5.47	5.55	5.46	− 1.97	0	0
	B-I add funds again in the warehouse STB-ICBC leads the market, rising continuously! D-There are too many orders to sell B-Cut the meat and it goes up STB-ICBC soars!	5.46	5.51	5.54	5.45	1.10	2	0
	STB-It will continue to soar next week STB-Financing is full at the price of 5.54 B-Buy in! B-The fundamentals are stable, and the asset quality continues to improve D-Cut a storehouse	5.47	5.57	5.58	5.46	2.20	2	1

In order to facilitate reading and avoid format problems, the Chinese text samples are translated into English

References

1. Nassirtoussi, A.K.; Aghabozorgi, S.; Wah, T.Y.; Ngo, D.C.L.: Text mining for market prediction: a systematic review. Expert Syst. Appl. **41**(16), 7653–7670 (2014). https://doi.org/10.1016/j.eswa.2014.06.009
2. Yang, K.; Yi, J.; Chen, A.; Liu, J.; Chen, W.: ConDinet++: full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images. IEEE Geosci. Remote Sens. Lett. **19**, 8015105 (2022). https://doi.org/10.1109/LGRS.2021.3093101
3. Thakkar, A.; Chaudhari, K.: Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks. Appl. Soft Comput. **96**, 106684 (2020). https://doi.org/10.1016/j.inffus.2020.08.019
4. Yang, K.; Yi, J.; Chen, A.; Liu, J.; Chen, W.; Jin, Z.: ConvPatchTrans: a script identification network with global and local semantics deeply integrated. Eng. Appl. Artif. Intell. **113**, 104916 (2022). https://doi.org/10.1016/j.engappai.2022.104916
5. Patel, D.; Thakkar, A.: A survey of unsupervised techniques for web data extraction. Int. J. Comput. Sci. **6**(2), 1–3 (2015)
6. Akyol, K.; Sen, B.: Modeling and predicting of news popularity in social media sources. Comput. Mater. Contin. **61**(1), 69–80 (2019). https://doi.org/10.32604/cmc.2019.08143
7. Galicia, A.; Talavera-Llames, R.; Troncoso, A.; Koprinska, I.; Martínez-Álvarez, F.: Multi-step forecasting for big data time series based on ensemble learning. Knowl-Based Syst. **163**, 830–841 (2019). https://doi.org/10.1016/j.knosys.2018.10.009
8. Fernandez-Basso, C.; Francisco-Agra, A.J.; Martin-Bautista, M.J.; Ruiz, M.D.: Finding tendencies in streaming data using big data frequent itemset mining. Knowl-Based Syst. **163**, 666–674 (2019). https://doi.org/10.1016/j.knosys.2018.09.026
9. Song, C.; Wang, X.K.; Cheng, P.F.; Wang, J.Q.; Li, L.: SACPC: a framework based on probabilistic linguistic terms for short text sentiment analysis. Knowl-Based Syst. **194**, 105572 (2020). https://doi.org/10.1016/j.knosys.2020.105572
10. Apala, K.R.; Jose, M.; Motnam, S.; Chan, C.C.; Liszka, K.J.; de Gregorio, F.: Prediction of movies box office performance using social media. In: 2013 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM 2013), IEEE, pp. 1209–1214 (2013). https://doi.org/10.1145/2492517.2500232
11. Golbeck, J.; Robles, C.; Turner, K.: Predicting personality with social media. In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262 (2011). https://doi.org/10.1145/1979742.1979614

12. Larkin, F.; Ryan, C.: Good news: using news feeds with genetic programming to predict stock prices. In: European Conference on Genetic Programming, Springer, Berlin, Heidelberg, pp. 49–60 (2008)
13. Mostafa, M.M.: More than words: social networks' text mining for consumer brand sentiments. Expert Syst. Appl. **40**(10), 4241–4251 (2013). https://doi.org/10.1016/j.eswa.2013.01.019
14. Kim, Y.; Jeong, S.R.; Ghani, I.: Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl. **6**(1), 2074–8523 (2014)
15. Thakkar, A.; Chaudhari, K.: A comprehensive survey on deep neural networks for stock market: the need, challenges, and future directions. Expert Syst. Appl. **177**(2), 114800 (2021)
16. Thakkar, A.; Chaudhari, K.: A comprehensive survey on portfolio optimization, stock price and trend prediction using particle swarm optimization. Arch. Comput. Methods Eng. (2020). https://doi.org/10.1007/s11831-020-09448-8
17. Thakkar, A.; Chaudhari, K.: Information fusion-based genetic algorithm with long short-term memory for stock price and trend prediction. Appl. Soft Comput. **128**, 109428 (2022). https://doi.org/10.1016/j.asoc.2022.109428
18. Thakkar, A.; Patel, D.; Shah, P.: Pearson Correlation Coefficient-based performance enhancement of Vanilla Neural Network for stock trend prediction. Neural Comput. Appl. **33**(24), 16985–17000 (2021)
19. Chaudhari, K.; Thakkar, A.: iCREST: international cross-reference to exchange-based stock trend prediction using long short-term memory. In: Applied Soft Computing and Communication Networks. Springer, Singapore, pp. 323-338 (2021)
20. Pavai, G.; Geetha, T.V.: Improving the freshness of the search engines by a probabilistic approach based incremental crawler. Inform. Syst. Front. **19**(5), 1013–1028 (2017)
21. Hernández, I.; Rivero, C.R.; Ruiz, D.: Deep Web crawling: a survey. WWW **22**(4), 1577–1610 (2019)
22. Ro, I.; Han, J.S.; Im, E.G.: Detection method for distributed web-crawlers: a long-tail threshold model. Secur. Commun. Netw. (2018). https://doi.org/10.1155/2018/9065424
23. Weng, Y.; Wang, X.; Hua, J.; Wang, H.; Kang, M.; Wang, F.Y.: Forecasting horticultural products price using ARIMA model and neural network based on a large-scale data set collected by Web crawler. IEEE Trans. Comput. Soc. Syst. **6**(3), 547–553 (2019). https://doi.org/10.1109/TCSS.2019.2914499
24. Arillotta, D.; Schifano, F.; Napoletano, F.; Zangani, C.; Gilgar, L.; Guirguis, A.; Corkery, J.M.; Aguglia, E.; Vento, A.: Novel opioids: systematic web crawling within the e-psychonauts' scenario. Front. Neurosci. **14**, 149 (2020). https://doi.org/10.3389/fnins.2020.00149
25. Singh, R.; Srivastava, S.: Stock prediction using deep learning. Multimed. Tools Appl. **76**(18), 18569–18584 (2017)
26. Vargas, M.R.; De Lima, B.S.; Evsukoff, A.G.: Deep learning for stock market prediction from financial news articles. In: 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), IEEE, pp. 60–65 (2017)
27. Kim, Y.: Convolutional neural networks for sentence classification. https://arxiv.org/abs/1408.5882
28. Sun, X.; Gao, F.; Li, C.; Ren, F.: Chinese microblog sentiment classification based on convolution neural network with content extension method. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, pp. 408–414 (2015)
29. Liao, S.; Wang, J.; Yu, R.; Sato, K.; Cheng, Z.: CNN for situations understanding based on sentiment analysis of twitter data. Procedia Comput. Sci. **111**, 376–381 (2017). https://doi.org/10.1016/j.procs.2017.06.037
30. Dos Santos, C.; Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 69–78 (2014)
31. Zhang, Y.; Jiang, Y.; Tong, Y.: Study of sentiment classification for Chinese microblog based on recurrent neural network. Chin. J. Electron. **25**(4), 601–607 (2016). https://doi.org/10.1049/cje.2016.07.002
32. Abdi, A.; Shamsuddin, S.M.; Hasan, S.; Piran, J.: Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Inform. Process. Manag. **56**(4), 1245–1259 (2019). https://doi.org/10.1016/j.ipm.2019.02.018
33. Yan, Y.; Yang, D.: A stock trend forecast algorithm based on deep neural networks. Sci. Program. **2021**(2), 1–7 (2021). https://doi.org/10.1155/2021/7510641
34. Schumaker, R.P.; Chen, H.: Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Trans. Inf. Syst. (TOIS) **27**(2), 1–19 (2009). https://doi.org/10.1145/1462198.1462204
35. Mungra, D.; Agrawal, A.; Thakkar, A.: A voting-based sentiment classification model. In: Choudhury, S.; Mishra, R.; Mishra, R.; Kumar, A. (Eds.) Intelligent Communication, Control and Devices. Advances in Intelligent Systems and Computing, p. 989. Springer, Singapore (2020)
36. Thakkar, A.; Mungra, D.; Agrawal, A.: Sentiment analysis: an empirical comparison between various training algorithms for artificial neural network. Int. J. Innov. Comput. Appl. **11**(1), 9 (2020). https://doi.org/10.1504/IJICA.2020.105315
37. Thakkar, A.; Mungra, D.; Agrawal, A.; Chaudhari, K.: Improving the performance of sentiment analysis using enhanced preprocessing technique and Artificial Neural Network. IEEE Trans. Affect. Comput. (2022). https://doi.org/10.1109/TAFFC.2022.3206891
38. Bollen, J.; Mao, H.; Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. Neth. **2**(1), 1–8 (2011). https://doi.org/10.1016/j.jocs.2010.12.007
39. Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K.: Predicting stock market index using fusion of machine learning techniques. Expert Syst. Appl. **42**(4), 2162–2172 (2015). https://doi.org/10.1016/j.eswa.2014.10.031
40. Ding, X.; Zhang, Y.; Liu, T.; Duan, J.: Deep learning for event-driven stock prediction. In: 24th International Joint Conference on Artificial Intelligence, (2015)
41. Li, G.; Zhang, A.; Zhang, Q.; Wu, D.; Zhan, C.: Pearson correlation coefficient-based performance enhancement of broad learning system for stock price prediction. IEEE Trans. Circuits Syst. II Express Br. **69**(5), 2413–2417 (2022). https://doi.org/10.1109/TCSII.2022.3160266
42. Huang, S.C.; Chuang, P.J.; Wu, C.F.; Lai, H.J.: Chaos-based support vector regressions for exchange rate forecasting. Expert Syst. Appl. **37**(12), 8590–8598 (2010). https://doi.org/10.1016/j.eswa.2010.06.001
43. Yu, Y.; Duan, W.; Cao, Q.: The impact of social and conventional media on firm equity value: a sentiment analysis approach. Dec. Support Syst. **55**(4), 919–926 (2013). https://doi.org/10.1016/j.dss.2012.12.028
44. Hagenau, M.; Liebmann, M.; Neumann, D.: Automated news reading: Stock price prediction based on financial news using context-capturing features. Dec. Support Syst. **55**(3), 685–697 (2013). https://doi.org/10.1016/j.dss.2013.02.006
45. Chatrath, A.; Miao, H.; Ramchander, S.; Villupuram, S.: Currency jumps, cojumps and the role of macro news. J. Int. Money Finance **40**, 42–62 (2014). https://doi.org/10.1016/j.jimonfin.2013.08.018
46. Kim, S.H.; Kim, D.: Investor sentiment from internet message postings and the predictability of stock returns. J. Econ Behav. Organ. **107**, 708–729 (2014). https://doi.org/10.1016/j.jebo.2014.04.015

47. Das, S.R.; Chen, M.Y.: Yahoo! for Amazon: sentiment extraction from small talk on the web. Manage. Sci. **53**(9), 1375–1388 (2007). https://doi.org/10.1287/mnsc.1070.0704
48. Tumarkin, R.; Whitelaw, R.F.: News or noise? Internet postings and stock prices. Finance Anal. J. **57**(3), 41–51 (2001). https://doi.org/10.1023/A:1018810005576
49. Zhang, X.; Zhao, J.; LeCun, Y.: Character-level convolutional networks for text classification. Adv. Neural Inf. Process. Syst. (NIPS) **28**, 649–657 (2015). https://doi.org/10.48550/arXiv.1509.01626
50. Boureau, Y.L.; Bach, F.; LeCun, Y.; Ponce, J.: Learning mid-level features for recognition. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 2559–2566 (2010)
51. Kingma, D.; Ba, J.: Adam: a method for stochastic optimization. https://arxiv.org/abs/1412.6980
52. Hsu, D.: Time series forecasting based on augmented long short-term memory. https://arxiv.org/abs/1707.00666
53. Gamboa, J.C.B.: Deep learning for time-series analysis. https://arxiv.org/abs/1701.01887
54. Panpoonsup, T.; Silpasuwanchai, C.; Pananookooln, C.; Dailey, M.: Evaluating the effectiveness of sentiment-based models for stock price prediction. Available at SSRN 4185507 (2022). https://doi.org/10.2139/ssrn.4185507
55. Li, X.; Wu, P.; Wang, W.: Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong. Inf. Process. Manage. **57**(5), 102212 (2020). https://doi.org/10.1016/j.ipm.2020.102212
56. Kim, J.M.; Lee, J.H.: Text document classification based on recurrent neural network using word2vec. J. Korean Inst. Intell. Syst. **27**(6), 560–565 (2017)
57. Chen, Y.; Yuan, J.; You, Q.; Luo, J.: Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 117–125 (2018). https://doi.org/10.1145/3240508.3240533
58. Sun, M.: Chinese text classification based on GRU-attention. Mod. Inform. Technol. **3**(03), 10–12 (2019)
59. Thakkar, A.; Chaudhari, K.: CREST: cross-reference to exchange-based stock trend prediction using long short-term memory. Procedia Comput. Sci. **167**, 616–625 (2020). https://doi.org/10.1016/j.procs.2020.03.328