The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2398-7294.htm

A stock price prediction method based on deep learning technology

Xuan Ji, Jiachen Wang and Zhijun Yan
School of Management and Economics, Beijing Institute of Technology, Beijing, China

Received 31 May 2020
Revised 26 August 2020
Accepted 7 September 2020

Abstract

Purpose – Stock price prediction is a hot topic and traditional prediction methods are usually based on statistical and econometric models. However, these models are difficult to deal with nonstationary time series data. With the rapid development of the internet and the increasing popularity of social media, online news and comments often reflect investors' emotions and attitudes toward stocks, which contains a lot of important information for predicting stock price. This paper aims to develop a stock price prediction method by taking full advantage of social media data.

Design/methodology/approach – This study proposes a new prediction method based on deep learning technology, which integrates traditional stock financial index variables and social media text features as inputs of the prediction model. This study uses Doc2Vec to build long text feature vectors from social media and then reduce the dimensions of the text feature vectors by stacked auto-encoder to balance the dimensions between text feature variables and stock financial index variables. Meanwhile, based on wavelet transform, the time series data of stock price is decomposed to eliminate the random noise caused by stock market fluctuation. Finally, this study uses long short-term memory model to predict the stock price.

Findings – The experiment results show that the method performs better than all three benchmark models in all kinds of evaluation indicators and can effectively predict stock price.

Originality/value – In this paper, this study proposes a new stock price prediction model that incorporates traditional financial features and social media text features which are derived from social media based on deep learning technology.

Keywords Text mining, Deep learning, Financial social media, Stock price prediction

Paper type Research paper

1. Introduction

Stock is a financial product characterized by high risk, high return and flexible trading, which is favored by many investors. Investors can get abundant returns by accurately estimating stock price trends. However, the stock price is influenced by many factors such as macroeconomic situation, market condition, major social and economic events, investors' preferences and companies' managerial decisions. Therefore, prediction of the stock price has always been the focus and difficult research topic. Statistical and econometric models

---

Funding: This work was supported by National Key Research and Development Plan of China (Grant No: 2017YFB1400101), National Natural Science Foundation of China (Grant No: 71572013, 71872013, 72072011) and Beijing Municipal Social Science Foundation (Grant No: 18JDGLB040).

International Journal of Crowd Science
Vol. 5 No. 1, 2021
pp. 55-72
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-05-2020-0012

are generally used in traditional stock price prediction, but these methods cannot deal with the dynamic and complex environment of the stock market. Since 1970, with the rapid development of computer technology, researchers have begun using machine learning to predict stock prices and fluctuations, helping investors determine investment strategies to reduce risk and increase returns.

The stock market is a highly complex time series scenario and has typical dynamic characteristics. There will be a lot of stock dynamic trading after the opening of the market and stock price will change accordingly. Moreover, the stock price is affected by many unpredicted factors, which results in a typical nonstationary stock price time-series data. Therefore, stock price prediction is one of the most challenging problems in all kinds of prediction research. In the past decades, scholars have studied stock price prediction from many perspectives, where the improvement of prediction models and the selection of model features are the two most important directions among them. Most of the early studies used econometric models, such as autoregressive integrated moving average (ARIMA) and autoregressive conditional heteroskedastic-autoregressive integrated moving average (ARCH-ARIMA) (Booth *et al.*, 1994; Engle, 2001), to predict stock price. However, it is difficult for econometric models to consider the impact of other factors on stock price fluctuations and they have strong assumptions about the data, which are often difficult to meet (Le and Xie, 2018). Therefore, machine learning has been widely used in stock price prediction in recent years and many more suitable models for stock prediction have been proposed. Many studies have shown that deep learning has superior efficiency than other models (Marmer, 2008) and neural network models excel regression and discriminant models (Refenes *et al.*, 1994). In terms of feature selection, some scholars explore the correlation between new features and stock price and some new features, including political factors, macroeconomic factors and investors' sentiment, etc., have been incorporated into the prediction model (Cervello-Royo *et al.*, 2015; Patel *et al.*, 2015).

Previous literature extensively investigates the stock price prediction methods and many advanced prediction models are proposed. However, existing approaches on stock price prediction have two main limitations. First, although the text features are used in the existing models to better incorporate the important information in social media, they are usually mined based on traditional text mining technologies, such as the bag-of-word model. These text mining technologies cannot consider the semantic and other information in social media which are helpful to improve the performance of prediction models. Second, the feature dimensionality reduction is a basic step when balancing text features and financial features in stock price prediction. However, previous price prediction methods usually adopt principal component analysis (PCA) and latent Dirichlet allocation (LDA) to reduce the feature dimension. PCA method has problems of information loss and is unable to process nonlinear data, while the LDA method cannot consider semantic information in social media. Thus, these two methods are not suitable for the stock price prediction (Bao *et al.*, 2017).

To fill the research gap discussed above, this paper proposes a new stock price prediction method based on deep learning technology, which integrates Doc2Vec, stacked auto-encoder (SAE), wavelet transform and long short-term memory (LSTM) model. Feature extraction of text information in social media can describe the emotional tendency of investors and help to predict the stock price more accurately. First, we classify the prediction features into two types, i.e. financial features and text features. We adopt the widely used financial features and extract text features from social media by deep learning technology. Second, Doc2Vec model is used to train original social media documents and obtain text feature vectors. Doc2Vec model can retain semantic information of documents and the relationship between different words, which overcomes the shortcomings of traditional text feature extraction methods (such as dictionary matching method, term frequency–inverse document frequency and LDA). Third, SAE is adopted to reduce dimension of text feature vectors, which balances the dimension of text features with financial features. Fourth, wavelet transform is used to transform the target variable (stock price) and to remove the random noise in the stock price time series data. Finally, stock finance features and excavated text features are taken as input features, and LSTM is adopted to predict the stock price.

The rest of this paper is structured as follows. We review the literature on stock price prediction in Section 2 and introduce our method in Section 3. We explain the research data and experimental process in Section 4. Finally, we conclude the paper with a summary and possible future research directions in Section 5.

2. Related literature

Our paper studies the stock price prediction method based on deep learning. The related research work is mainly about the prediction model and feature selection of the prediction model. This section will review the literature from these two aspects.

2.1 Prediction model

The improvement of prediction models has always been one of the most important research directions of stock price prediction. Stock price prediction methods mostly adopt an econometric model or machine learning model. These two models have been continuously improved to be more suitable for processing financial time series data in the complex stock market.

In terms of econometric models, Booth *et al.* used the ARIMA model to predict stock price by six explanatory variables, including macroeconomic factors and lag factors. The results showed that these variables were helpful to improve the accuracy (Booth *et al.*, 1994). Breidt *et al.* believed that ARCH, generalized ARCH (GARCH) or standard (short-memory) stochastic volatility models were not appropriate to predict stock price (Breidt *et al.*, 1998). They proposed a new time series prediction method that can deal with conditional variances, called the long memory stochastic volatility model, which was superior to other models. (Zhang *et al.*, 2018) proposed an enhanced ARIMA-GARCH model based on differential information (Zhang and Zhang, 2016). By adding the approximate differential information of the dependent variable lag and taking stock price change trend information into account, the ability to predict the direction of price change is improved.

With the development of machine learning technology, many scholars try to solve the problem with new emerging technology instead of traditional prediction models to predict stock price more accurately. Maknickash and Maknickiene used a recursive neural network (RNN) to construct a stock price prediction model and optimized the selection of RNN parameters, such as the number of neurons and the number of iterations (Maknickas and Maknickiene, 2019). Nelson *et al.* (2017) used the LSTM model and multiple stock analysis indicators to predict the rise and fall of the stock price in the future (Nelson *et al.*, 2017). The result showed that the performance of LSTM was better than that of the traditional machine learning model and non-time series model. Peng *et al.*, 2019 mainly focused on the preprocessing method of financial time series data, including interpolation, wavelet de-noising and normalization on the data and tried various parameter combinations of the LSTM model (Peng *et al.*, 2019). They found that the optimized model had a low computational complexity and significantly improved prediction accuracy. Vo *et al.* compared the effects of LSTM, Bi-directional LSTM (Bi-LSTM) and gated recurrent unit on the stock price prediction. They found that the Bi-LSTM model read the data one more time backward which helped improve prediction accuracy, particularly in forecasting sequential data such as financial time series (Vo *et al.*, 2019).

In addition, existing literature also combines statistical econometric models with machine learning models or use more than two kinds of machine learning models at the same time to predict stock price. Compared with a single model, these models usually have better performance. Achkar *et al.* (2018) considered two different model combination methods, back propagation algorithm-multi-layer perception (BPA-MLP) and LSTM-RNN. Using the stock price data of Facebook, Google and the price data of Bitcoin, they found that the LSTM-RNN model was better than BPA-MLP (Achkar *et al.*, 2018). Bao *et al.* first reduced noise in original time series stock data through wavelet transform and then predict by LSTM model (Bao *et al.*, 2017). The results showed that the performance of the integrated model was better than other similar models. M'ng and Mehralizadeh proposed a prediction model named wavelet principal component analysis-neural network (WPCA-NN), which combined wavelet transform, PCA and artificial neural network to de-noising, removing the random noise in the stock price sequence (M'ng and Mehralizadeh, 2016). The results showed that the performance of the WPCA-NN was better than traditional prediction methods. KIM T. and KIM H. proposed an LSTM-CNN model based on feature combination, using stock time series and stock trend graphs as input features (Kim and Kim, 2019). The results showed that the LSTM-CNN model was superior to the single model in predicting stock prices.

In summary, econometric models and machine learning models are the two most widely used methods in stock price prediction. However, it is difficult to deal with nonlinear time series problems by econometric models, while traditional machine learning models mostly take single period data as a sample and ignore a lot of implicit information developing over time (Baek and Kim, 2018). Deep learning technology is a new emerging technology that can effectively process time-series data and multi-period data. At the same time, the combination of multiple models usually has better performance than a single model and is becoming the main direction in stock price prediction.

2.2 Feature selection of prediction model

Feature selection and feature engineering are helpful to enrich the data set and to mine important information from the original data set, which can improve prediction accuracy. This is also a hot research direction that scholars care about. Previous studies generate different useful features from the original data set to improve the stock price prediction performance, such as emotion of stock investors (Delong *et al.*, 1990; Shleifer and Vishny, 1997), stock movement graphs (Quan, 2013; Singh and Srivastava, 2017) and major economic and political events (Ding *et al.*, 2015; Zubiaga, 2018). Among them, the emotion of stock investors is usually one of the most adopted features in stock price-prediction models.

The investors' emotion for a stock and the overall stock market often has an important impact on stock price fluctuation (Nassirtoussi *et al.*, 2014). Therefore, previous studies normally use natural language processing (NLP) technology to analyze stock social media documents and obtain investors' emotions, which provides new important features for stock price prediction models. Schumaker and Chen proposed a new method named proper noun scheme, which marked important nouns (Schumaker and Chen, 2009). They classified nouns into seven types, including date, location, money, organization, percentage, person and time. The results showed that the proper noun scheme performed better than word-of-bag and named entity recognition. Kraus and Feuerriegel collected the financial text disclosed by companies on the trading day and used sequence modeling to deal with company disclosures (Kraus and Feuerriegel, 2017). Then, they combined RNN and LSTM models to predict stock price. The results showed that the introduction of financial text could effectively improve prediction accuracy. Zhou *et al.* used the word-of-bag model to extract five emotional attributes of the stock market investors, such as disgust, joy, sadness and fear (Zhou *et al.*, 2018). They found the results of the K-means model were significantly better than the baseline models, including the one taking purely financial time series as input features. LDA is another less frequently used but interesting technique in stock price prediction. Jin *et al.* used LDA to extract topics from the text and the representative topics were adopted as input features of the prediction model, which helped improve the prediction accuracy (Jin *et al.*, 2013).

Feature dimensionality reduction is also one of the most important research directions of feature extraction. Xie *et al.* proposed a dual dimension reduction model of joint mutual information to improve the PCA algorithm. The model first used mutual information to preliminarily screen a large number of features and then determined the number of PCA principal elements for secondary dimension reduction based on complex correlation coefficient and cumulative variance contribution rate (Xie *et al.*, 2020). The prediction results showed that the improved method was better than the traditional feature reduction model. Hagenau *et al.* attempted to use more expressive features to represent text. They used dictionaries to extract English word stems and then used Chi-square and Bi-normal-separation to calculate the interpretive power of features, retaining only those features with strong interpretive power (Hagenau *et al.*, 2013). The results showed that the feature dimension reduction method could significantly improve the prediction accuracy and reduce the overfitting problem in machine learning models. Huang *et al.* extracted nouns, verbs and compound phrases from the text and then based on the thesaurus, made the conversion of synonyms for each word to achieve the purpose of dimension reduction (Huang *et al.*, 2010).

In summary, existing research mostly adopts traditional text feature extraction methods (such as word-of-bag, named entity recognition and LDA) to add new text features for stock price prediction. Although the features extracted by these methods can represent the emotions of investors to some extent, these extracted features normally cannot represent document semantic information, context and other information in social media (Nassirtoussi *et al.*, 2014). Deep learning technology can retain the semantic information and better extract effective information from original documents. Therefore, this paper adopts deep learning technology to extract text features and achieve more accurate price prediction.

3. A new stock price prediction method

We propose a new stock price prediction model (Doc-W-LSTM) based on deep learning technology, which integrates Doc2Vec, SAE, wavelet transform and LSTM model. It uses stock financial features and text features to predict future stock prices. The model mainly includes several steps:

- Selecting features for the prediction model;
- Training the text feature vector by Doc2Vec and reducing the dimension of the text feature vector by SAE;
- Denoising the stock price time series data based on wavelet transform method;
- Predicting the stock price with LSTM model (Figure 1).

3.1 Feature selection

Our model has two types of input features including financial features and social media text features. Financial features are descriptions of the daily trading data of stocks, which can

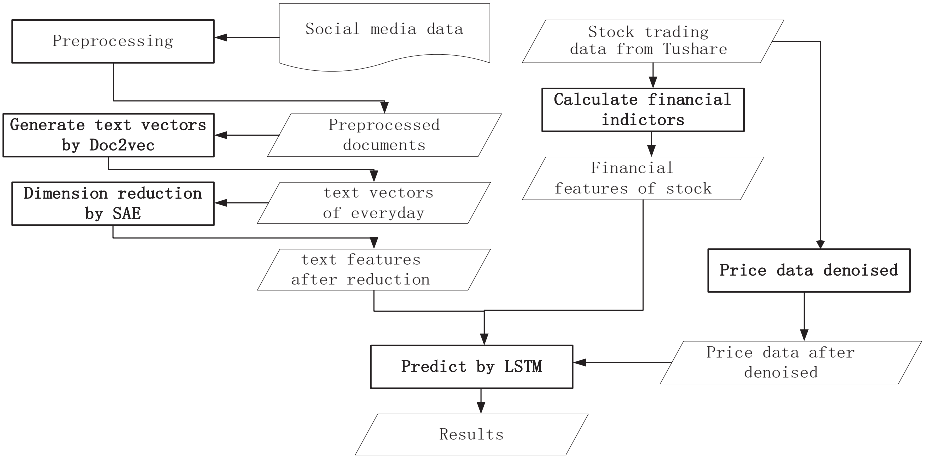
Preprocessing
Social media data
Stock trading data from Tushare
Generate text vectors by Doc2vec
Preprocessed documents
Calculate financial indictors
Dimension reduction by SAE
text vectors of everyday
Financial features of stock
text features after reduction
Price data denoised
Predict by LSTM
Price data after denoised
Results

Figure 1.
Stock price prediction method

reflect basic information of stock prices. Text features of social media refer to the text feature vector containing effective information in social media, which includes investors' comments in financial social media and the news published by companies.

Financial features are described by daily transaction data and financial technical indicators (Bao *et al.*, 2017) and include 21 features. Daily transaction data include open/close price, low/high price, trading volume, change amount and change rate, which are daily first-hand trading information in the stock market. These characteristics can directly reflect the historical trading situation of stocks. Financial technical indicators refer to indicators calculated based on stock trading data, including CCI (commodity channel index), ATR (average true range), SMI (stochastic momentum index), etc. These indicators can often reflect some regular characteristics of stock movements. Table 1 gives the list of financial features.

Text features are extracted from social media and can represent investors' attitudes toward the stock, the overall trend of public opinions and the company's decision information (Nassirtoussi *et al.*, 2014; Kraus and Feuerriegel, 2017) and some samples of text are shown in excerpt cases of text from social media (Translated).

> Hurry up! Buy it, just think it as giving yourself in advance retirement salary. Meinian Health and Zhonggong Education are the two long-term stocks I followed this year. They didn't disappoint me. Although I didn't buy much, I was very happy.

> Meinian health medical treatment develops rapidly, and sets up health physical examination center all over the country. Physical examination crowd are becoming more and more, and physical checks must be prepared in advance, so the potential of Meinian health is huge. Its performance will get better and better, like that of Maotai, which may be the real reason why Alibaba has become the second largest shareholder.

The proposed Doc-W-LSTM method uses the Doc2Vec model to extract high-dimensional text feature vectors. However, the trained text feature vectors will have a large dimension, while the financial feature only has 21 dimensions. The dimension imbalance between these two kinds of features will affect the prediction effect of the proposed method, thus we will reduce the dimension of text feature vectors to make two types of features have the same dimension.

Feature	Description
Daily transaction data	
Open/close price	Nominal daily open/close price
High/low price	Nominal daily highest/lowest price
Volume	Daily trading volume
Price_change	Change volume
P_change	Change rate
Technical indicators	
MACD	Moving average convergence divergence: displays trend following characteristics and momentum characteristics
CCI	Commodity channel index: helps to find the start and the end of a trend
ATR	Average true range: measures the volatility of price
BOLL	Bollinger band: provides a relative definition of high and low, which aids in rigorous pattern recognition
EMA20	20 days exponential moving average
MA5/MA10	5/10 days moving average
V_MA5/V_MA10	5/10 days trading volume average
MTM6/MTM12	6/12 months momentum: helps pinpoint the end of a decline or advance
ROC	Price rate of change: shows the speed at which a stock's price is changing
SMI	Stochastic momentum index: shows where the close price is relative to the midpoint of the same range
WVAD	Williams's variable accumulation/distribution: measures the buying and selling pressure

Table 1.
Stock financial features

3.2 Word segmentation and text feature extraction

The proposed method first extracts the text feature vector from social media and then uses the text feature vector to describe the influence of the emotional tendency of stock investors and company decisions on stock price volatility. We collect two kinds of texts from financial social media, namely, daily comments of investors and company information published by social media platforms and then preprocess and extract the text features from social media. The daily comments of investors are posted on social media and represent their feelings about the company's stock and expectations of the stock market. Then the social medial platform collects and publishes companies' financial information online.

First, the social media documents are preprocessed to remove useless characters, such as special symbols and meaningless words. There is no space between two adjacent words in a Chinese sentence, so word segmentation is a necessary step. The Jieba package in Python is used to separate each word in a sentence.

Second, we choose Doc2Vec to extract features from the text. Doc2Vec is a method proposed in 2014 to map various length text into the fixed-length vector, which can mine the semantic and emotional information hidden in the text (Le and Mikolov, 2014). Doc2Vec is based on a neural network language model to learn the vector expression of each text, mainly including two models, i.e. DM (distributed memory) and DBOW (distributed bag of words). Both models use the context and paragraph features to estimate the probability distribution of the occurrence of a certain word and then generate the text feature vector of the document. Probability can be used to express the similarity of contents and emotions in different paragraphs.

DM model predicts the vector of the next word by combining paragraph vector and word vector, that is, it predicts the probability distribution of target word under the condition of given paragraph vector and context. A fixed-length sliding window is set during the training. The probability distribution of the next word in the training window is used and the window is moved back one word after the training. The DBOW model ignores the context relationship of the original input documents. Namely, it does not slide the window from the beginning of the document but predicts a random word in the paragraph. In other words, at each iteration, a training window will be sampled from the text and then a word will be randomly sampled from the window as the prediction task. Kim *et al.* found that the combination of DM and DBOW would achieve better performance (Kim *et al.*, 2019). Thus, we will also select the best-performing set of models in both the DM and DBOW tests and splice the text vectors they generate as input features for our next step.

Doc2Vec model can be used for a lot of unstructured text data mining and NLP research, such as false comment identification, document classification, document emotion analysis and other tasks. In this paper, the Doc2Vec model is used to extract the features of social media text and form text feature vectors. As non-text features are collected at daily intervals, but multiple posts are posted on social media every day, we divide multiple posts on the same date into a group and take the values of each dimension of the text feature vectors for the enantiomorphic posts within the group as the text feature vectors of the day (Bollen *et al.*, 2011). Based on the Doc2Vec model, the high-dimensional text features on a daily basis can be obtained and then the dimension of the text features will be reduced to reach the same dimension as the financial features.

3.3 Text feature dimension reduction

Doc2Vec model normally generates a text feature vector with a large dimension, which can better represent the semantic information in the original document. In general, the text feature vector usually has hundreds or even thousands of dimensions. However, the proposed method only has 21 dimensions of financial features. The imbalanced dimension of two types of features leads to two serious problems. On the one hand, this imbalance will weaken the importance of the financial feature in the prediction model. However, the financial features in stock price prediction are very important, while other new features usually provide additional useful information. On the other hand, the large dimension of the text feature vector will affect the training speed of the proposed method. Therefore, we introduce SAE to compress the dimension of text feature vectors and make sure that two kinds of features have the same dimension and valuable information in the original text feature vector is still retained as much as possible.

Auto-encoder is a kind of neural network that replicates the input signal in the output layer as much as possible (Zhang *et al.*, 2018). Its input vector and output vector have the same dimension. Auto-encoder can effectively encode the input data to generate the hidden layer and then generate the output data through decoding. The encoded hidden layer vectors can well represent the important information of the original vectors in low dimensional space.

An auto-encoder consists of three layers of a neural network. The first layer and the third layers are the input layer and the output layer, respectively. Then the second layer is the hidden layer, which generates high-level features of the data. The purpose of auto-encoder training is to make the input and output vectors as similar as possible. The first training step is to map the input vector to the hidden layer and the second training step is to reconstruct the vector by mapping the hidden layer vector to the output layer. These two steps can be expressed as:

$$h(x) = f(W_1 x + b_1) \tag{1}$$

$$y = g\left(W_2 h(x) + b_2\right) \tag{2}$$

where x is the vector of high-dimensional text trained by Doc2Vec. $h(x)$ is the hidden layer vector generated by the auto-encoder. W_1 and W_2 are the coefficient vectors of the hidden layer and output layer and b_1 and b_2 are the constant terms of the hidden layer and output layer. f and g correspond to encoding function and decoding function, respectively. They are activation functions, such as Sigmoid, ReLU and hyperbolic tangent function. We use the Sigmoid function as the activation function (Chen *et al.*, 2014). y is the output vector through coding and decoding, which has the same dimension as x. The goal of auto-encoder training is to make y and x as similar as possible.

The dimension of the text vector generated by Doc2Vec is very large. We use SAE to reduce it and to get a 21-dimension text vector which has the same dimension with financial features. SAE is composed of multiple auto-encoders stacked layer by layer (Schölkopf, 2007). Each layer is based on the expression of the last layer, which can learn deep expressions of original data and be more suitable for complex reduction tasks. We build a four-layer SAE that combines with three auto-encoders. The concrete structure is $\{\mathrm{x}, \mathrm{m}_1, \mathrm{m}_2, \mathrm{k}\}$, where m_1 is the hidden layer vector of the first auto-encoder. m_2 is the hidden layer vector of the second auto-encoder. k is the hidden layer vector generated by the third auto-encoder. Then k is the text feature vector finally obtained by dimension reduction. The information loss in the original high-dimension vector can be minimized by SAE, where the dimension size relation is $\mathrm{k} < \mathrm{m}_2 < \mathrm{m}_1 < \mathrm{x}$. The specific structure of the SAE we used is shown in Figure 2. The circles in Figure 2 represent neurons and dotted lines represent connections between different neurons.

3.4 Noise reduction for time series data

Due to the complexity of stock market fluctuation, the stock price is often full of random noise, which will lead to large price volatility and then result in overfitting problems. We hope to eliminate some noise with strong randomness while preserving the data trend. In general, noise reduction of time series data is to eliminate many small fluctuations in the original data through function transformation. It helps smooth the curve of the original data without changing the overall fluctuation trend.

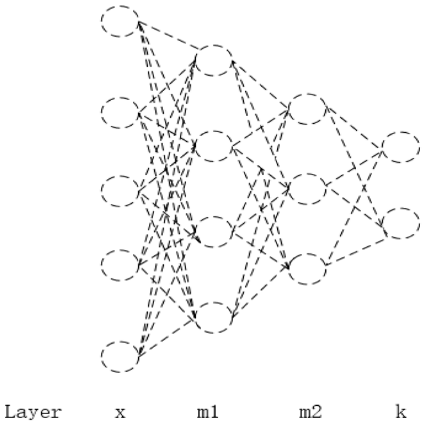

Figure 2.
Structure of SAE

As a commonly used noise reduction method, wavelet transform can better deal with non-stationary time series data and preserve the characteristics of original data as much as possible. It is widely used in prediction tasks in financial scenarios (Papagiannaki *et al.*, 2005; Ramsey, 1999). Therefore, we choose Haar wavelet transform as the noise reduction method for the stock price. This method can decompose data according to time and frequency and has an acceptable processing time, with the time complexity of O(n) (Abramovich *et al.*, 2002).

The basic principle of wavelet transform is to generate some wavelet signals which contain important information and noise after transforming the original data. The signal coefficient of important information is larger and the signal coefficient of noise is smaller. The algorithm will automatically select a suitable threshold. The wavelet signals greater than the threshold is considered to contain important information and should be retained, while the signals less than the threshold are considered as noise and will be removed.

3.5 Prediction model

LSTM neural network is an improved model of RNN. The input data of LSTM and RNN will time dimension, which can improve the performance of time series prediction. Compared with RNN, LSTM adds three different gates, i.e. forget gate, input gate and output gate, to solve the gradient disappearance problem, which has been widely applied in time series modeling. Therefore, we choose it as the final prediction model.

LSTM is composed of multiple neurons. In each neuron, data first enters forget gate. The forget gate determines which input information will be forgotten so it will not affect the update of the next neuron. In the second step, the input gate decides which information is allowed to be added. The output of the previous neuron and input of the local neuron are processed by the sigmoid function and tanh function to generate two results. And then which information needs to be updated is decided based on these two results. The results will be saved for the output gate. Finally, the output gate determines which result obtained in the input gate can be generated. The results from the output gate of one neuron will be inputted to the next neuron, etc.

The input data of LSTM is a three-dimension array, representing time dimension, sample dimension and feature dimension, respectively. The time dimension represents the sliding time window, the sample dimension represents the sample size of training and testing and the feature dimension represents the number of input features. We choose 7 days as a time window to predict the close price on day 8. The input features are 42 dimensions, half of which are financial features and the other half are text features.

4. Experiment and results

We collected investors' comments and news of 15 companies from one famous social media platform ("Oriental Fortune website") and obtained stock transaction data from the Tushare financial database. Then mean absolute error (MAE), root mean square error (RMSE) and R-squared (R^2) were selected to evaluate the performance of the proposed prediction method.

4.1 Data Acquisition

We crawled investors' comments and companies' news of the top 15 listed medical companies from the "Oriental Fortune website" generated between January 2010 and November 2019. A total of 530,813 documents were obtained, which include comments and news. After collecting the social media text data, we first conducted preprocessing, which consisted of three steps. First, documents with less than 20 Chinese characters in length were deleted. Second, identical documents were removed. Third, some continuous expressions were compressed. For example, "good good good" in investors' comments were changed to "good." After text preprocessing, 342,118 documents were left.

We chose to predict the stock price of the company "Meinian Health." That is because "Meinian Health" ranks the second among all listed medical companies in total profit and its investors' comments and official news are very active. At the same time, "Meinian Health" went to the public in 2005, so we can collect all documents after 2010. We used documents of 14 companies except "Meinian Health" for text vector training and then generated the text vector of "Meinian Health." Table 2 is the descriptive information of social media documents after data preprocessing.

We also collected daily transaction data of "Meinian Health" from the Tushare financial database, including open, close, high, low, trading volume, change volume and change rate. In addition, we calculated financial technical indicators based on daily transaction data (Bao *et al.*, 2017).

4.2 Metrics

We used MAE, RMSE and R^2 as measures to evaluate the performance of the prediction methods. MAE measures error without considering the directions of the predicted values. RMSE measures the average magnitude of estimation error in predicted values. MAE and RMSE are measures of closeness which evaluates the accuracy of the predicted value to the actual price. R^2 measures the linear correlation between two variables and eliminates the influence of dimension on different regression problems. We hope the model has low MAE and RMSE and a high R^2. The three metrics are defined as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_{true}^{(i)} - y_{predict}^{(i)}\right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_{true}^{(i)} - y_{predict}^{(i)}\right)^2} \tag{4}$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_i \left(y_{true}^{(i)} - y_{predict}^{(i)}\right)^2}{\sum_i \left(\bar{y} - y_{true}^{(i)}\right)^2} \tag{5}$$

where $y_{true}^{(i)}$ represents the true value of the target variable of sample i, $y_{predict}^{(i)}$ represents the predicted value of the target variable of sample i, $\bar{y}$ represents the mean of the target variable's true value of all samples and m represents the total number of samples.

Table 2. Text feature description

Company name	The no. of documents	Average length
"Meinian health"	38,486	289
Other 14 companies	303,632	305

4.3 Experiment

Doc2Vec is an unsupervised algorithm, but it includes hyper-parameters such as model type (DM or DBOW), vector dimension and the size of the sliding window. The model type of Doc2Vec, dimension and the window size are denoted as m, s and w, respectively.

According to the research of Kim *et al.* (2019), the text vector dimension of Doc2Vec usually chooses a value slightly lower than the average length of the document, so we chose 100 and 200 as the candidate value of the parameter s. Too large or too small sliding windows can affect the performance of the model, so we chose 5 and 10 as the candidate values of the parameter w (Kim *et al.*, 2017). DM has no sliding window parameters, so three hyper-parameters are combined to produce six different scenarios for Doc2Vec parameter optimization.

Lau *et al.* proposed that the tuning of Doc2Vec should be carried out in combination with specific models and tasks (Lau and Baldwin, 2016). We chose the LSTM model as the prediction model to tune the hyper-parameters of the Doc2Vec model. For different hyper-parameter scenarios, we used the LSTM model to predict the stock price and to identify the optimal parameter combination of the Doc2Vec model.

We used the other 14 companies' text as training data of Doc2Vec and used the trained Doc2Vec model to generate the vector of "Meinian Health" as the predicted object. We chronologically divided the text data of "Meinian Health" into three parts, i.e. the first 80% data is train set, the middle 10% data is validation set and the last 10% data is test set. We input different vectors generated by Doc2Vec into LSTM without financial features, used the train set and the validation set to train the model and to adjust parameters and used the test set to verify the model performance. The results of different parameter scenarios are shown in Table 3.

It can be seen from the Table 3 that the DM model has the best performance when s is equal to 200, while the DBOW model has the best performance when s is equal to 100 and w is 5. Kim *et al.* found that the combination of DM and DBOW would achieve better performance (Kim *et al.*, 2019). Therefore, we examined the prediction performance of the DM-DBOW model with the optimal parameters identified in previous steps. Moreover, we investigated the prediction performance with the introduction of a four-layer SAE model, of which structure is $\{x, m^1, m^2, k\}$. The results are shown in Table 4.

Table 4 shows that the combination of DM and DBOW can achieve better results than single model. However, the vector dimension generated by Doc2Vec is too high and contains

Table 3.
Parameters selection of Doc2Vec model

Parameters combination	MAE	RMSE
m = DM, s = 100	1.358	1.646
m = DM, s = 200	1.191	1.561
m = DBOW, s = 100, w = 5	1.131	1.658
m = DBOW, s = 200, w = 10	1.338	1.724
m = DBOW, s = 100, w = 10	1.530	1.959
m = DBOW, s = 200, w = 5	1.284	1.780

Table 4.
The comparison of DM-DBOW and SAE + DM-DBOW

Model	MAE	RMSE
DM-DBOW	1.073	1.339
SAE + DM-DBOW	0.894	1.239

a lot of noise, which will result in serious over-fitting problems. The introduction of auto-encoder can help to handle this problem. Table 4 shows that MAE and RMSE are improved after the text vector is processed by auto-encoder, which indicates better prediction performance.

We derived the text features of prediction model based on DM-DBOW and SAE model and integrated them with the financial features to form the feature matrix of the prediction model. Then the noise of target variable (stock closing price) is smoothed by wavelet transform. Finally, we used "Meinian Health" as the predicted object. We chronologically divided the text and financial data of "Meinian Health" into three parts, i.e. the first 80% data is train set, the middle 10% is validation set and the last 10% data is test set. The train set and the validation set are used to train the model and to adjust parameters and the test set is used to verify performance of models.

In the proposed prediction methods, LSTM is chosen as the prediction model, which also has some parameters to be identified. The most commonly used parameters of LSTM include hidden layers, dropout, number of neurons, optimizer, batch sizes and epochs. The hidden layer is set to 1–3 layers according to experience and the computing power of the machine. Dropout is usually between 0.2 and 0.5. According to Kolmogorov's theorem, the number of neurons in the hidden layer is set as double of the input dimensions plus one (Greff *et al.*, 2017). Finally, we adjusted the parameters of LSTM where the optimal parameter combination was 2 hidden layers, 7 time windows, 85 neurons, 0.5 dropout, Adam optimizer, 4 batch sizes and 50 epochs.

4.4 Results

To verify model performance improvement by combining text features and financial features, we compare the performance of the LSTM model with and without text features. The LSTM model without text features is denoted as LSTM-F. The experimental results are shown in Table 5.

The result shows that the performance of our model is better than that of the LSTM model using only financial features in MAE, RMSE and R^2. It suggests that effective text information mining in social media can effectively improve the performance of the prediction model. To further test the proposed method's convergence and robustness, a piecewise time series prediction method is introduced. Data are divided into 10 groups. Each group of data is predicted and paired T-test is carried out. Experimental results also prove the effectiveness of this method as shown in Table 6.

Table 5. Performance comparison of Doc-W-LSTM and LSTM-F

Model	MAE	RMSE	R^2
Doc-W-LSTM	0.019	0.110	0.957
LSTM-F	0.046	0.579	0.774

Table 6. Performance comparison of Doc-W-LSTM and LSTM-F with t-test

Model	MAE	RMSE	R^2
Doc-W-LSTM	0.029^{***}	0.325^{**}	0.907^{***}
LSTM-F	0.060	0.756	0.747

Notes: $^{**}p < 0.05$; $^{***}p < 0.001$

In addition, we use ARIMA, RNN and LSTM models for comparison. The reason why we choose these three models for comparison is that ARIMA, RNN and LSTM models consider the time dimension of data and are well-behaved models in time series problems (Vo *et al.*, 2019). ARIMA model only uses financial features and the other two models use the combination of 21-dimension text vector that is directly trained by Doc2Vec and financial features. The results are shown in Table 7 and Figure 3.

The results show that the ARIMA model cannot process the non-stationary time series data and the fitting is very poor. RNN and LSTM models consider a variety of influential factors and time dimension at the same time and can generally fit the test data. The metrics of the Doc-W-LSTM model (MAE = 0.019, RMSE = 0.110, $R^2 = 0.957$) are better than other baseline models. Figure 3 intuitively observes that the Doc-W-LSTM model fits the real curve better than other models, proving that the proposed method can effectively predict the fluctuation of stock prices.

5. Conclusion

We propose a new method to predict stock prices. This method adopts Doc2Vec to train financial social media documents and to extract text feature vectors. Then, SAE is used to reduce the dimension of text vectors to avoid a serious imbalance between text features and financial features. Moreover, to avoid the impact of random noise in stock price data on the prediction model, we use Haar wavelet transform to generate denoised stock price time-series data. Finally, we combine the text features and financial features and use the LSTM model to predict future stock prices. Experimental results show that the proposed method is superior to other baseline methods in MAE, RMSE and R^2. It suggests that our method which incorporates text feature information can better predict stock prices.

The main contribution of this paper includes two parts. First, we propose a new stock price prediction method combining text features from social media, which improves the performance of traditional methods. Social media content contains a lot of important information about the stock. The stock financial index variables can only represent the development trend of the stock price, but the feeling of investors can describe the potential trend of the stock price, which is usually neglected in traditional prediction methods. We use the deep learning technology to extract text features, which can represent investors' sentiment and help to improve prediction performance greatly. Second, we use SAE to solve the problem of unbalanced stock features and text features, which helps to improve the accuracy of stock price prediction methods. The traditional dimensionality reduction methods are mainly statistical methods based on word frequency or PCA, but these methods will cause information loss in original data (Nassirtoussi *et al.*, 2014). In our method, the dimension of text features is reduced by the SAE method, which is proved to be an excellent method to reduce the data dimension and preserve information from the original data as much as possible (Wang *et al.*, 2016).

This study has several limitations that can provide new directions for future studies. First, we only collected social media text data from one platform. Although we collected as

Table 7.
Testing result of different models

Model	MAE	RMSE	R^2
ARIMA	1.465	1.455	−0.140
RNN	0.435	0.301	0.882
LSTM	0.385	0.240	0.906
Doc-W-LSTM	0.019	0.110	0.957

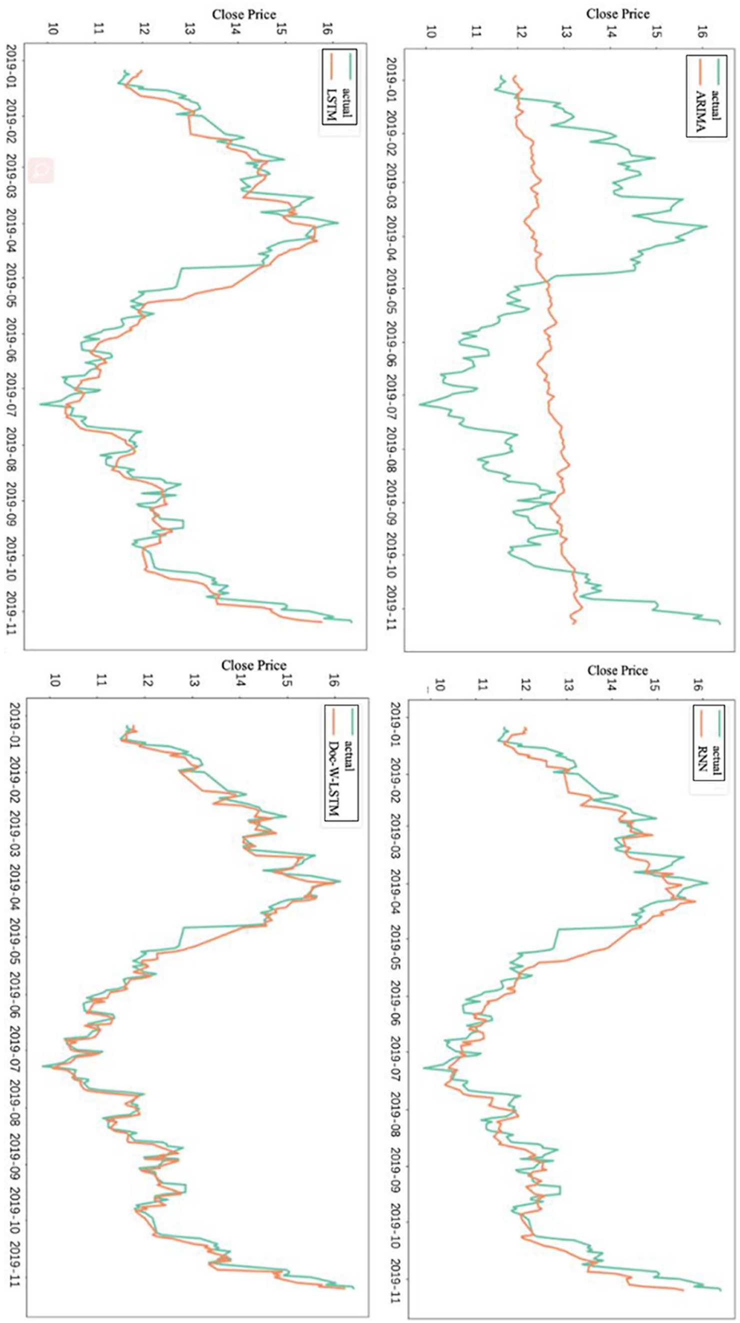

Figure 3.
Fitting curve of
different models

much data as possible from large companies, the investors of other platforms may present different emotions and one website is less representative. We will try to collect more financial social media documents from different platforms in the future. Second, only one stock is selected for prediction in our study. In the future, the relationship among multiple stocks can be considered and the prices of multiple stocks can be predicted to further verify the robustness of the proposed method.

References

Abramovich, F., Besbeas, P. and Sapatinas, T. (2002), "Empirical Bayes approach to block wavelet function estimation", *Computational Statistics and Data Analysis*, Vol. 39 No. 4, pp. 435-451.

Achkar, R., Elias-Sleiman, F., Ezzidine, H., Haidar, N. and Ieee (2018), "Comparison of BPA-MLP and LSTM-RNN for stocks prediction", in *2018 6th International Symposium on Computational and Business Intelligence*, pp. 48-51.

Baek, Y. and Kim, H.Y. (2018), "ModAugNet: a new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module", *Expert Systems with Applications*, Vol. 113, pp. 457-480.

Bao, W., Yue, J. and Rao, Y.L. (2017), "A deep learning framework for financial time series using stacked autoencoders and long-short term memory", *Plos One*, Vol. 12 No. 7, p. e0180944.

Bollen, J., Mao, H. and Zeng, X. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2 No. 1, pp. 1-8.

Booth, G.G., Martikainen, T., Sarkar, S.K., Virtanen, I. and Yliolli, P. (1994), "Nonolinear dependence in Finnish stock returns", *European Journal of Operational Research*, Vol. 74 No. 2, pp. 273-283.

Breidt, F.J., Crato, N. and de Lima, P. (1998), "The detection and estimation of long memory in stochastic volatility", *Journal of Econometrics*, Vol. 83 Nos 1/2, pp. 325-348.

Cervello-Royo, R., Guijarro, F. and Michniuk, K. (2015), "Stock market trading rule based on pattern recognition and technical analysis: forecasting the DJIA index with intraday data", *Expert Systems with Applications*, Vol. 42 No. 14, pp. 5963-5975.

Chen, Y., Lin, Z., Zhao, X., Wang, G. and Gu, Y. (2014), "Deep learning-based classification of hyperspectral data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 7 No. 6, pp. 2094-2107.

Delong, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. (1990), "Noise trader risk in financial-markets", *Journal of Political Economy*, Vol. 98 No. 4, pp. 703-738.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015), "Deep learning for Event-Driven stock prediction", in Yang, Q. and Wooldridge, M. (Eds), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2327-2333.

Engle, R. (2001), "GARCH 101: the use of ARCH/GARCH models in applied econometrics", *Journal of Economic Perspectives*, Vol. 15 No. 4, pp. 157-168.

Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R. and Schmidhuber, J. (2017), "LSTM: a search space odyssey", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, pp. 2222-2232.

Hagenau, M., Liebmann, M. and Neumann, D. (2013), "Automated news reading: stock price prediction based on financial news using context-capturing features", *Decision Support Systems*, Vol. 55 No. 3, pp. 685-697.

Huang, C.J., Liao, J.J., Yang, D.X., Chang, T.Y. and Luo, Y.C. (2010), "Realization of a news dissemination agent based on weighted association rules and text mining techniques", *Expert Systems with Applications*, Vol. 37 No. 9, pp. 6409-6413.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W. and Ramakrishnan, N. (2013), "Forex-Foreteller: currency trend modeling using news articles", *19th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 1470-1473.

Kim, T. and Kim, H.Y. (2019), "Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data", *Plos One*, Vol. 14 No. 2, p. e0212320.

Kim, H.K., Kim, H. and Cho, S. (2017), "Bag-of-concepts: comprehending document representation through clustering words in distributed representation", *Neurocomputing*, Vol. 266, pp. 336-352.

Kim, D., Seo, D., Cho, S. and Kang, P. (2019), "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec", *Information Sciences*, Vol. 477, pp. 15-29.

Kraus, M. and Feuerriegel, S. (2017), "Decision support from financial disclosures with deep neural networks and transfer learning", *Decision Support Systems*, Vol. 104, pp. 38-48.

Lau, J.H. and Baldwin, T. (2016), "An empirical evaluation of Doc2vec with practical insights into document embedding generation", *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin, Germany, pp. 78-86.

Le, Q.V. and Mikolov, T. (2014), "Distributed representations of sentences and documents", The 31st International Conference on Machine Learning (ICML-14*)*, pp. 1188-1196.

Le, L. and Xie, Y. (2018), "Recurrent embedding kernel for predicting stock daily direction", in Sill, A. and Spillner, J. (Eds), *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies*, pp. 160-166.

Maknickas, A. and Maknickiene, N. (2019), "Support system for trading in exchange market by distributional forecasting model", *Informatica*, Vol. 30 No. 1, pp. 73-90.

Marmer, V. (2008), "Nonlinearity, nonstationarity, and spurious forecasts", *Journal of Econometrics*, Vol. 142 No. 1, pp. 1-27.

M'ng, J.C.P. and Mehralizadeh, M. (2016), "Forecasting east Asian indices futures via a novel hybrid of Wavelet-PCA denoising and artificial neural network models", *Plos One*, Vol. 11, p. e0156338.

Nassirtoussi, A.K., Aghabozorgi, S., Teh, Y.W. and Ngo, D.C.L. (2014), "Text mining for market prediction: a systematic review", *Expert Systems with Applications*, Vol. 41 No. 16, pp. 7653-7670.

Nelson, D.M.Q., Pereira, A.C.M. and de Oliveira, R.A. (2017), "Stock market's price movement prediction with LSTM neural networks", in *2017 International Joint Conference on Neural Networks*, pp. 1419-1426.

Papagiannaki, K., Taft, N., Zhang, Z.L. and Diot, C. (2005), "Long-term forecasting of internet backbone traffic", *IEEE Transactions on Neural Networks*, Vol. 16 No. 5, pp. 1110-1124.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015), "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", *Expert Systems with Applications*, Vol. 42 No. 1, pp. 259-268.

Peng, Y., Liu, Y. and Zhang, R. (2019), "Modeling and analysis of stock price forecast based on LSTM", *Computer Engineering and Application*, Vol. 55, pp. 209-212. (in Chinese).

Quan, Z.Y. (2013), "Stock prediction by searching similar candlestick charts", in Chan, C.Y., Lu, J., Norvag, K. and Tanin, E. (Eds), *2013 IEEE 29th International Conference on Data Engineering Workshops*, pp. 322-325.

Ramsey, J.B. (1999), "The contribution of wavelets to the analysis of economic and financial data", *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, Vol. 357 No. 1760, pp. 2593-2606.

Refenes, A.N., Zapranis, A. and Francis, G. (1994), "Stock performance modeling using neural networks – a comparative-study with regression-models", *Neural Networks*, Vol. 7 No. 2, pp. 375-388.

Schölkopf, B., Platt, J. and Hofmann, T. (2007), "Greedy layer-wise training of deep networks", *Advances in Neural Information Processing Systems*, Vol. 19, pp. 153-160.

Schumaker, R.P. and Chen, H. (2009), “Textual analysis of stock market prediction using breaking financial news: the AZFinText system”, *ACM Transactions on Information Systems*, Vol. 27 No. 2.

Shleifer, A. and Vishny, R.W. (1997), “The limits of arbitrage”, *The Journal of Finance*, Vol. 52 No. 1, pp. 35-55.

Singh, R. and Srivastava, S. (2017), “Stock prediction using deep learning”, *Multimedia Tools and Applications*, Vol. 76 No. 18, pp. 18569-18584.

Vo, N.N.Y., He, X., Liu, S. and Xu, G. (2019), “Deep learning for decision making and the optimization of socially responsible investments and portfolio”, *Decision Support Systems*, Vol. 124, UNSP 113097.

Wang, Y., Yao, H. and Zhao, S. (2016), “Auto-encoder based dimensionality reduction”, *Neurocomputing*, Vol. 184, pp. 232-242.

Xie, X., Lei, X. and Zhao, Y. (2020), “Application of mutual information and improved PCA dimensionality reduction algorithm in stock price forecasting”, *Computer Engineering and Applications*, in Chinese.

Zhang, G.S. and Zhang, X.D. (2016), “A Differential-Information based ARMAD-GARCH stock price forecasting model”, *Systems Engineering – Theory and Practice*, Vol. 36, pp. 1136-1145 (in Chinese).

Zhang, Q., Yang, L.T., Chen, Z. and Li, P. (2018), “A survey on deep learning for big data”, *Information Fusion*, Vol. 42, pp. 146-157.

Zhou, Z., Ke, X. and Jichang, Z. (2018), “Tales of emotion and stock in China: volatility, causality and prediction”, *World Wide Web-Internet and Web Information Systems*, Vol. 21, pp. 1093-1116.

Zubiaga, A. (2018), “A longitudinal assessment of the persistence of twitter datasets”, *Journal of the Association for Information Science and Technology*, Vol. 69 No. 8, pp. 974-984.

Corresponding author
Zhijun Yan can be contacted at: yanzhijun@bit.edu.cn
