Expert Systems With Applications 178 (2021) 115019

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction

Nan Jing [a], Zhao Wu [b], Hefei Wang [c,*]

[a] Department of Information Management, SHU-UTS SILC Business School, Shanghai University, Shanghai 201800, People's Republic of China
[b] Department of Information Management, SHU-UTS SILC Business School, Shanghai University, Shanghai 201800, People's Republic of China
[c] International College, Renmin University of China, Beijing 100872, People's Republic of China

ARTICLE INFO

Keywords:
Investor sentiment
Deep learning
Stock market prediction
LSTM
CNN

ABSTRACT

Whether stock prices are predictable has been the center of debate in academia. In this paper, we propose a hybrid model that combines a deep learning approach with a sentiment analysis model for stock price prediction. We employ a Convolutional Neural Network model for classifying the investors' hidden sentiments, which are extracted from a major stock forum. We then propose a hybrid research model by applying the Long Short-Term Memory (LSTM) Neural Network approach for analyzing the technical indicators from the stock market and the sentiment analysis results from the first step. Furthermore, this work has conducted real-life experiments from six key industries of three time intervals on the Shanghai Stock Exchange (SSE) to validate the effectiveness and applicability of the proposed model. The experiment results indicate that the proposed model has achieved better performance in classifying investor sentiments than the baseline classifiers, and this hybrid approach performs better in predicting stock prices compared to the single model and the models without sentiment analysis.

1. Introduction

There is a long-standing debate as to whether stock prices are predictable. Under the Efficient Market Hypothesis (Fama, 1965), stock prices reflect all available information. This implies that stock prices are unpredictable using historical information, and rules out the possibility of finding undervalued stocks or predicting trends in the market through either fundamental or technical analysis. Yet there is also a large body of evidence that points to the contrary. Fundamental analysis and technical analysis remained major tools for retail and institutional investors alike. More specifically, theory and evidences suggest that those who are better at processing information have an edge in stock market investment (Pedersen, 2015). Traditionally, such information processing and analyzing is the goal of financial analysts. However, there have been an abundance of research suggesting that financial analysts are subject to conflicts of interest that may lead to very biased recommendations (see Michaely & Womack, 1999, and subsequent research on this topic). With the rise of social media, "the institution of financial analysis risks becoming de-professionalized, in the same way that many jobs became commoditized by the use of new tools or access to information, the era of DIY [do-it-yourself] financial analysis is dawning."[1] Traditionally the domain of professional forecasters, financial analysis is increasingly being performed and broadcast by investors themselves. Scholars and practitioners are also turning their attention to the popularity of online investment forums to seek if additional information can be extract from investor discussions that help predict stock price movements. Indeed, there is plenty of empirical evidence that suggests investor sentiments in particular may be a critical factor in explaining stock price movements (Schumaker & Chen, 2009, Sprenger, Tumasjan, Sandner, & Welpe, 2014, Chen, De, Hu, & Hwang, 2014).

There are several studies that combine textual data from online social networks and historical data of stock prices to make predictions (Khadjeh, Aghabozorgi, Wah, & Ngo, 2014). The majority of them have concentrated on predicting the future movement of stocks as a classification task, most of which is a binary-class (e.g., up &down) classification problem (Junqué de Fortuny, De Smedt, Martens, & Daelemans, 2014). Meanwhile, very few of them attempted to present models to predict the future price directly based on time series analysis. Moreover, these studies always take news as the only source to analyze the investor sentiment (Huang, Liao, Yang, Chang, & Luo, 2010; Jin et al., 2013; Junqué de Fortuny et al., 2014; Li, Xie, Chen, Wang, & Deng, 2014). However, Friesen and Weller (2006) argue that the cognition and

* Corresponding author.
E-mail addresses: jingnan@shu.edu.cn (N. Jing), allenwu@shu.edu.cn (Z. Wu), wanghefei@ruc.edu.cn (H. Wang).

[1] Quote by Horace Dediu, former analyst, now blogger at Asymco.

https://doi.org/10.1016/j.eswa.2021.115019
Received 16 May 2020; Received in revised form 7 March 2021; Accepted 8 April 2021
Available online 13 April 2021

Table 1
Related work on the stock market prediction by machine learning. Popular techniques include SVR and ANN.

References	Data	Period	Frequency	Technique	Output	Performance measures
Hillebrand and Medeiros (2010)	NYSE	3 January 1995-31 December 2005	Daily	Bagging, log-linear, NN	Volatility	RMSE
Kara et al. (2011)	ISE National 100	2 January 1997-31 December 2007	Daily	SVM, ANN	Movement (Increase, Decrease)	Prediction accuracy
Kazem et al. (2013)	NASDAQ	12 September 2007 –11 November 2011	Daily	SVR + FA	Price	MSE/MAPE
Adebiyi, Adewumi, and Ayo (2014)	NYSE	18 August 1988–25 February 2011	Daily	ANN	Price	Forecast Error
Göçken et al. (2016)	BIST 100	8 June 2005-27 May 2013	Daily	ANN, HS	Index	MAE/RMSE/MARE/MSRE/ RMSRE/MAPE/MSPE
Qiu, Song, and Akagi (2016)	Nikkei 225	November 1993-July 2013	Monthly	ANN + GA	Return	MAE/NMSE/RMSE/MI
Zhang et al. (2016)	SZSE, NASDAQ	4 January 2010-22 May 2015	Daily	SVM, AdaBoost, GA	Trend (Buy, Sell, Others)	Modified accuracy
Chong et al. (2017)	KOSPI	4 January 2010-30 December 2014	5 min	DNN	Return	MSE
Tashiro et al. (2019).	TSE	1 July 2013-30 June 2014	30 sec	Improved CNN	Two classes (Up, Down)/Three classes (Up, Down, Neutral)	F1-score, Precision,
Chen and Ge (2019)	HKEX	1 January 2005-31 December 2017	Daily	LSTM/AttLSTM	Price	Accuracy
Sirignano and Cont (2019)	NASDAQ	1 January 2014-31 March 2017	Daily	LSTM	Direction (up, down.)	Accuracy
Our paper	SSE	3 time intervals	Daily	LSTM+CNN	Price	MAPE

understanding of the same piece of news vary from investors to investors because of the biases like conservatism, overconfidence, and loss-aversion.

In this paper, to overcome these limitations, we propose a hybrid model for stock market analysis and price prediction by combining a deep learning approach with a sentiment analysis model. First, we calculate the sentiment factor by classifying the investors' hidden sentiments based on the Convolutional Neural Network model. Furthermore, an LSTM neural network approach is applied for one-day-ahead closing price prediction, taking the technical indicators derived from stock quotes and the investor sentiment factor in a major stock forum as its input. Our goal is to design and build a more accurate model to predict the stock price and implement this model to evaluate its accuracy based on the real-world stock data from the Shanghai Stock Exchange (SSE).

The Chinese A share market has a distinct feature that retail investors account for over 80% of the trading volume. Therefore, we believe we are most likely to find evidence of technical indicators and investor sentiments having some predictive power for stocks listed in SSE. Indeed, we find strong evidences that investors sentiments combined with technical indicators have predictive power of future stock prices. Moreover, we demonstrate that our sentiment classification has additional predictive power over using technical indicators alone. Earlier studies on investor sentiment analysis suggest that sentiments on stock forums tend to follow naïve momentum strategy, thus making these groups of investors simple trend followers. We use technical indicators such as trend and momentum indicators to control for this time series momentum effect. Our results show that controlling for the momentum factors, our sentiment classification provide additional predictive power for stock price movement.

The remainder of the paper is organized as follows. Section 2 provides an overview of recent literature related to both financial market prediction and social media sentiment analysis. In Section 3, we introduce the hybrid model and give a detailed introduction to the sentiment analysis model and the deep learning approach in this hybrid model. Section 4 presents the experiments that are conducted in this work to validate the proposed model and discusses the experiment results in comparison with other classifier and similar prediction models. Finally, Section 5 concludes the findings of this work and points out the direction for its future efforts.

2. Literature review

Using time series analysis for stock market prediction has always been a central topic in academic finance research. Classic models include the autoregressive conditional heteroscedasticity (ARCH) model, the generalized autoregressive conditional heteroscedasticity (GARCH) model, and more recently using machine learning approaches such as Support Vector Machine (SVM), Artificial Neural Network (ANN) for understanding and predicting the financial market volatility (Engle, 1982; Bollerslev, 1986; Chong, Han, & Park, 2017). Researchers have also attempted to incorporate alternative factors from online social and news media in their proposed models for financial market analysis (Cavalcante, Brasileiro, Souza, Nobrega, & Oliveira, 2016). Some success has been achieved using sentiment analysis to identify and extract investors' subjective attitudes towards the financial market and a specific listed company. More recent works used deep learning methods for text mining and sentiment analysis (Khadjeh et al., 2014). This section briefly reviews the related literatures on stock market prediction especially by applying sentiment analysis in social media and machine learning methods on time series analysis.

2.1. Stock market prediction using machine learning time series analysis

Machine learning methods have been applied to financial time series analysis in recent years, in augmentation to the traditional classic models such as ARCH and GARCH models. These approaches have proven to be effective for the high-dimensionality and non-linearity of the stock market data. Table 1 summarizes the studies that incorporate the use of machine learning methods related to our paper.

According to Kara, Acar Boyacioglu, and Baykan (2011), both ANN and SVM are significantly effective in predicting the stock price movement, and selecting model parameters. Kazem, Sharifi, Hussain, Saberi, and Hussain (2013) also present a chaotic firefly algorithm to optimize the hyperparameters of the SVR model, which outperforms other genetic algorithm-based, firefly-based SVR models and ANN significantly by producing the lowest mean squared error (MSE) and mean absolute percentage error (MAPE) in predicting stock prices from NASDAQ, indicating that individual machine learning approaches are inadequate for making accurate predictions. Furthermore, two special Neural Networks, CNN and LSTM, are widely used in recent works (Tashiro, Matsushima, Izumi, & Sakaji, 2019; Chen & Ge, 2019; Sirignano & Cont, 2019).

Table 2
Related work on investor sentiment analysis in social media. Popular techniques include Support Vector Machines (SVM), Naive Bayes (NB), and Sentiment dictionary.

References	Text	Source	Period	Technique	Output	Measures
Li (2010)	Corporate filings	Securities and Exchange Commission	1994–2007	Naive Bayes	Positive, Neutral, Negative, Uncertain	Success rate
Huang et al. (2010)0	News headlines	Taiwan electronic newspapers	June 2005-October 2005	Weighted association rule	Significance degree	Precision/Recall
Groth and Muntermann (2011)	Disclosures	Quoted companies	Not mentioned	SVM	Positive, Negative	Accuracy/Precision/ Recall/F1/AUC
Yu et al. (2013)	Posts	Blog, forum, news and micro blog	1 July-30 September 2011	Naïve Bayes	Positive, Negative	Accuracy/F-measure
Jin et al. (2013)	News	Bloomberg	April 2010-March 2013	Latent Dirichlet Allocation (LDA)	30 Topics	Precision/Recall
Junqué de Fortuny et al. (2014)	News	Flemish newspapers	June 2005-March 2012	Lexicon	Positive, Negative	Accuracy/AUC
Li et al. (2014)	News	FINET	January 2003-March 2008	Sentiment dictionary	Positive, Negative, Neutral	Accuracy
Wang et al. (2018)	Posts	Sina Weibo	2012–2015	Sentiment dictionary	Positive, Negative, Neutral	Accuracy/Precision/ Recall/F1/AUC

Several scholars employ various optimization algorithm to improve prediction performance, while an increasing number of studies focus on combining econometric and deep learning approaches instead of a single model to improve prediction performance. Enke and Mehdiyev (2013) propose a fuzzy inference neural network with evolution based on fuzzy clustering which is tested on predicting the future prices of S&P 500. Their results suggest that the newly suggested model works better than other neural network-based models, with the lowest relative mean squared error (RMSE) of 0.9834. Kim and Won (2018) compare several hybrid models combining econometric models with neural networks to predict the realized volatility of KOSPI 200 index. Among these proposed models, the GEW-LSTM model combining all three models, GARCH, EGARCH, and EWMA with LSTM, shows the best prediction performance which has the lower errors (MAE, MSE, HMAE, and HMSE) than the single and double GARCH hybrid models. The majority of above referenced studies take advantage of technical indicators like the Relative Strength Index (RSI) and Moving Average, Convergence-Divergence (MACD) as the prediction model input in empirical analysis (Cavalcante et al., 2016). Nevertheless, only a few of studies combine these indicators with other additional inputs, one of which is the public sentiment extracted from online social media.

2.2. Investor sentiment analysis in social media

Prior work in the time series analysis has recognized the significance of investor sentiment in modeling financial markets. For the study of sentiment analysis, models based on machine learning and deep learning are extensively used (Li, Fong, Zhuang, & Khoury, 2015). Huang et al. (2010) analyze financial new headlines by weighted association rules to help investors make investment decisions. Jin et al. (2013) build a prediction model combines news events and financial data to predict the movement of foreign currency. Groth and Muntermann (2011) make use of corporate disclosures by four machine learning techniques for risk management. An interesting finding shows that SVM performs the best compared to Naïve Bayes, k-Nearest Neighbor, and Neural Network. In general, the majority of scholars quantify the investor sentiment by analyzing news from public media outlets such as Taiwan electronic newspapers (Huang et al., 2010), Bloomberg (Jin et al., 2013), Flemish newspapers (Junqué de Fortuny et al., 2014), and FINET (Li et al., 2014). However, the opinion formed from a specific news event varies from investor to investor. Some studies attempt to extract the sentiment implied in the posts from online social media such as Twitter, blogs, and forums where investors share their views and opinion more directly (Yu, Duan, & Cao, 2013; Wang, Xu, & Zheng, 2018).

Table 2 lists related work on investment sentiment analysis. In

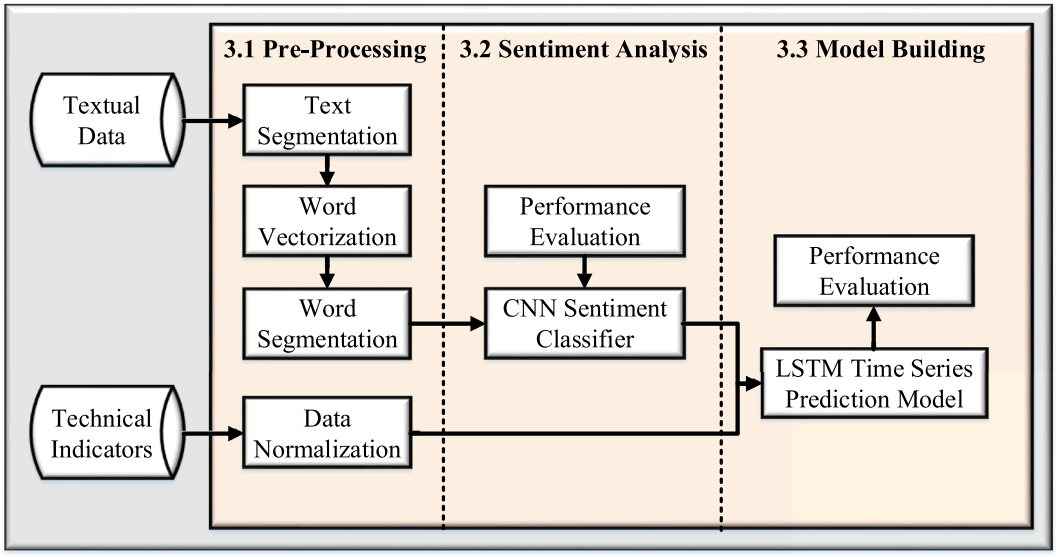

Fig. 1. The structure of the hybrid model.

Table 3
Illustration of pre-processing for the Chinese Language.

Type	data
Sentence	公司的三季报出来了, 我感觉数据还行。 (The company's third-quarter report is available now, and in my opinion, its data are satisfactory.)
Words	公司|的|三季报|出来|了|, |我|感觉|数据 |还行|。 (The company |'s | third-quarter report | is available | now, and in | my opinion | data | are satisfactory|.)
Useful Words	公司|三季报|出来|我|感觉|数据 |还行| (company | third-quarter report | available | my | opinions | data | satisfactory)

addition to the conventional machine learning techniques, the neural networks have proven effective in such text classification tasks (Al-Smadi, Qawasmeh, Al-Ayyoub, Jararweh, & Gupta, 2018; Chen, Yan, & Wong, 2018).

It is found in earlier work on investor sentiment analysis that stock forum sentiments tend to be trend following (Dewally, 2003). To control for the momentum effect, we use technical indicators as controls. There are quite a few studies that use such indicators in stock market predictions, but none to the best of our knowledge that combine the technical indicators with investors sentiments. To fill these gaps, this paper proposes a hybrid model that combines the investor sentiment derived from social media with the technical indicators like Moving Average (MA), Relative Strength Index (RSI) and Momentum Index (MOM) to predict the time series of stock prices.

3. A hybrid prediction model based on the LSTM approach and CNN classifier

As mentioned earlier, the Artificial Neural Network (ANN) is considered suitable for stock market prediction due to its capability of handling nonlinear, discontinuous, and high-frequency data, without the prior knowledge regarding the distribution of the input data. In this paper, we propose a hybrid prediction model that combines a deep learning approach with a sentiment analysis model. This model is composed of three main components: pre-processing data, analyzing investor sentiments and building a prediction model, as shown in Fig. 1. The first component processes the textual and technical data used for the model. Then, the second component employs a sentiment analysis model based on a Convolutional Neural Network to classify the textual data from the online social network. Finally, by combining the sentiment analysis results and the stock quotes, the hybrid model applies LSTM neural network approach for predicting stock prices.

3.1. Pre-processing the textual and technical data

3.1.1. Textual data of Chinese posts in the forum

Text pre-processing filters noise in textual data and transforms the textual data into data that the sentiment classifiers can understand. Compared with the English language, there are several differences in Chinese when carrying out sentiment analysis, one of which is the word segmentation of the Chinese sentence while there is an apparent division between words in English sentences. Therefore, the first step of pre-processing is to split these textual data into words. Then, eliminating stop word, which means filtering the words without any contribution to the sentiment of a sentence, for dimension reduction and increased efficiency in the following learning models.

Table 3 takes a sentence of a post as an example to illustrate the pre-processing for the Chinese language. The first row is the raw sentence data, which means that "the company's third-quarter report is available now, and in my opinion, its data are satisfactory" in English. The second row are the words after segmentation and the last row are the words after filtering out the stop words. To apply these words in deep learning algorithms, it is necessary to transfer data of such words into a vector

Table 4
Word vector representation.

Words	d_1	d_2	...	d_k
公司(company)	−0.38	0.30	…	−0.73
三季报(third-quarter report)	0.71	−0.35	…	0.13
出来(available)	−0.28	1.31	…	0.45
我(my)	0.31	0.78	…	−1.43
感觉(opinion)	1.26	−0.67	…	0.80
数据(data)	0.65	1.22	…	0.74
还行(satisfactory)	1.23	0.65	…	0.77
Average	0.50	0.46	…	0.10

Table 5
Technical features in detail.

Types	Technical indicators	Abbreviation
Trend indicators	Moving average (5)	MA(5)
	Moving average (30)	MA(30)
	Moving average (60)	MA(60)
	Exponential moving average (5)	EMA(5)
	Exponential moving average (30)	EMA(30)
	Exponential moving average (60)	EMA(60)
	Moving Average Convergence/ Divergence (6,15,6)	MACD(6,15,6)
	Moving Average Convergence/ Divergence (12,26,9)	MACD(12,26,9)
	Moving Average Convergence/ Divergence (30,60,30)	MACD (30,60,30)
Momentum indicators	Relative Strength Index (14)	RSI(14)
	Williams' %R (14)	WILLR(14)
	Momentum index (14)	MOM(14)
	Chande Momentum Oscillator (14)	CMO(14)
	Ultimate Oscillator (7,14,28)	ULTOSC (7,14,28)
Volumeindicators	On Balance Volume	OBV
	Chaikin A/D Oscillator (3,10)	ADOSC(3,10)

representation. In this model, we apply Word2vec, a deep learning-based tool that was released by Google in 2013, to get the word vectors (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This tool uses two major architectures, the continuous bag of words (CBOW) model and the Skip-gram model to obtain the vector representation of words. The CBOW architecture predicts the central word based on the context, and the Skip-gram model predicts the context based on the central word. When building the word2vec model, the paper selects a Chinese corpus obtained from Wikipedia for training. Assuming that there are n words in the text feature of a sentence, a two-dimensional matrix of $n \times k$ is obtained after using word2vec. The number of rows n is the number of words in a sentence, and the number of columns k is the dimension of the word vectors. According to Mikolov, Chen, Corrado, and Dean (2013), the range of the dimension is generally from 100 to 800. All the values are investigated in the experiment to find an optimal classification accuracy. Table 4 shows the word vectors of the previous sentence. Each word has been assigned a corresponding vector based on the corpus. The cosine similarity between the vectors indicates the level of similarity between the words represented by m variables in those vectors (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013).

3.1.2. Technical indicators derived from stock quotes

In time series analysis of the financial market, there are several popular technical indicators that contribute to predicting the market trend and stock prices. The most mainstream technical indicators in algorithmic stock trading includes trend indicators, momentum indicators, and volume indicators (Hu et al., 2015). In addition, with different parameter settings, these technical factors are converted into specific features (Da Costa, Nazário, Bergo, Sobreiro, & Kimura, 2015). Based on that, this paper selects 16 technical indicators with different parameter settings in three different types (trend, momentum and volume), as listed in Table 5.

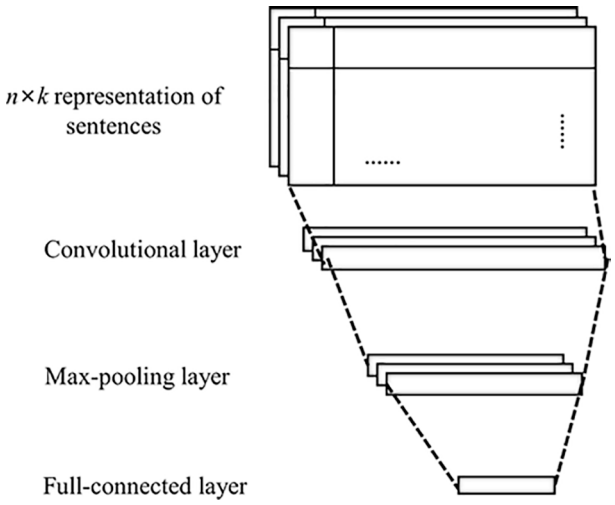

Fig. 2. Sentiment analysis model based on CNN. The size of the network's input is $n \times k$.

These technical features are scaled to the range [0, 1] after data normalization by Eq. (1):

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

3.2. Sentiment analysis using the Convolutional Neural Network

3.2.1. Sentiment analysis model based on CNN

After data preparation, we build a sentiment analysis model based on the Convolutional Neural Network (CNN) to classify the sentiment polarity of posts in the forum. As a special type of neural networks, CNN is often applied in the Natural Language Processing (NLP) domain (Collobert et al., 2011; Hassan & Mahmood, 2018). According to Lee, Jeong, Seo, Kim, and Kang (2018), applying a reduced number of layers with a relatively large number of filters (e.g. 128 in their work) often achieves higher classification accuracies in text classification. Therefore, we set 128 filters on the convolutional layer and our CNN classifier contains the data input, one convolutional layer, one max-pooling layer, and one fully connected layer, as shown in Fig. 2.

In the input layer, given an unclassified text t consisted of n words, R^k is the k-dimension word vector and $w^i \in R^k$ is the i-th word's vector in this text. Based on this, t is represented as Eq. (2):

$$w^{1:n} = w^1 \oplus w^2 \oplus \cdots \oplus w^n \tag{2}$$

where $\oplus$ represents the concatenation operator. For example, "公司三季报出来我感觉数据还行" (The company's third-quarter report is available now, and in my opinion, its data are satisfactory.) is composed of "公司+三季报+出来+我+感觉+数据+还行" (company + third-quarter report + available + my + opinions + data + satisfactory), which means this text is represented by the vector of these words.

Next, in the convolutional layer, the convolution operation is performed by sliding a convolution kernel with the size of $h \times k$ at the input layer to obtain a feature map c. Such process is shown in Eq. (3):

$$c^j = f(m \cdot w^{j:j+h-1} + b) \tag{3}$$

where c^j represents the j-th feature value after the conventional process. f is the convolutional kernel function which is nonlinear. m is a convolution kernel ($m \in R^{h \times k}$, h represents the number of words in m, while k represents the dimension of the word vector), and b is the bias, both of which will be learned in the training stage.

To extract the most important features, max-over-time pooling operation is applied in pooling layer. The maximum value of the feature is captured by the following equation:

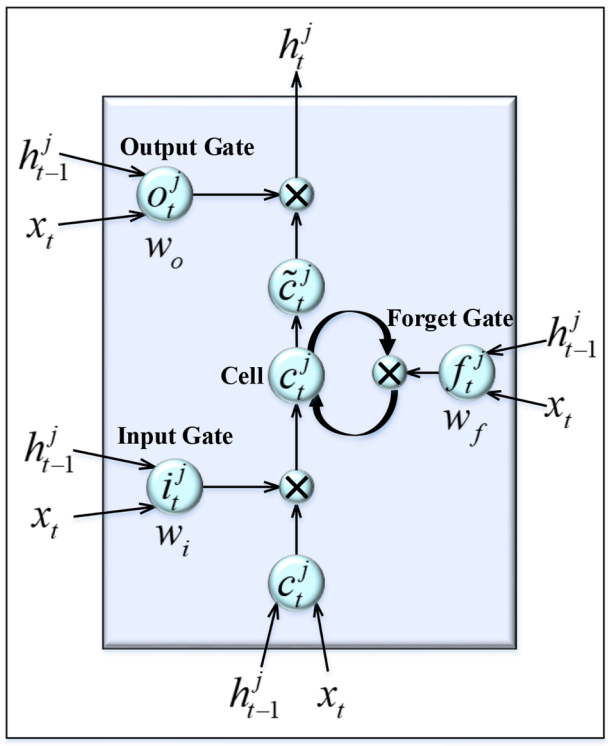

Fig. 3. The architecture of our LSTM.

$$\hat{c} = max(c^j) \tag{4}$$

After the max-pooling process, the output features $v = [\hat{c}_1, \hat{c}_1, \cdots, \hat{c}_k]$ is the input of the fully-connected layer. The final layer applies soft-max classifier to obtain the classification result, as shown in Eq. (5):

$$P(y = l|v; \theta) = \frac{e^v \cdot \theta_l}{\sum_{k=1}^{K} e^v \cdot \theta_k} \tag{5}$$

$P(y = l|v; \theta)$ is the probability that this text belongs to the label l. In this case, the labels are defined as negative and positive. A part of textual data is labelled for model training and testing. The data in a training set is used to construct a sentiment classifier, and the data in a test set is used to evaluate its performance. Meanwhile, four classifiers by Support Vector Machines (SVM), Naive Bayes (NB), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) Neural Network are also employed with the same data in the empirical study for comparisons. Subsequently, we calculate a daily sentiment factor (SF^t) based on the predicted sentiment of textual data by the Eq. (6):

$$SF^t = \frac{num^t_+ - num^t_-}{num^t} \tag{6}$$

where t represents the date and num^t represents the total number of the posts in date t. num^t_+ and num^t_- mean the total number of the posts classified into positive and negative sentiment respectively in date t.

3.2.2. Classification model evaluation index measures

For classification tasks, the most commonly used measures are precision and recall rates. However, precision and recall rates alone cannot show the model performance completely, since the actual data sets are always imbalanced. Therefore, we calculate the F-measure, the harmonic average of precision and recall rates. These measures are defined in the following equations:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

Table 6
The detailed information of hyperparameters and their ranges.

Hyperparameter	Description	Range
Layers size	the number of layers in a LSTM model	(1,3,5,10)
Hidden units size	the number of hidden units per layer	(5,10,20,30,50)
Learning rate	how much to change the model in response to the estimated error	(0.0001,0.001,0.01,0,1)
Epochs size	the number of times that the LSTM passes through the entire training	(10,50,100,200)
Batch size	the number of samples that will be propagated through the LSTM	(10, 20, 40, 60, 80, 100)

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F-measure = \frac{2 \times precision \times recall}{precision + recall} \tag{9}$$

where TP represents the "true positive", which means the total number of positive samples which are correctly predicted. FP and FN stand for the "false positive" and the "false negative", which mean the total number of samples that incorrectly predicted in the positive class and the negative class respectively.

3.3. Stock prediction based on LSTM Neural Network

3.3.1. Prediction model based on LSTM

As a special type of recurrent neural networks (RNNs), LSTM is particularly designed to avoid the problem of exploding and vanishing gradients for sequential data (Hochreiter & Schmidhuber, 1997). Memory cells and gates are applied to exploit long-term information and ignore the useless information to discover the relationships in time series data. This study builds a time series prediction model combining the sentiment factor and technical indicators based on the LSTM neural network. At time t, the proposed prediction model defines the jth neuron in the LSTM as shown in Fig. 3. x_t is the input data while h_t is the hidden state at time t. There are a memory cell c_t^j and three gates: the input gate i_t^j, the forget gate f_t^j, and the output gate o_t^j in the jth neuron. Besides, $\widetilde{c}_t^j$ is the input modulate gate to filter new information. i_t^j, f_t^j, o_t^j, c_t^j, and $\widetilde{c}_t^j$ are defined in the following equations:

$$i_t^j = \sigma(U_i x_t + w_i h_{t-1} + b_i)^j \tag{10}$$

$$f_t^j = \sigma(U_f x_t + w_f h_{t-1} + b_f)^j \tag{11}$$

$$o_t^j = \sigma(U_o x_t + w_o h_{t-1} + b_o)^j \tag{12}$$

$$c_t^j = f_t^j c_{t-1}^j + f_t^j \widetilde{c}_t^j \tag{13}$$

$$\widetilde{c}_t^j = tanh(U_c x_t + w_c h_{t-1} + b_c)^j \tag{14}$$

$$h_t^j = o_t tanh(c_t^j) \tag{15}$$

Table 7
A portion of the selected industries in SSE.

Industry	The number of stocks	Market Value (million yuans)	Average price
Finance	65	8,970,000	6.69
Manufacturing	829	8,960,000	9.52
Mining	51	2,920,000	6.76
Energy	66	1,260,000	5.77
Transportation	73	1,240,000	5.45
Construction	50	1,080,000	6.2

where U_i, U_f, U_o, U_c, w_i, w_f, w_o, and w_c are weight matrices in this neurons, b_i, b_f, b_o, and b_c are the bias terms, and $\sigma(\cdot)$ is a sigmoid function. Firstly, the forget gate f_t^j is used to control how much information from the previous cell state is forgotten by outputting a value from zero to one. Subsequently, the neuron employs the input gate i_t^j to control the added amount of new information. After filtering the information of the previous cell state and new information, the output gate o_t^j determines the output h_t^j of this neuron. Therefore, our prediction model eliminates the effects of the meaningless information and store necessary long-term information. In each stock, sentiment factors and technical indicators are applied as the model input and the one-day-ahead closing price is the output of the proposed model. Besides, hyperparameter setting is of great importance to the performance of the prediction models (Greff, Srivastava, Koutnik, Steunebrink, & Schmidhuber, 2017; Lin et al., 2019). Therefore, in our model, we employ a grid search approach to obtain a better hyperparameter setting. For hyperparameter optimization, Grid search performs an exhaustive search process by a specified subset of the hyperparameter setting of the proposed model. The performance of the proposed model under each hyperparameter setting will be evaluated by the Mean Absolute Percentage Error. According to Greff et al. (2017) and Lin et al. (2019), we select five hypermeters (Layers size, Hidden units size, Learning rate, Epochs size, and Batch size) for the proposed model. The detailed information of hyperparameters and their ranges is shown in Table 6.

3.3.2. Prediction model evaluation measure

According to Göçken, Özçalıcı, Boru, and Dosdoğru (2016) and Kim and Won (2018), this work uses the Mean Absolute Percentage Error (MAPE) to evaluate the performance of the prediction model. These MAPE is defined in the following equations:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y}_t}{y_t}\right| \tag{16}$$

where $\widehat{y}_t$ and y_t represent the predicted and the actual closing prices respectively, and n is the total number of predictions.

4. Experiment and results

In order to evaluate our research model, this paper constructs the experiment procedure as follows. Firstly, we test a single LSTM model using only technical indicators to predict the stock price, serving as a comparison for our hybrid model. Secondly, we generate the proposed hybrid model, including three major phases. Data gathering phase mainly introduces the gathering of history stock data and chooses the stock forums on Eastmoney.com, one of the most popular financial information provider websites in China. The second phase is sentiment analysis, which transforms Chinese posts gathered in the previous stage into a daily sentiment factor by the Convolutional Neural Network model. In this phase, we employ 10-fold cross validation to evaluate the classification performance of our sentiment analysis model based on the CNN model. Four models based on Logistic, SVM, RNN, LSTM are conducted with the same dataset as comparisons. Once the sentiment factors and technical indicators are collected, in the final phase, we input the data to train the LSTM model for stock price prediction. After that, we evaluate the prediction results with MAPE in stock and industry level. Finally, we compare the forecast performance of proposed hybrid model with the single model and other similar model in previous works. For robustness, we randomly selecting five stocks of each industry for twenty times to conduct the prediction experiments.

4.1. Data gathering and description

Traditional predictions in the stock market are often constrained to a

Table 8
The stock universe in the SSE.

Industry	Finance	Manufacturing	Mining	Energy	Transportation	Construction
Symbol	600000.SH	600060.SH	600028.SH	600027.SH	600004.SH	600502.SH
	600036.SH	600475.SH	600188.SH	600617.SH	600115.SH	601117.SH
	601099.SH	600660.SH	600508.SH	600874.SH	600350.SH	601618.SH
	601229.SH	600771.SH	600583.SH	600917.SH	601188.SH	601800.SH
	601788.SH	601777.SH	601899.SH	601016.SH	603223.SH	603007.SH

Table 9
The indicator description of 600000.SH.

indicator	mean	std	min	max
MA(5)	12.2350	1.8891	9.3160	16.8100
MA(30)	12.1332	1.7633	9.6310	16.6567
MA(60)	12.0197	1.5818	9.8718	16.4070
EMA(5)	12.2349	1.8864	9.3421	16.7981
EMA(30)	12.1366	1.7438	9.6183	16.5546
EMA(60)	12.0384	1.5687	9.8315	16.1869
MACD(6,15,6)	0.0361	0.1583	−0.3078	0.7754
MACD(12,26,9)	0.0553	0.1774	−0.2860	0.6800
MACD(30,60,30)	0.0982	0.2518	−0.3924	0.6490
RSI(14)	53.6051	12.7818	23.5967	87.8203
WILLR(14)	−49.0061	29.5965	−100.0000	0.0000
MOM(14)	0.0787	0.5200	−1.3600	2.5900
CMO(14)	7.2102	25.5636	−52.8066	75.6406
ULTOSC(7,14,28)	49.8058	6.4460	34.5248	83.4663
OBV	6,161,362	4,449,190	−3285540	15,472,476
ADOSC(3,10)	44,321	339,289	−556759	1,975,632

specific stock or the market index. Although the outputs of these models are in alignment with the actual prices, it is likely to be only effective to the specific time series data and may cause model over-fitting. Therefore, in order to demonstrate the generalizability and applicability of the proposed model, we select six industries, whose market value is up to ¥108,0000 million, from the Shanghai Stock Exchange (SSE), as shown in Table 7.

Subsequently, in each experiment trial, five stocks are randomly selected from each industry to conduct the experiment using our proposed model. Table 8 shows the random stock selection for one experiment trial. This work gathers the stock quotes to calculate the technical indicators and textual data from a popular forum. The period of research is from Jan. 01, 2017 to Jul. 31, 2019 and incorporates 628 trading days. After that, ninteen other random selections are done and the average MAPE of this experiment will be shown in experiment results and we use two other time intervals to evaluate the performance of the proposed model in different period. The remainder of this section will describe the textual data and technical indicators of these stocks in the first random selection in details.

4.1.1. Textual data from Eastmoney.com

This work gathers the textual data from the stock forum in eastmoney.com. As one of the most popular financial websites in China, the post data of each stock in the eastmoney.com stock market can express the sentiments of Chinese investors (Wang et al., 2018). We implement a web crawler based on Python 3.7 to extract the textual data in forum posts. The Python library Request is imported to open URLs and Beautiful Soup is imported for screen-scraping HTML and XML. We obtained a total of around 880 thousand posts from Jan. 01, 2017 to Jul. 31, 2019 automatically by the web crawler for each experiment trial, all of which are written in Chinese. We use a widely adopted Chinese sentiment corpus ChnSentiCorp to train and test our sentiment analysis model. ChnSentiCorp contains three thousand positive textual data and three thousand negative textual data in Chinese. According to Dang, Zhang, and Chen (2010) and Wong (2015), to evaluate the reliability and stability of the sentiment analysis model, 10-fold cross violation is employed to measure its classification performance. Moreover, additional four models including two deep learning models, RNN and LSTM, and two machine learning techniques, SVM and logistic regression, are conducted using the same data for comparison.

Table 10
Performance comparison of each sentiment classifier model.

	CNN	Logistic	SVM	RNN	LSTM
Average precision	0.875	0.79	0.823	0.83	0.844
Average recall	0.823	0.786	0.764	0.806	0.807
Average F-measure	0.8482	0.7880	0.7924	0.8178	0.8251

4.1.2. Technical indicators from Shanghai stock Exchange

For these thirty stocks in six industries, this work obtains all the daily historical data from the API provided by DataYes. Subsequently, we use these stock quotes to calculate the technical indicators. The equations of these technical indicators are from investopedia.com. As an illustration, Table 9 shows the mean, maximum, minimum, and standard deviation (std) of the indicators for 600000.SH from Jan. 01, 2017 to Jul. 31, 2019.

After that, all these indicators are scaled to the range [0, 1] by Eq. (1) in Section 3.1 for eliminating the error caused by the difference of dimensions.

4.2. Experiment results and discussion

4.2.1. Comparison of the sentiment analysis model

In this part, we compare the performance among our proposed model and other models for sentiment analysis on the collected dataset. Subsequently, we apply 10-fold cross validation to evaluate the proposed model using CNN, as well as four different models (logistic, SVM, RNN and LSTM). These four models are also used frequently in sentiment analysis (Khadjeh et al., 2014). Comparing the results from all these evaluations, we find 400-dimensional word vector space gives the best classification performance. That being said, table 10 gives a brief description of the performance measures with average precision, average recall, and average F-measure for all five sentiment classifiers.

As shown in Table 10, after the 10-fold cross validation, sentiment analysis model based on the CNN has the highest average value of precision and recall among these five models. In general, models with neural networks outperform the models using traditional machine learning techniques. The model based on CNN outperforms other models, with the highest average F-measure value of 0. 0.8482. This difference might not seem large enough, however, when considering over 150 million posts, this degree of improvement can achieve an accurate judgment of millions of posts. Therefore, the comparison suggests that it is effective to apply the CNN for analyzing the sentiment hidden in the textual data. These results show that the proposed model including the data preprocessing and classification for textual data collected from the forum is viable and instrumental for sentiment analysis. In summary, we apply the trained CNN to classify the sentiment polarity of unlabeled textual data.

4.2.2. Stock price predictions and analysis

To test the prediction performance of the proposed hybrid model, we apply the sentiment factor and technical indicators of thirty stocks in six

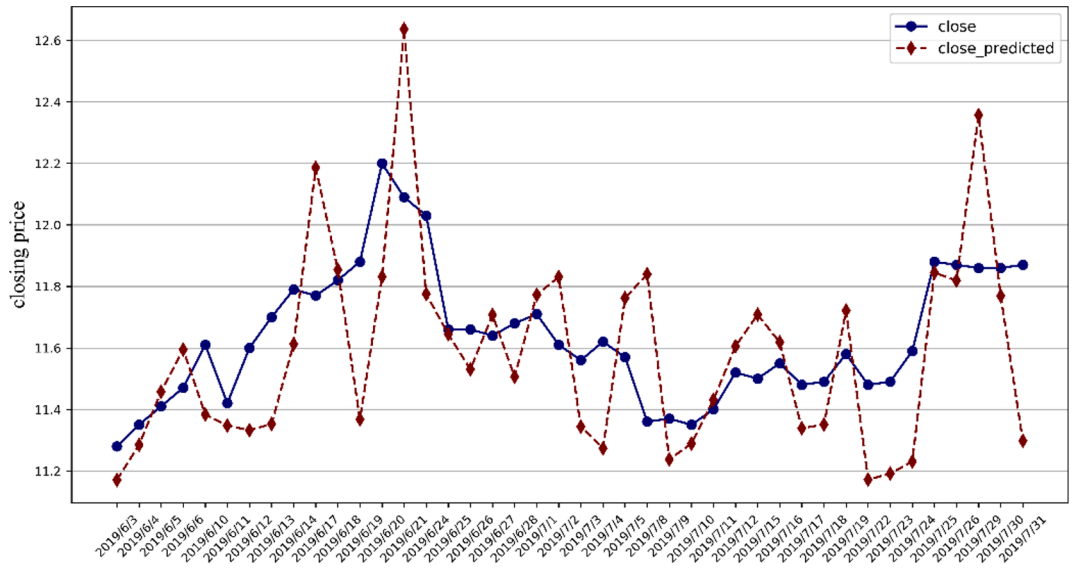

Fig. 4. Actual closing price and predicted price of 600000.SH in finance.

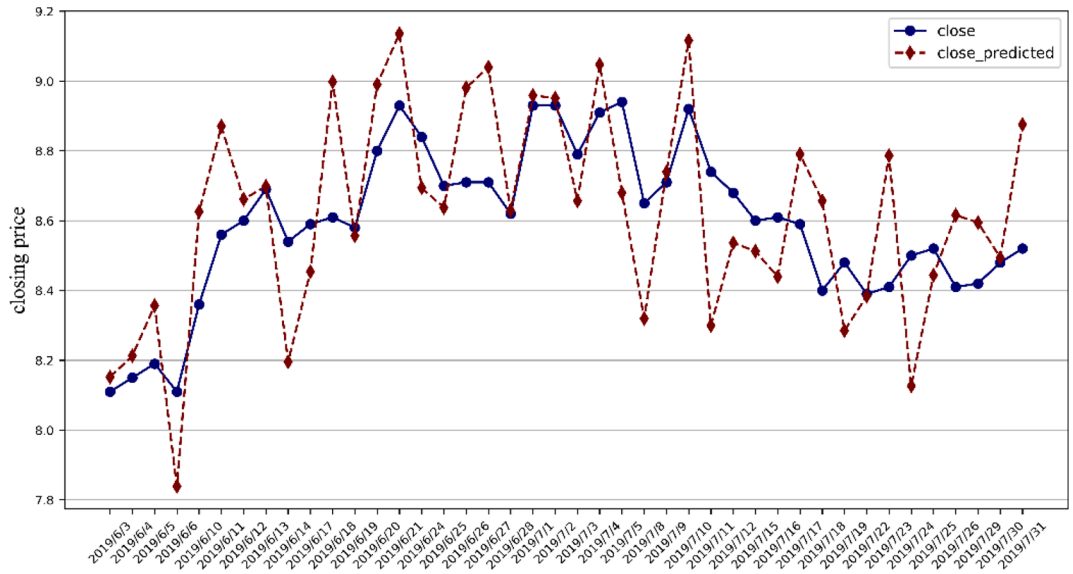

Fig. 5. Actual closing price and predicted price of 600060.SH in manufacturing.

industries from Shanghai Stock Exchange (SSE) as the input data for the one-day-ahead prediction of the closing price. We divide the data into two sets. Data from January 01, 2017 to May. 31, 2019 is used as training data set and data of June and July of 2019 is used as testing data set. The results show that the best combination of hyperparameters after grid search is {Layers size = 5, Hidden units size = 20, Learning rate = 0.01, Epochs size = 100, and Batch size = 10}. Firstly, Figs. 4–9 illustrate the comparison between the predicted price and the actual closing price of one stock (600000.SH, 600060.SH, 600028.SH, 600027.SH, 600004.SH, 600502.SH) from each industry.

As shown in Figs. 4–9, the prediction results of the proposed model are aligned with the actual data, especially in 600004.SH and 600502.SH. Moreover, to be more accurate, we calculate MAPE to evaluate the error between predicted price and actual closing price, shown in Table 11—16.

Table 11–16 show that the MAPE of thirty stocks are below 0.05. It is true that the market value of each industry is different, but we select only five stocks of each industry which cannot have the same impact of the total industries. Especially with the use of machine learning approaches, such difference on the industry level has little effect on the results (Kazem et al., 2013; Zhang, Li, & Pan, 2016). Consequently, we calculate the average value using an arithmetic mean of MAPE in each industry and the overall MAPE in Table 17.

At the industry level, the average value of MAPE in each industry remains steady, which indicates that industry may have little influence on the performance of the proposed model. To evaluate the performance in different conditions, we conduct experiments in two other time intervals. The first time interval (2013.6–2015.12) contains a significant bear market between 2015.6 and 2015.12, when the Chinese stock market have fallen almost 50% from the peak in 2015. As mentioned above, the second time interval goes from 2017.1 to 2019.7. The third interval is the recent Covid-19 pandemic period, when the stock market was also overwhelmed by panic. Besides, nineteen other random selections are completed and the results of experiments in these three periods are shown in Table 18.

Subsequently, to compare with other prediction models, we choose four most used prediction models in similar works as mentioned in the literature review, which are Support Vector Regression, Support Vector

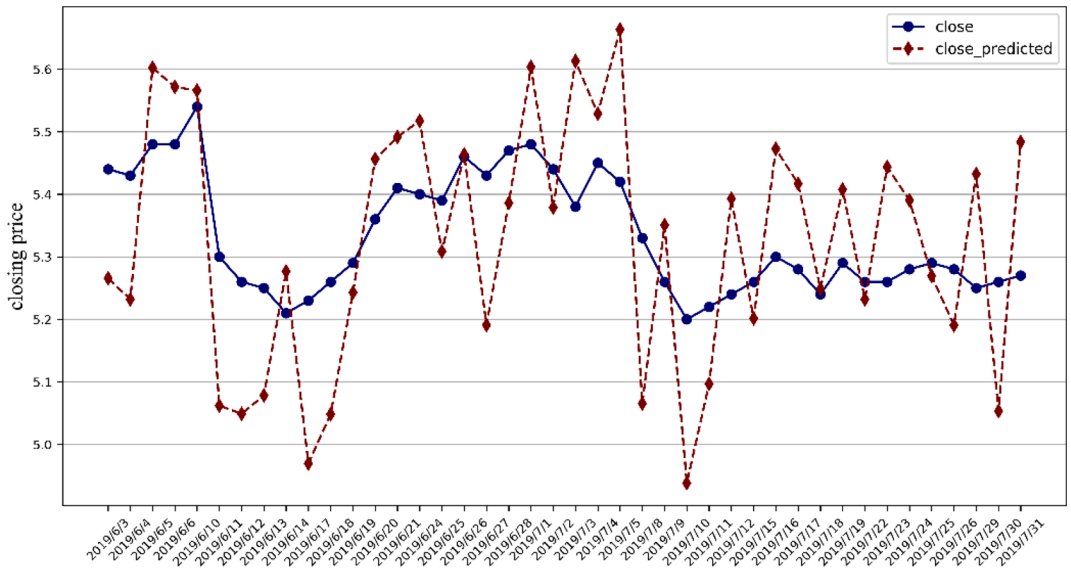

Fig. 6. Actual closing price and predicted price of 600028.SH in mining.

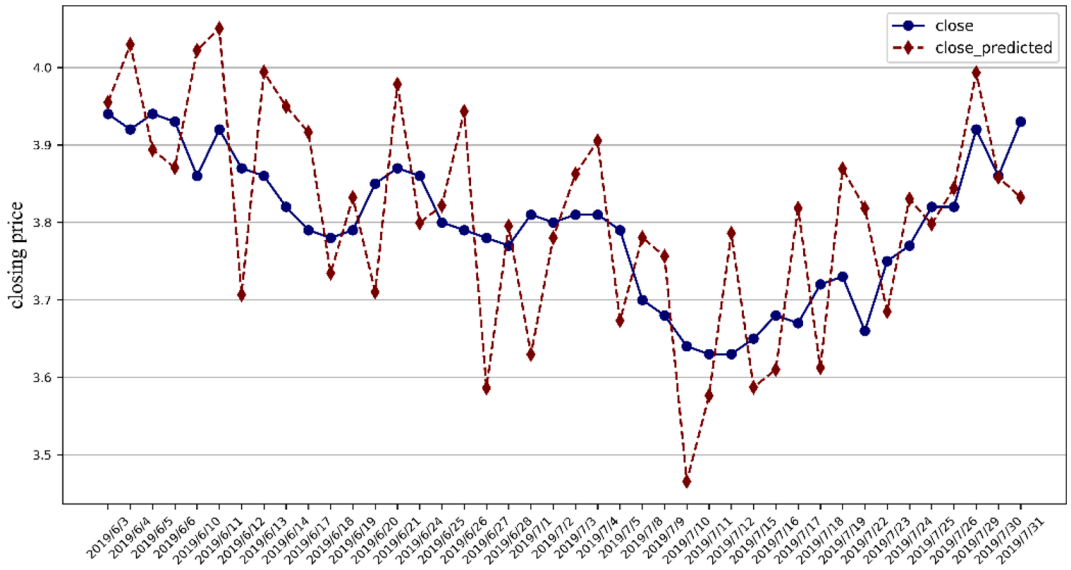

Fig. 7. Actual closing price and predicted price of 600027.SH in energy.

Regression with Genetic Algorithm, Convolutional Neural Networks, and Convolutional Neural Networks with Genetic Algorithm. Then, we implement these models with the same dataset of twenty stock combinations (used in previous experiments) in three periods (2013.6–2015.12, 2017.1–2019.7, 2017.10–2020.4) and the split of training set and test set is the same as in Table 18 as well. We apply the MAPE to measure the model performance on test set. Table 19 lists the similar models proposed in previous works and their results using the same dataset to compare with our model. We also use a single LSTM model based on only technique indicators as a control experiment to show the impact of investor sentiment to the stock price prediction.

Compared to other prediction models with the same datasets, Table 19 shows that the proposed model is ranked first with the MAPE value of 0. 0449. Also, the MAPE of the hybrid model is less than the single LSTM model, which shows that the model has a better prediction performance when integrating with the investor sentiment. Overall, based on experimental results, we conclude that the hybrid model proposed in this paper outperforms other models. The hybrid model integrating deep learning with investor sentiment analysis performs more effectively and accurately in predicting the stock price in the financial market.

5. Conclusion and future work

In this paper, we propose a novel hybrid model combining a deep learning approach with a sentiment analysis model to predict the stock price in the Chinese A-share market. The major contribution of this paper is in the integration of the Long Short-Term Memory (LSTM) Neural Network approach for stock prediction and the Convolutional Neural Network model for sentiment analysis. We propose a sentiment analysis model to extract various investors' perspectives towards these stocks. The hybrid model includes three components: pre-processing data, analyzing sentiment and building the prediction model. In the first phase, we vectorize textual data for sentiment analysis and calculate the technical indicators for the final prediction. Secondly, a sentiment analysis model based on CNN is trained and tested for analyzing investor sentiments. Lastly, we integrate the sentiment factor and technical indicators as the inputs of a LSTM neural network to predict

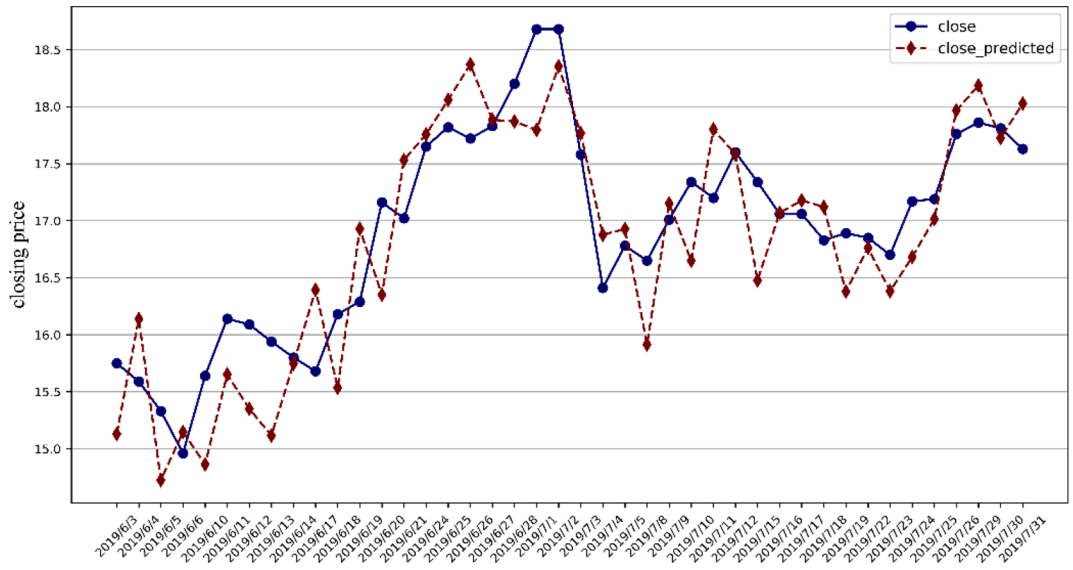

Fig. 8. Actual closing price and predicted price of 600004.SH in transportation.

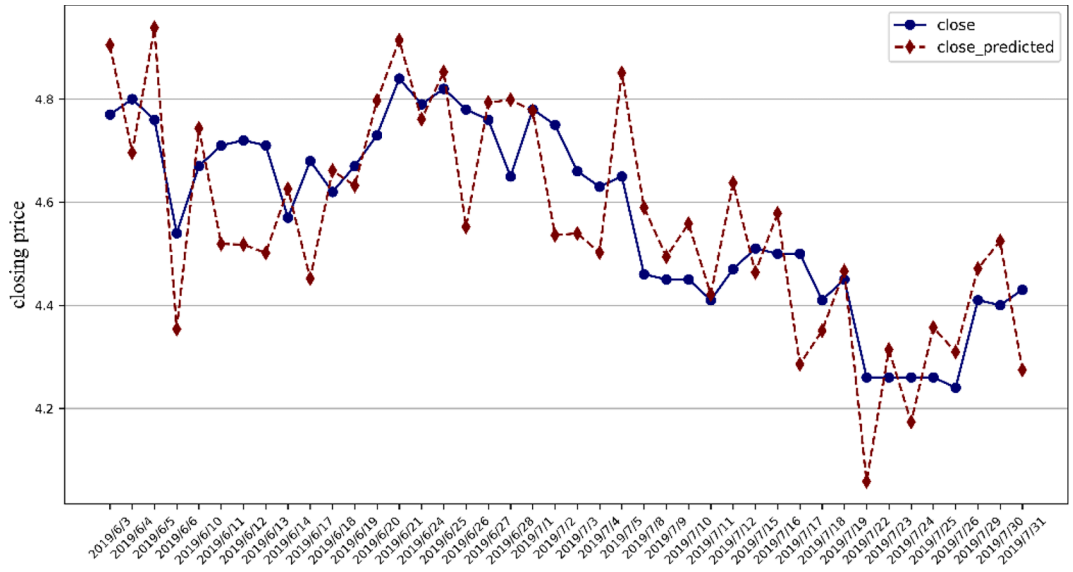

Fig. 9. Actual closing price and predicted price of 600502.SH in construction.

Table 11
MAPE of prediction result in finance.

	600000.SH	600036.SH	601099.SH	601229.SH	601788.SH
MAPE	0.0177	0.0251	0.0215	0.0252	0.0225

Table 12
MAPE of prediction result in manufacturing.

	600060.SH	600475.SH	600660.SH	600771.SH	601777.SH
MAPE	0.0208	0.0256	0.0246	0.0282	0.0225

Table 13
MAPE of prediction result in mining.

	600028.SH	600188.SH	600508.SH	600583.SH	601899.SH
MAPE	0.0412	0.0425	0.0431	0.0333	0.0411

Table 14
MAPE of prediction result in energy.

	600027.SH	600617.SH	600795.SH	600917.SH	601016.SH
MAPE	0.0254	0.0259	0.0246	0.0268	0.0245

Table 15
MAPE of prediction result in transportation.

	600004.SH	600115.SH	600350.SH	601188.SH	603223.SH
MAPE	0.0251	0.0245	0.0236	0.0245	0.0289

Table 16
MAPE of prediction result in construction.

	600502.SH	601117.SH	601618.SH	601800.SH	603007.SH
MAPE	0.0241	0.0292	0.0261	0.0197	0.0242

Table 17
Average MAPE of prediction result.

	Finance	Manufacturing	Mining	Energy	Transportation	Construction	Average
MAPE	0.0224	0.0244	0.0254	0.0247	0.0253	0.0247	0.0245

Table 18
The MAPE results of experiments of twenty random selections in the three periods.

Training set	Test set	Min	Max	Median	Average
2013.6–2015.10	2015.11–2015.12	0.0279	0.0712	0.0460	0.0413
2017.1–2019.5	2019.6–2019.7	0.0217	0.0870	0.0383	0.0403
2017.10–2020.2	2020.3–2020.4	0.0203	0.1061	0.0498	0.0528

Table 19
Comparison of average MAPE of test set between the proposed model and other models.

Model	2015.11–2015.12	2019.6–2019.7	2020.3–2020.4	Average
Support Vector Regression	0.0866	0.0973	0.1043	0.0961
Genetic Algorithm-Support Vector Regression	0.075	0.0994	0.0923	0.0889
Convolutional Neural Networks	0.1058	0.0906	0.0811	0.0925
Genetic Algorithm- Convolutional Neural Networks	0.0658	0.0673	0.0723	0.0685
LSTM	0.0542	0.077	0.0713	0.0675
Hybrid Model (CNN-LSTM)	0.0413	0.0405	0.0528	0.0449

the one-day-ahead closing prices.

In the empirical study, we conduct an experiment to test the classification performance of the CNN model. The sentiment analysis model that uses Logistic, SVM, RNN, and LSTM are trained as controls, with the same data in training and test sets. The result shows that our sentiment analysis model outperforms other models in term of F-measure. Subsequently, the prediction model is built based on the thirty stocks from the Shanghai Stock Exchange (SSE). Our hybrid model with a lower MAPE value of 0.0449 is superior to the single model and other prediction models proposed by the previous work in the literature. We conclude that the combination of investor sentiments and technical indicators based on the LSTM neural network can build a prediction model that brings more accurate prediction for the future stock prices.

In summary, the contributions of the proposed model can be found in the three aspects. The hybrid model that combines a deep learning model approach with a sentiment analysis model provides a comprehensive understanding for the prediction in financial market from both the time series analysis level and the investor sentiment level. The combination of the CNN model for sentiment analysis and the LSTM neural network approach for price prediction give more insight for features extraction from two large set of raw data, stock quotes and textual data. The data preprocessing stage provides a complete process for programmatically manipulating the textual data and converting it to word vectors. Furthermore, the CNN model helps complements the proposed hybrid model by analyzing investor sentiments from stock discussion forums. Furthermore, it is our hope that the procedure and steps to build such a hybrid model provides a well-versed example for a more comprehensive prediction for the financial market based on the time series data analysis including stock quotes and textual data.

However, there still exist several limitations of this study. First of all, we must explain that the rather exceptional prediction accuracy was made using data only for China, a "peculiar" market with trading restrictions and state control over the economy. The predictability of our model partially relies on the fact that the strategy is hardly implementable given the trading restrictions in China. We use the closing price and stock forum discussions of the current trading day to predict the closing price of the next trading day. To exploit such predictability and implement trading strategies accordingly, one needs to do intraday trading, or the so-called T + 0 trading, which is prohibited in China. Finding market predictability for those that are not implementable is not uncommon. It can also be understood as a form of "limits to arbitrage". The study should be replicated to other markets to be reasonably validated.

In addition, more types of social network can be used as data sources to analyze the investor sentiment, such as Twitter, Instagram, Snapchat and online financial news. Furthermore, we applied the grid search for parameter optimization, while optimization algorithms such as differential evolution algorithm and particle swarm optimization are worth investigating to further improve the model's predictive power.

CRediT authorship contribution statement

Nan Jing: Conceptualization, Methodology, Validation, Supervision. **Zhao Wu:** Software, Formal analysis, Data curation, Writing - original draft. **Hefei Wang:** Validation, Funding acquisition, Writing - review & editing.

Declaration of Competing 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.

References

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics, 2014*, 1–7.

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science, 27*, 386–393.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., & Oliveira, A. L. I. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications, 55*, 194–211.

Chen, H., De, P., Hu, Y., & Hwang, B. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies, 27*(5), 1367–1403.

Chen, J., Yan, S., & Wong, K.-C. (2018). Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*.

Chen, S., & Ge, L. (2019). Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction. *Quantitative Finance, 19*(9), 1507–1515.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187–205.
Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(1), 2493–2537.
Da Costa, T. R. C. C., Nazário, R. T., Bergo, G. S. Z., Sobreiro, V. A., & Kimura, H. (2015). Trading System based on the use of technical analysis: A computational experiment. *Journal of Behavioral and Experimental Finance, 6*, 42–55.
Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems, 25*(4), 46–53.
Dewally, M. (2003). Internet investment advice: Investing with a rock of salt. *Financial Analyst Journal, 59*, 65–77.
Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.
Enke, D., & Mehdiyev, N. (2013). Stock market prediction using a combination of stepwise regression analysis, differential evolution-based fuzzy clustering, and a fuzzy inference neural network. *Intelligent Automation & Soft Computing, 19*(4), 636–648.
Fama, E. (1965). The behavior of stock-market prices. *The Journal of Business, 38*(1), 34–105.
Friesen, G., & Weller, P. A. (2006). Quantifying cognitive biases in analyst earnings forecasts. *Journal of Financial Markets, 9*(4), 333–365.
Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction. *Expert Systems with Applications, 44*, 320–331.
Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2222–2232.
Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680–691.
Hassan, A., & Mahmood, A. (2018). Convolutional recurrent deep learning model for sentence classification. *IEEE Access, 6*, 13949–13957.
Hillebrand, E., & Medeiros, M. C. (2010). The benefits of bagging for forecast models of realized volatility. *Econometric Reviews, 29*(5–6), 571–593.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., & Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. *Applied Soft Computing, 36*, 534–551.
Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y., & Luo, Y.-C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications, 37*(9), 6409–6413.
Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., & Ramakrishnan, N. (2013). Forex-foreteller: Currency trend modeling using news articles. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1470–1473.
Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management, 50*(2), 426–441.
Kara, Y., Acar Boyacioglu, M., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications, 38*(5), 5311–5319.
Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing, 13*(2), 947–958.
Khadjeh, N. A., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.
Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103*, 25–37.
Lee, G., Jeong, J., Seo, S., Kim, C., & Kang, P. (2018). Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems, 152*, 70–82.
Li, F. (2010). The Information content of forward-looking statements in corporate filings-A Naïve Bayesian machine learning approach. *Journal of Accounting Research, 48*(5), 1049–1102.
Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2015). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing, 20*(9), 3411–3420.
Lin, H., Shi, C., Wang, B., Chan, M. F., Tang, X., & Ji, W. (2019). Towards real-time respiratory motion prediction based on long short-term memory neural networks. *Physics in Medicine and Biology*.
Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*(1), 14–23.
Michaely, R., & Womack, K. (1999). Conflict of interest and the credibility of underwriter analyst recommendations. *The Review of Financial Studies, 12*(4), 653–686.
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.
Pedersen, L. (2015). *Efficiently inefficient: How Smart Money Invests and Market Prices are Determined*. Princeton University Press.
Qiu, M., Song, Y., & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. *Chaos, Solitons & Fractals, 85*, 1–7.
Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems, 27*(2), 1–19.
Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance, 9*(9), 1449–1459.
Sprenger, T., Tumasjan, A., Sandner, P., & Welpe, I. (2014). Tweets and trades: The Information content of stock microblogs. *European Financial Management, 20*(5), 926–957.
Tashiro, D., Matsushima, H., Izumi, K., & Sakaji, H. (2019). Encoding of high-frequency order information and prediction of short-term stock price by deep learning. *Quantitative Finance, 19*(9), 1499–1506.
Wang, Q., Xu, W., & Zheng, H. (2018). Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing, 299*, 51–61.
Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition, 48*(9), 2839–2846.
Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926.
Zhang, X., Li, A., & Pan, R. (2016). Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine. *Applied Soft Computing, 49*, 385–398.