Multimedia Tools and Applications (2023) 82:27415–27449
https://doi.org/10.1007/s11042-023-14587-8

Exploring the impact of investor's sentiment tendency in varying input window length for stock price prediction

Zhongtian Ji[1] · Peng Wu[2] · Chen Ling[1] · Peng Zhu[1]

Received: 31 July 2021 / Revised: 30 June 2022 / Accepted: 31 January 2023 /
Published online: 14 February 2023

Abstract
Stock price prediction is one of the most important aspects of business investment plans, and has been an attractive research topic for both researchers and financial analysts. Many previous studies indicated the effectiveness of social media sentiment in stock price predictions through time series modelling. However, the time series information hidden in consecutive trading days has not been fully explored. In this paper, we build a stock price prediction model based on attention-based Long Short Term Memory (ALSTM) network using price data, technical indicators and sentiment information from social media. We employed a novel method to feed the deep network with long time series data to learn the deep sequential information of stock price movement. A fine-tuned BERT sentiment classification model and a sentiment lexicon are proposed to extract deep sentiment tendency of social media posts. We conducted experiments on 28 stocks within three years' transaction period, and the results show that: (1) evaluated by the indicators of the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the accuracy, our proposed method outperforms the baseline models in both validation and test data sets; (2) models incorporating stock prices, technical indicators and sentiment features perform better than models that only use partial data source; (3) the fine-tuned

✉ Peng Wu
wupeng@njust.edu.cn

Zhongtian Ji
jizhongtian@njust.edu.cn

Chen Ling
lingchen@njust.edu.cn

Peng Zhu
pzhu@njust.edu.cn

[1] School of Economics and Management, Nanjing University of Science &Technology, 200. Xiaolinwei road., Nanjing 210094 Jiangsu, China

[2] School of Intelligent Manufacturing, Nanjing University of Science &Technology, 200. Xiaolinwei road., Nanjing 210094 Jiangsu, China

BERT model performs better in sentiment classification task, and the exploitation of the sentiment features computed with the use of BERT model also led to higher predicting accuracy compared with the features calculated using sentiment lexicon; and (4) setting the input window length to 5-day achieves the best performance in average prediction accuracy.

Keywords Social media · Sentiment analysis · Fine-tuned BERT · Stock price prediction · Attention-based LSTM

1 Introduction

Stock price prediction is an important task in the planning of investment activities. However, it remains a challenging problem to build an effective stock price prediction model, considering that stock prices are affected by multiple factors. In addition to historical prices and a series of technical indicators, the current stock price is also affected by social sentiment. The overall social mood toward a company may be one of the most significant variables affecting its stock price. Nowadays, with the rapid development of social media, an increasing number of investor posts are released on social media, making large amounts of sentiment data available.

Many prior studies have confirmed the validity of investor's sentiments in stock market predictions [4, 55, 61, 63], even in the Bitcoin exchange market [87]. However, the social media information comprises texts in loose and unrestricted format which grow in a dynamic way. Therefore, this study attempts to integrate and make use of as much content as possible in the social environment of stock market to develop an effective stock prediction method that fully utilizes time series information.

Other drawbacks of previous studies involve using only snapshots of the dataset at time point t to predict another time point in the future [12, 83] and using models that were not tailored for deep sequential information [55]. This ignores the time series relationships among consecutive trading days before time point t, which is also a significant information hidden in the historical time series. LSTM network [29] is designed to learn sequential information, which has been verified to be superior to other models for the task of extracting effective information from complex financial time series data [35, 58]. Therefore, we believe it will help to improve the performances of our prediction method.

To address these questions, we employ four approaches that 1) propose a fine-tuned BERT sentiment classification model and a sentiment lexicon to construct sentiment analysis, 2) convert sentiment information into novel representation feature as model input, 3) build a ALSTM-based architecture to learn the deep sequential information via varying input window length, and 4) conduct experiments on a large scale of social media posts concerning 28 stocks for a period of three years.

This study makes four contributions, namely: (1) we introduce an ALSTM-based architecture for stock price prediction using stock price data, technical indicators and sentiment information, which performs better than the baseline models in both validation and test data sets using three different evaluation metrics; (2) we compare the model performance using different data source, the real effectiveness of sentiment analysis in stock prediction is demonstrated; (3) we propose a fine-tuned BERT sentiment classification model which shows good performance in sentiment classification task, and the exploitation of the sentiment feature computed with the use of the BERT model also lead to higher predicting accuracy compared

with the feature calculated using sentiment lexicon; and finally, (4) we compare the predicting accuracy when using different input window length and found that setting time window to 5-day can improve the average predicting performance for all proposed models. The highest average predicting accuracy of 28 stocks is achieved when using sentiment feature calculated by the fine-tuned BERT model.

The rest of the paper is presented in the following. Section 2 introduces some related works on stock predictions based on price data and technical indicators, predictions combining sentiment analysis, and also predictions using long input window length. Section 3 describes our proposed methodology. Section 4 presents the detailed experimental process and assesses experimental results. Section 5 presents the discussion and implications. Finally, the last section concludes our contribution and proposes future works.

2 Related work

This section summarizes studies on (1) Domain 1: Stock predictions based on price data and technical indicators, (2) Domain 2: Stock predictions based on sentiment analysis, and (3) Domain 3: Stock predictions based on long input window length. Several research gaps are concluded through the summary.

2.1 Stock predictions based on price data and technical indicators

Stock market prediction has been an important task in both academics and businesses. Based on the Efficient Market Hypothesis (EMH) [18], some of the early studies propose that it is impossible, given the risk it may face, to achieve above-market returns over the long term. Therefore, the prediction accuracy of the stock market will not exceed 50% [71]. However, the EMH has been questioned ever since [31, 62], especially with the rapid development of machine learning models [5, 21, 64, 85]. Prediction accuracy of 56% is generally considered as satisfying results [73, 77].

Despite Fama's hypothesis, there are two different philosophies of trading for stock market prediction [8]: fundamental analysis and technical analysis. The former analysis macroeconomic factors, a company's financial conditions, while the latter assumes that future performance are related to certain historical patterns [75] like time-series prices. Several technical indicators are defined to represent these patterns including the moving average (MA) [24], exponential Moving Average (EMA) [37], momentum [43], Bollinger band [23], etc.

Some researchers tried to make stock predictions based on historical prices only [93, 94] or predict by using a small dataset [22]. Due to the low instance test set, the result may be insufficient. Stock markets generate large-scale trading data every day, providing large amounts of training data for deep neural network [47]. Fischer and Krauss [20] applied an LSTM-based model for financial time series predictions, and the result shows that the LSTM network performs better than memory-free classification models, i.e., a random forest, a logistic regression classifier, and a deep neural net.

Studies in Table 1 cover 4 main aspects of work: (a) stock market selection; (b) feature selection; (c) input window length; and (d) predicting method adoption. Each column corresponds to one aspect. As for selection of stock market, these studies choose a continuous period of time for training and testing. As for feature selection, it can be classified as price data (e.g. [28, 86]), or technical indicators (e.g. [93]), or both of them (e.g. [54, 59]). Input window

Table 1 Summary of studies based on price data and technical indicators

Author	Market Objects	Market Period	Feature set	Input window length	Predicting Method	Performance criteria	Main Findings
Yeh, et al. [85]	Taiwan Capitalization Weighted Stock Index	2002.10–2005.12	Price data	1d	SVR, ARIMA, FNN	RMSE	The proposed model performs better than baselines.
Güreşen, et al. [25]	NASDAQ Stock Exchange index	2008.10.07–2-009.06.26	Daily stock exchange rates of NASDAQ	4d	MLP, ANN, GARCH,	MSE, MAD	Multilayer Perceptron yields the best results.
Zuo and Kita [93]	NIKKEI stock average and Toyota motor corporation stock price	1985.02.22–2-008.12.30	Technical indictors	–	Bayesian network, AR, MA, ARMA ARCH	RMSE, correlation coefficient	Compared with the traditional time series predicting algorithm, the prediction accuracy of the proposed model is improved by 15–20%.
Zuo and Kita [94]	FTSE100 index	2005.01–2007.12	Index data	1d	Bayesian network	Accuracy	The correction rate of the proposed model is almost 60% and the profit is much greater than other models.
Ballings, et al. [5]	5767 publicly listed European companies	2009–2010	Price data, Technical indictors	–	LR, NN, K-nearest neighbors, SVM,RF, AdaBoost, Kernel Factory.	AUC	The Random Forest model performs best followed by others.
Patel, et al. [59]	2 stocks and 2 indexs	2003.01–2012.12	Price data, Technical indictors	10d	ANN, SVM, RF, Bayesian network	Accuracy, F1-score	The predicting performance of all models improves when using technical indicators as trend - determined data.
Chandra and Chand [9]	ACI Worldwide, Staples, and Seagate in NASDAQ	2006.12–2010.10	Price data	17d	RNN, ANN	RMSE	Compared with feedforward networks, recurrent neural networks have better

Table 1 (continued)

Author	Market Objects	Market Period	Feature set	Input window length	Predicting Method	Performance criteria	Main Findings
							generalization ability in time series tasks.
Yong, et al. [86]	Singapore Stock Market Index	2010.01.01–2-017.01.03	Price data	10d	Feed-forward DNN	RMSE, MAPE, Profit, SR	The proposed method gets 70.83% profitable trades
Nelson, et al. [54]	5 stocks	2008–2015	Price data, Technical indictors	15 m	LSTM, Multi-Layer Perceptron, RF, pseudo-random model	Accuracy, Precision, Recall, F1	LSTM achieves better accuracy than other Machine Learning methods
M, et al. [28]	5 stocks in NSE	2011.01.03–2-016.12.30	Price data	200d	LSTM, RNN, CNN, MLP	MAPE	Neural networks outperform the linear model (ARIMA).
Fischer and Krauss [20]	S&P 500 stocks	1992.12–2015.10	Volume-weighted--average-prices	240d	LSTM, RF, DNN, LOG	Mean return, Standard deviation, Annualized Sharpe ratio, Accuracy	LSTM networks outperforms memory-free classification methods
Zhang Xiaolin and Tan [88]	Stocks in China's A-share	2006.01.01–2-017.12.31	Technical indictors	–	LSTM, SVR, MLP	AR, IR, IC	The LSTM model performs better in extracting information from raw dataset in predicting stocks future return than several advanced models
Pang X., et al. [58]	Shanghai A-shares composite index, SZSE	2006.01.01–2-016.10.19	Price data	10d	Embedded layer + LSTM	Accuracy, MSE	The LSTM networks with embedded layer has better performance.
Baek and Kim [3]	S&P500, Korea Composite Stock Price Index 200 (KOSPI200) and 20 stocks	2008.02.01–2-017.07.26	Price data, Technical indictors	20d	LSTM, RNN	MSE, MAPE, MAE	A robust model is proposed for financial time-series prediction.
	3 indexs		Price data	–	LSTM, RNN	RMSE, MAE, NMSE	

Table 1 (continued)

Author	Market Objects	Market Period	Feature set	Input window length	Predicting Method	Performance criteria	Main Findings
Zhang Y. a., et al. [92]		2010.01.06–2-018.04.27					The proposed model outperforms baselines in predicting accuracy with higher directional symmetry.
Kumar A., et al. [41]	Google stock	2012.01–2017.01	Price data	–	LSTM, RNN	Loss	The proposed model are good at handing the time-based problems.
our study	28 stocks	2016.11.18–2-019.11.18	Price data, Technical indicator, Sentiment index from GuBa	1d, 3d, 5d, 7d, 10d, 15d, 30d	RNN, ALSTM, SVR	Accuracy, RMSE, MAE	ALSTM model using price data, technical indicators and sentiments from social media achieves highest accuracy when the input window length is set to 5-day.

length is the length of the input vector (e.g., 3d represents a 3-day time window). Some abbreviations are used for this field: 'm' is minutes and 'd' is days. A null value means no relevant information mentioned. As for predicting method adoption, it can be classified as (1) reduced-form models, such as ARIMA (e.g. [85]), GARCH (e.g. [25]) or (2) machine learning models, including Bayesian network (e.g. [94]), SVM (e.g. [5]), SVR (e.g. [88]), or (3) deep learning models, such as ANN (e.g. [9]), RNN (e.g. [3]), LSTM (e.g. [41, 58, 92]).

2.2 Stock predictions based on sentiment analysis

Sentiment analysis, which is mainly designed to understand what others are thinking [57], has been proved effective in many applications including movie reviews [39, 40, 80], product reviews [38] and public opinions [70, 81]. Nowadays, sentiment information extracted from social media for stock market prediction has also been proved to be effective [46, 60]. There are two main sources for the researchers to merge the information extracted from the text content into their financial models. In previous studies, the main source was the news [45, 67, 68], and in recent studies, social media sources [48]. Bollen, et al. [6] conducted the most influential study to gauge specific dimensions of Twitter sentiments in predicting Dow-Jones index and achieved higher predicting accuracy. Since this seminal study, sentiment extracted from Twitter [52, 82], Yahoo! Finance [56], Sina Weibo [83], GuBa [48], etc. has been proven to be highly correlated with the stock market. Xing, et al. [84] mentioned that it is insufficient for investors to make investments only based on public sentiment and other factors must also be considered in prediction models.

There are two main perspectives on sentiment analysis of text contents: sentiment lexicon [15, 30] and natural language processing [1, 32]. Picasso, et al. [61] extracted two distinct sets of sentiment features from sentiment texts based on the dictionary of Loughran and Mcdonald [50] and AffectiveSpace2 [7] separately. The former is a specific dictionary for financial applications while the latter is a vector space model which is designed to extract sentiments from structured content. Their results show that combining sentiments with price technical indicators outperforms using price data only. The employment of AffectiveSpace feature as input achieved higher accuracy, while the use of the features calculated by Loughran and McDonald dictionary achieves higher returns.

As shown in Table 2, these studies include 5 main aspects of work: (a) stock market selection; (b) feature selection; (c) input window length; (d) sentiment analysis method adoption; and (e) predicting method adoption. As for selection of stock market, these studies also focus on a continuous period of time. As for feature selection, these studies add sentiment information into feature set in the form of (1) polar sentiments (e.g. [45], (2) sub-categorical sentiments (e.g. [61, 69]), or (3) sentiment index (e.g. [31]). As for input window length, these studies also focused on a fixed input window length (e.g. [6, 48]). As for sentiment analysis method adoption, it can be classified as sentiment lexicon (e.g. [82]) or natural language processing (e.g. [56]). As for predicting method adoption, machine learning models, including SVM (e.g. [47]), SVR (e.g. [52]) and (2) deep learning models, such as LSTM (e.g. [12]), RNN (e.g. [83]) are commonly used.

2.3 Stock predictions based on long input window length

Stock prediction can be viewed as a time series problem when using long input window length for model training. Given a univariate or a multivariate time-series, one may treat the entire

Table 2 Summary of studies based on sentiment analysis

Author	Market Objects	Market Period	Feature set	Input window length	Sentiment analysis method	Predicting Method	Performance criteria	Main Findings
Sehgal and Song [69]	52 stocks	6 month time period	Sub-categorical sentiment from Yahoo Finance	–	Natural Language Processing	Decision Tree, Naive Bayes, Bagging	Recall, Precision, F1	The stock performance is highly correlated with its recent online sentiments.
Bollen, et al. [6]	Dow-Jones index	The presidential election; Thanksgiving day in 2008	Price data, Polar and Sub-categorical sentiment from Twitter	3d	Sentiment lexicon, Natural Language Processing	Granger Causality, FNN	MAPE, Accuracy	Calm increases the accuracy to 86.7%.
Oh and Sheng [56]	10 stocks	2010.05.11–2010.08.07	Sentiment index from Yahoo Finance and Stocktwits	–	Natural Language Processing	Decision Tree	Accuracy	Micro blog sentiments own predictive power on future movements of stock price.
Vu and Chang [82]	Four tech companies	2011.04.01–2011.05.31	Price data, Sentiment index from Twitter	–	Sentiment lexicon	Decision Tree	Precision, Recall, F1	Sentiment improves the accuracy in predicting rising and falling of stocks.
Li X., et al. [45]	22 stocks in HSI	2003.01–2008.03	Price data, Polar sentiment from News	1d	Sentiment lexicon	SVM	Accuracy	The models with sentiment analysis perform best in both validation set and testing set.
Junqué de Fortuny, et al. [31]	11 stocks	2007.01.01–2012.03.25	Price data, Technical indicator, Sentiment index from News	1d	Natural Language Processing	SVM	Accuracy, AUC	Employing state-of-the-art text-mining methods can predict stock price movements more accurately.
Wang, et al. [83]	SSE Composite Index	2012.01–2015.12	Technical indicators, Sentiment index from Sina Weibo	1d	Sentiment lexicon	ANN, SVM, RF, DBN, CNN, RNN	F1, Precision, Recall, Accuray, AUC	The AUC value of the proposed method increased by at least 14.2% comparing to the baseline models.
Chen M.-Y., et al. [12]	3 stocks	2016.01.04–2017.12.29	Polar sentiment from News	1d	Sentiment lexicon	LSTM	Accuracy	The proposed method can predict the future trend of stock price with higher accuracy.
Picasso, et al. [61]	20 stocks	2017.07.03–2-018.06.14	Price data, Technical indicator, Sub-categorical sentiment from News	1d	Sentiment lexicon, Natural Language Processing	NN, SVM, RF	Accuracy, Recall	Combining sentiments with price technical indicators outperforms using price data only.
Li X., et al. [47]	12 stocks	2003.01–2008.03	Price data, Technical indicators, Sentiment index from News	10d	Sentiment lexicon	LSTM, SVM, MKL	Accuracy, F1	The proposed LSTM model performs better than the MKL and SVM in all evaluation criteria.

Table 2 (continued)

Author	Market Objects	Market Period	Feature set	Input window length	Sentiment analysis method	Predicting Method	Performance criteria	Main Findings
Maqsood, et al. [52]	15 stocks	2000.01–2018.10	Price data, Polar sentiment from Twitter	1d	Sentiment lexicon	SVR, Linear regression, CNN	RMSE, MAE	Sentiments of local and global events improves the stock prediction performance.
Li Y., et al. [48]	CSI 300 index	2009.01.01–2-014.10.31	Price data, Sentiment index from GuBa	10d	Natural Language Processing	LSTM, logistic regression, SVM, Naïve Bayes	Accuracy, Recall, F1	LSTM model owns more predictive power with investor sentiment.
our study	28 stocks	2016.11.18–2-019.11.18	Price data, Technical indicator, Sentiment index from GuBa	1d, 3d, 5d, 7d, 10d, 15d, 30d	Sentiment lexicon, Natural Language Processing	RNN, ALSTM, SVR	Accuracy, RMSE, MAE	ALSTM model using price data, technical indicators and sentiments from social media achieves highest accuracy when the input window length is set to 5-day.

Table 3 Summary of studies based on long input window length

Author	Market Objects	Market Period	Feature set	Input window length	Input data form	Predicting Method	Performance criteria	Main Findings
Oliveira, et al. [15]	PETR4	2000.01.04–-2009.08.18	Price data, Technical indicators, Macroeconomic series	5d, 10d, 15d, 22d	High dimension vector	ANN	RMS, MRPE	The best predicting performance was achieved with a 5-day quotes and a 1-day predicting horizon.
Nguyen, et al. [55]	18 stocks	2012.07.23–-2013.06.19	Price data, Sub-categorical sentiment from message board	2d	One-dimensional vector	SVM	Accuracy	Proposed model with sentiments outperforms other baseline methods in average predicting accuracy.
Verma, et al. [79]	NIFTY50 Index, NIFTY Bank/Auto/IT/-Energy Index	2013.01.01–-2017.01.31	Price data, News	1d, 2d, 5d	High dimension vector	LSTM, SVM	MCC, Accuracy	The LSTM model outperforms linear SVM on different stock indices data.
Shynkevich, et al. [72]	S&P 500 stock market index	2002.01.29–-2008.12.23	Technical indicators	3d, 5d, 7d, 10d, 15d, 20d, 25d, 30d	One-dimensional vector	SVM, ANN, kNN	Accuracy, Return, Winning rate, Sharpe ratio	When the length of the input window is equal to the forecasting horizon, the best predicting performance is obtained.
Zhang L., et al. [89]	50 stocks	2007–2016	Price data	3d, 5d, 10d, 15d, 20d	High dimension vector	AR, LSTM, SFM	Average square error	The proposed model outperforms the AR and the conventional LSTM models on real price data.
Lee C. and Soo [42]	TWSE index, 4 stocks in TWSE	2001.01.01–-2017.02.13	Price data, Technical indicators, News	15d	High dimension vector	CNN+LSTM	RMSE, Profit	The proposed recurrent convolutional neural networks outperforms the technical analysis only.
Mourelatos, et al. [53]	Stock of National Bank of Greece (ETE)	2009.12.22–-2014.12.11	FTSE100, DJIA, GDAX, NIKKEI225, EUR/USD, Gold	1d, 2d, 5d, 10d	High dimension vector	GA-SVR, LSTM	Return, Volatility, Sharpe Ratio, Accuracy	The LSTM network has advantages in most scenarios.
Eapen, et al. [17]	S&P 500 stock market index	2008.01.02–-2018.11.27	Price data	14d, 28d, 50d, 56d	High dimension vector	CNN, BiLSTM, SVR	Accuracy, MSE	The proposed method increases predicting performance by 9% and by over a factor of six

Table 3 (continued)

Author	Market Objects	Market Period	Feature set	Input window length	Input data form	Predicting Method	Performance criteria	Main Findings
								upon SVR model on S&P 500 dataset.
Kim T. and Kim [34]	S&P 500 ETF data	2016.10.14–-2017.10.16	Price data	30 m	High dimension vector	LSTM, CNN, LSTM-CNN,	RMSE, RMAE, MAPE	Incorporating temporal and image features can effectively reduce the prediction error.
Rezaei, et al. [66]	4 stock indexs	2010.01–20-19.09	Price data	250d	One-dimensional vector	CNN, LSTM, CNN-LSTM, EMD, CEEMD	RMSE, MAE, MAPE	CNN alongside LSTM can improve the predicting performance comparing with baseline models.
Long, et al. [49]	14 stocks	2012.03–20-18.06	Price data, Technical indicators	30d	High dimension vector	CNN, BiLSTM	Accuracy, AUC	Compared with baselines, the proposed method achieved best predicting performance in forecasting price movements.
Zhang Y. a., et al. [92]	Stocks in the SSE50 index	2016.01.01–-2019.03.31	Price data	5d	High dimension vector	ANN, LSTM	RMSE, MAPE	The LSTM model outperforms baseline models with the best predicting performance.
our study	28 stocks	2016.11.18–-2019.11.18	Price data, Technical indicator, Sentiment index from GuBa	1d, 3d, 5d, 7d, 10d, 15d, 30d	High dimension vector	RNN, ALSTM, SVR	Accuracy, RMSE, MAE	ALSTM model using price data, technical indicators and sentiments from social media achieves highest accuracy when the input window length is set to 5-day.

time-series as a sample. There has been a lot of interest in predicting through long input window length, and it remains an active research area [15, 91].

Nguyen, et al. [55] extract information from two consecutive days for stock movement prediction. In their study, features of each day are considered to be a parallel relationship and used for the training of SVM. Shynkevich, et al. [72] employ technical indicators to describe the information about the past trend of the stock price. In their research, indicators are regarded as a snapshot of the current situation which also reflect the past behaviour over a certain period of time. Several machine learning algorithms are proposed to train these input features which are calculated from price data through different time span. With the rapid development of computer engineering, deep learning algorithms have been widely used in financial time series modelling tasks. Instead of using indicators calculated from different input window length, these studies consider higher-dimensional input data [17, 34], allowing deep learning networks to learn the hidden sequential information.

As shown in Table 3, the 5 main aspects in these studies include: (a) stock market selection; (b) feature selection; (c) input window length; (d) input data form; and (e) predicting method adoption. Stock market selection and feature selection are trivial. As for input window length, these studies use a relatively long time period (e.g. [49]), or several optional lengths for comparison (e.g. [53, 89]). As for input data form, it can be categorized as one-dimensional vector (e.g. [55, 72]) or high dimension vector (e.g. [42]). As for predicting method adoption, LSTM (e.g. [66, 79]) is most commonly used.

2.4 Summary

Through summarizing and comparing previous researches in above three domains, we identified three issues that warrant further investigation, which follows here.

The first issue is that many previous studies make prediction barely using stock price data and several technical indicators. The booming development of social media accelerates the dissemination of users' opinions and sentiments [44]. Investors tend to seek for emotional help [19], leading the impact of sentiment opinions more significant than usual. Hence, sentiment analysis on social media posts own greater significance in stock prediction task.

The second issue is that the sentiment analysis approaches lack an in-depth understanding of the sentiment text content. Some of the semantics-based methods use sentiment lexicon to analysis the sentiment. However, since the sentiment of the whole content is judged by limited keywords, the deep sentiment in the text may be neglected due to the imperfection of the sentiment lexicon. In order to extract the deep sentiment, an efficient method should be developed. Therefore, we utilize BERT [16] in our sentiment analysis process, inasmuch as it has yielded better results for many NLP tasks including sentiment classification.

The last issue is that the previous studies fail to explore the impact of using long input window lengths on prediction performance. Although many previous studies consider taking long input window length for models to learn, the length number is usually fixed [45, 90], or the input data form lacks time series information [55]. The change of input window length may also result in variation in prediction performance but is seldom considered. Hence, it is of vital importance to discuss the difference of using different input window lengths in prediction.

To settle the three issues, this study build a prediction model based on ALSTM networks using three data sources as input: price data, technical indicators and sentiment feature. The

sentiment feature is extracted from social media posts through two different sentiment analysis methods for comparison. The first one is a manually predefined sentiment polarity lexicon in the financial field, and the second one is a fine-tuned BERT sentiment classification model. Different length of input window is organized to feed the ALSTM networks for comparison. To our knowledge, this paper is one of the earliest attempts to reveal the impact of sentiment analysis via different window lengths for stock price prediction.

3 Methodology

An overview of the research framework is shown in Fig. 1. First, the sentiment posts are analysed and sentiment indicator for each transaction day are calculated. Then the sentiment indicators combining with the time series stock prices and technical indicators are organized as model input. Through learning the past N days' features, the closing price of $N + 1$ day is predicted. Details of each part are explained in the following subsections.

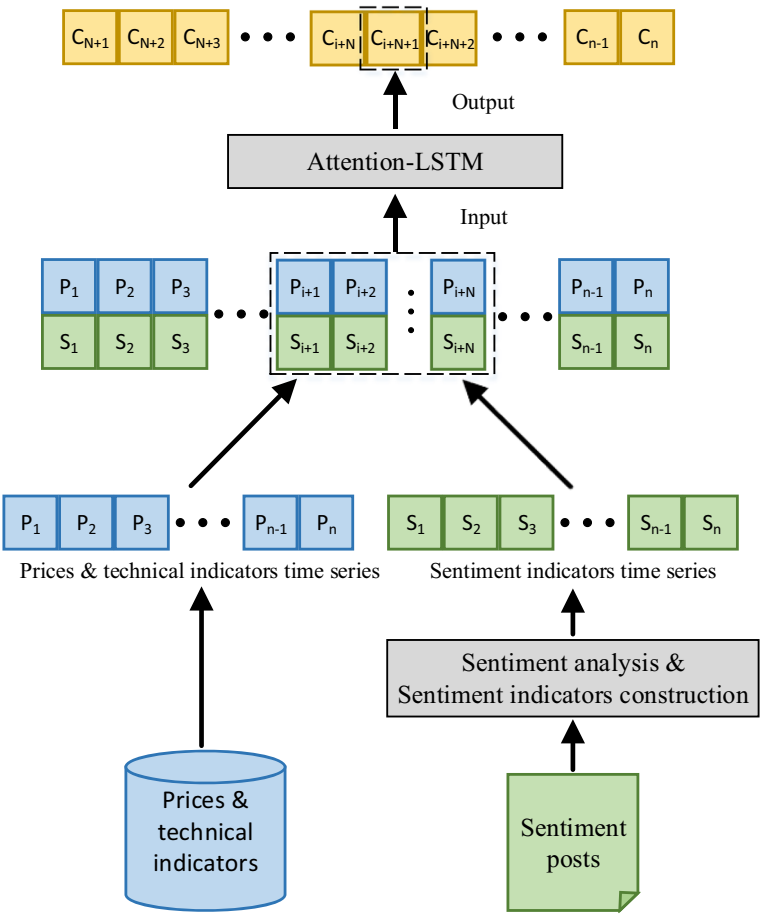

Fig. 1 Illustration of research framework

3.1 Price and technical indicators

In this study, 6 stock price indictors and 8 technical indicators are selected to construct the indicator set.

The stock price data comprises open, close, high, low price, turnover rata and trading volume. Technical indicators are wildly used for market states analysis [3]. Therefore, besides historical prices, we also employ several technical indicators as extra inputs for ALSTM networks which are shown in Table 4. These indicators can reflect the stock trends from multiple aspects, which provides rich stock market signals for the ALSTM networks to learn. However, these technical indicators may not have exact values at every single day due to the different time configuration. Therefore, transaction days with missing values are removed to ensure the integrity of the time series data.

3.2 Sentiment analysis

The sentiments analysis module in Fig. 1 classifies sentiment posts into three categories: positive, negative, and neutral, according to the beliefs or expectations expressed: a positive post means that the mentioned stock price is supposed to rise in the nearly future, or it shows the poster's tendency in buying this stock; a negative post indicates the expectation in price falling or the tendency of selling this stock; and a neutral post means no obvious expectation or recommendation shown in the post and poster has no tendency in trading. These user-generated text contents are processed by two sentiment analysis methods in this study for comparison: a manually constructed sentiment lexicons and a fine-tuned BERT model for sentiment classification.

3.2.1 Sentiment lexicon

Sentiment dictionaries have been widely used in transforming sentimental contents into representations. In this experiment, the National Taiwan University Sentiment Dictionary was used as basic lexicon and extra finance related terms were manually added. These terms are regarded as rise/fall relevant terms which were summarized from online posts and relevant studies for making up the lack of relevance between the original lexicon and the stock market. The new lexicon contains two polar sentiments: positive and negative. Words which are not exist in our lexicon is regarded as the third sentiment dimension – neutral. Based on the natural language processing, three steps are employed to process these online posts. The first step is Chinese word segmentation and unwanted word removal. Unwanted word such as stop words and special characters (@, #, $ etc.) has no role during classification process. By this step, the text sequences for each post is obtained. The second step is sentiment word matching. Through

Table 4 Meanings of technical indicators

Names	Meanings
PE	price-to-earnings ratio
PB	Price-to-book ratio
PS	Price-to-sales ratio
ROE	Return on equity
MA5	5-day moving average
MA10	10-day moving average
EMA5	5-day exponential moving average
EMA10	10-day exponential moving average

this process, the text sequences are matched with our sentiment lexicon, which mark words with tags "positive", "negative" and "neutral". The third step is post sentiment calculation. The sentiment polarity of post j is calculated through Eqs. (1)–(4).

$$PosCount_j = \sum_{i=1}^{T} Pos_{(i,j)} \tag{1}$$

$$NegCount_j = \sum_{i=1}^{T} Neg_{(i,j)} \tag{2}$$

j represents the number of posts; i represents the ith word in text sequence. The $Pos_{(i,\ j)}$ or $Neg_{(i,\ j)}$ indicates weather the ith word is positive or negative respectively. When the word appears in the positive part of our lexicon, $PosCount_j$ is employed as the total positive number. When the word appears in the negative part of our lexicon, $NegCount_j$ is employed as the total negative number. In this study, $PosCount_j$ and $NegCount_j$ are used to represent the extent of expectation on rise and fall.

$$D_j = PosCount_j - NegCount_j \tag{3}$$

$$Sent_j = \begin{cases} \text{Positive if } D_j > 0 \\ \text{Neutral if } D_j = 0 \\ \text{Negative if } D_j < 0 \end{cases} \tag{4}$$

Through Eqs. (3) and (4), the magnitudes of $PosCount_j$ and $NegCount_j$ are compared. When $PosCount_j$ is larger than $NegCount_j$, it means the post has more expectation in the rising of the stock price, and vice versa. D_j is calculated in accordance with $PosCount_j$ and $NegCount_j$ to classify the polarity of the post j into positive, negative and neutral. These marks of sentiment polarity are employed to construct sentiment indicators.

3.2.2 BERT-based sentiment classifier

Besides sentiment lexicon, we also employ BERT, a pre-trained language model based on deep bidirectional Transformers [78], to perform sentiment classification task. We also take advantage of fine-tuned BERT for sentence-level sentiment classification as it has produced state-of-art results for many NLP tasks [26]. The output of this multi-class, single-label sentiment classifier is the predicted probability of each class, and we get the final predicted category (positive, negative or neutral) according to the output probability.

A natural idea for fine-tuning is to further pre-train BERT with target domain data [74] since BERT was trained in the general domain. In this study, we directly fine-tune the pre-trained BERT model with task-specific dataset, which is constructed using randomly selected data from GuBa dataset. The sentiment polarity of each text is manually labelled in the following process. First, we unified the sentimental annotation guideline in the financial fields. Second, a group of five coders completes the first round of sentiment annotation. Then another group of five coders completes the second round of sentiment annotation for the same text contents. Inconsistencies in annotation are judged by a five-coder verification team under final discussion. Finally, it was used in the fine tuning process for the specific task. In this way, we

reduce the limitation of the model performance and endow the model with rich sentiment knowledge.

3.2.3 Construction of sentiment indicators

Sentiment indicators are constructed through sentiment indicators construction method in Fig. 1 based on the sentiment classification results. Following [2, 10, 33], we adopt the bullishness indicator, which is defined as Eq. (5),

$$B_t = \frac{M_t^{pos} - M_t^{neg}}{M_t^{pos} + M_t^{neg}} \tag{5}$$

where $M_t^c = \sum_{i \in D(t)} w_i x_i^c$ is the weighted sum of messages of type $c \in \{pos, neg, neu\}$ in the time interval $D(t)$. x_i^c is equal to one when post i is type c and zero otherwise, and w_i is the weight of the post. Antweiler and Frank [2] reveal that the alternative weighting schemes make no difference to conclusions and employ the equal weighting. Therefore, we also regard M_t^c as the number of posts of different categories. Antweiler and Frank [2] propose another bullishness indicator, which is shown in Eq. (6):

$$B_t^* = \ln\left[\frac{1 + M_t^{pos}}{1 + M_t^{neg}}\right] \tag{6}$$

In order to reflect the number of investors expressing a certain sentiment, they provide an alternative method of calculation, as shown in Eq. (7):

$$B_t^* \approx B_t \ln(1 + (M_t^{pos} + M_t^{neg})) \tag{7}$$

The second measurement of B_t^* outperforms the other one in their research. However, neutral posts are not considered in these bullishness indicators. The neutral posts can also reflect the investors' attention on a particular stock even if they do not contain obvious expectations or beliefs. Considering a more comprehensive investor attention, we propose the following investor sentiment indicator B_t^{all}, as is shown in Eq. (8),

$$B_t^{all} = B_t \ln(1 + M_t) \tag{8}$$

where M_t is the total number of posts at time interval D(t). M_t changes with the investors' attention and is not influenced by the sentiment classification methods.

3.3 Attention-based LSTM networks

In this study, attention-based LSTM networks are chosen as prediction model. LSTM has similar architecture with Recurrent Neural Network (RNN). Recurrent neural network is able to learn temporal patterns from sequential data through internal loops. Weights is learned by backpropagation which has difficulties in retaining long-term information, and may confronts the problem of vanishing (or exploding gradients). LSTM models were proposed to solve these problems [29], and the biggest difference is that there exist three more gates in LSTM.

These gates determine whether each data can pass through the gate and enable LSTM networks to learn long-term dependencies. These three gates are the input gate, forget gate, and output gate. An input gate indicates whether new information can be added into the LSTM memory. A forget gate decides what information should be abandon. An output gate controls whether to output the state. The calculations for the integral process are performed as the following formulas:

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

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

$$\widetilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \tag{11}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{12}$$

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

$$h_t = o_t * \tanh(C_t) \tag{14}$$

where, W_f, W_i, W_c, W_o are weight matrices, b_f, b_i, b_c, b_o are bias vectors, h_t is the memory cell value at time t, σ calculates how much data to keep, f_t is the value of the forget gate layer, i_t shows the values of the input gate, $\widetilde{C}_t$ is the total data reserved at time t, C_t indicates the current cell state, o_t is the output gate layer. The LSTM model comprises these memory blocks and is capable to learn longer temporal patterns.

Attention mechanism is introduced to the LSTM networks, which will adaptively assign different attention weights to different features. After forming the feature vector $H = \{h_1, h_2, \ldots h_T\}$ through the hidden layer, the attention mechanism will look for the attention weight α_i of h_i, and the attention mechanism formula is as follows:

$$e_i = \tanh(W_h h_i + b_h), e_i \in [-1, 1] \tag{15}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{t} \exp(e_i)}, \sum_{i=1}^{t} \alpha_i = 1 \tag{16}$$

where W_h is the weight matrix of h_i. The output of the attention mechanism can be obtained as:

$$\left[h_1^*, h_2^*, \ldots, h_T^*\right] = [h_1, h_2, \ldots, h_T] * [\alpha_1, \alpha_2, \ldots, \alpha_T] \tag{17}$$

where the above $*$ operation is number multiplications componentwise. That is, $h_j^* = h_j * \alpha_j, j = 1, 2, \ldots, T$.

4 Experiments

4.1 Dataset

Two datasets are employed in stock price prediction process. The first one is the stock prices and technical indicators dataset, and the second one – the sentiment information dataset. Stock prices and technical indicators come from the RESET Financial Database (www.resset.com), while the sentiment information comes from GuBa (http://guba.eastmoney.com).

4.1.1 Stock price and technical indicator dataset

All 28 pharmaceutical stocks in the CSI 300 are chosen to conduct experiment. Stock historical prices and technical indicators are collected for a period of three years (from November 18, 2016 to November 18, 2019). Stock codes and company names are shown in the Table 5.

Table 5 Stock codes and company names

Stock codes	Company names
000423.SZ	DEEJ
000538.SZ	YUNNAN BAIYAO
000661.SZ	CHANGCHUN HIGH-TECH
000963.SZ	HUADONG MEDICINE
002001.SZ	NHU
002007.SZ	HUALAN BIO
002044.SZ	HEALTH 100
002252.SZ	SHANGHAI RAAS
002294.SZ	SALUBRIS
002411.SZ	BICON
002422.SZ	INDUSTRY GROUP
002773.SZ	KANGHONG PHARMACEUTICAL
300,003.SZ	LEPU MEDICAL
300,015.SZ	AIER EYE HOSPITAL
300,122.SZ	ZFSW
300,142.SZ	WALVAX
600,085.SH	Tongrentang Chinese Medicine
600,196.SH	FOSUN PHARMA
600,276.SH	HENGRUI MEDICINE
600,332.SH	BYS
600,436.SH	PIEN TZE HUANG
600,535.SH	TASLY HOLDING GROUP
600,566.SH	JUMPCAN
600,867.SH	TONGHUA DONGBAO PHARMACEUTICAL
600,998.SH	JOINTOWN PHARMACEUTICAL GROUP
601,607.SH	SHANGHAI PHARMA
603,259.SH	WuXi AppTec
603,858.SH	BUCHANG PHARMA

There are three reasons for choosing the 28 pharmaceutical stocks in the CSI 300 stocks:

1. Csi 300 stocks have higher capitals comparing with others in the whole A-share market, which means there are more discussions in the GuBa.
2. Negative news about pharmaceutical and biological companies continues to emerge. Increasing attention has been drawn from Chinese society, such as the fraud case of DEEJ and the expired honey case of Tongrentang Chinese Medicine.
3. Choosing stocks in the same industry can reduce the negative impact of the industry factors on stock price prediction.

4.1.2 GuBa dataset

For sentiment indicators constructing, expectations and beliefs need to be extracted from online posts. Text contents of the 28 stocks are collected from GuBa during the same three-year period to build our sentiment information dataset. GuBa is the most representative internet stock message board in China where investors usually share company news, stock price movement predictions, facts, and comments (usually with strong emotional tendencies) on specific company events. Each stock has its own GuBa page where the stock-related posts can be easily accessed. Two examples of GuBa posts published by investors during the three-year period are shown in Fig. 2. The first post shows negative sentiment obviously and the other shows strong optimism about the stock price future trends.

The stock market is closed for weekends and holidays. The posts published from 2:40 pm of the previous transaction date to 2:40 pm of the current transaction date are assigned to the current transaction date. Transaction date over 24 hours are divided by the number of days it covers. Each stock has transaction dates for a three-year period in our dataset.

However, as in other sentiment information sources, posts on the GuBa are also messy. The post content is usually varying in length, riddled with many spelling mistakes, uncommon expressions, redundant HTML links and irrelevant information. Table 6 tabulates the statistics of each transaction date concerning the min, median, mean, max and the total number of the number of posts for each stock after a clean-up pre-processing. Over this three-year period, we accumulated a total of 1,451,272 pieces of data.

Fig. 2 Two GuBa posts published by investors

Table 6 Statistics of each transaction date

Stocks	The number of posts				Total number of posts
	Min	Media	Mean	Max	
DEEJ	1	45	73	1635	79,487
YUNNAN BAIYAO	1	18	30	1373	32,880
CHANGCHUN HIGH-TECH	1	35	47	875	51,427
HUADONG MEDICINE	1	19	34	904	36,236
NHU	1	31	53	435	56,385
HUALAN BIO	1	20	30	458	32,063
HEALTH 100	1	13	28	1053	29,673
SHANGHAI RAAS	1	22	97	2099	104,854
SALUBRIS	1	11	18	502	18,973
BICON	1	9	24	447	24,539
INDUSTRY GROUP	1	21	31	1123	33,043
KANGHONG PHARMACEUTICAL	1	3	6	214	4941
LEPU MEDICAL	1	20	33	381	35,117
AIER EYE HOSPITAL	1	15	28	3259	29,173
ZFSW	1	16	31	246	6623
WALVAX	1	32	44	249	23,201
Tongrentang Chinese Medicine	1	8	19	1833	10,965
FOSUN PHARMA	1	35	51	1298	55,646
HENGRUI MEDICINE	1	51	67	1337	71,628
BYS	2	114	201	2484	220,086
PIEN TZE HUANG	1	36	47	590	50,772
TASLY HOLDING GROUP	1	20	32	360	34,351
JUMPCAN	1	12	27	423	27,410
TONGHUA DONGBAO PHARMACEUTICAL	1	17	36	970	37,518
JOINTOWN PHARMACEUTICAL GROUP	1	9	15	192	14,717
SHANGHAI PHARMA	1	12	22	242	22,427
WuXi AppTec	2	65	204	5177	119,245
BUCHANG PHARMA	1	102	171	5397	187,892

4.2 Baseline setup

In the experiment, Support Vector Regression (SVR) and recurrent neural networks (RNN) are used as baselines.

4.2.1 Support vector regression

First designed by Cortes and Vapnik [14] as a classifier, SVR is employed to capture nonlinear relationship and has a global optimum. Previous studies have reported the effectiveness of SVR in financial time series forecasting problems [27, 64].

In a regression task, given a time-series data set $F = \{(\mathbf{x}_k, y_k)\}_{k=1}^{n}$ derived from an unknown function $y = g(\mathbf{x})$, we need to determine a function $y = f(\mathbf{x})$ based on F and to minimize the difference between f and the unknown function g. The main idea of SVR is to build a mapping $\mathbf{x} \rightarrow \phi(\mathbf{x})$ to a new feature space $\mathbf{X}'$according to the mapping scheme. The nonlinear relationship is then transformed into a linear relationship between the new feature $\phi(\mathbf{x})$ and label y in the new created space. The SVR model can be obtained as

$$y = f(\mathbf{x}, \boldsymbol{\alpha}, b) = \sum_{k} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) + b \tag{18}$$

Where $\mathbf{x}_k$ is support vectors in data set F and y_k is the corresponding labels. $K(\mathbf{x}_k, \mathbf{x}) = \phi(\mathbf{x}_k) \cdot \phi(\mathbf{x})$ is the kernel function and "·" is the inner product in feature space $\mathbf{X}'$. Learning process on the given data set F is to find the support vectors and determining the parameters $\boldsymbol{\alpha}$ and b. There requires no need for explicit calculation for the new feature $\phi(\mathbf{x})$, since a kernel function is employed in training and forecasting. The most widely used kernel is the radial basis function (RBF) with a width of σ as shown in Eq. (19):

$$\mathrm{K}(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|/2\sigma^2\right) \tag{19}$$

A grid-search and cross-validation process is employed to get the optimal model, and the parameter grid consists of penalty $C = \{0.1, 1, 2, 5, 10\}$ and kernel parameter gamma = $\{0.01, 0.1, 0.2, 0.5, 0.8\}$.

4.2.2 Recurrent neural networks

Recurrent neural networks (RNN) [51] are wildly employed in stock market predictions [11]. RNN is a type of neural network where connections between the calculating units form a directed circle. Same task is performed for every element in a sequence and the output depends only on the previous calculation.

In our RNN model, the input value of the tth day $x_t = (x_{t,1}, \cdots, x_{t,m})$ is iterated over the following equations,

$$h_t = \tanh(Ux_t + Wh_{t-1} + b) \tag{20}$$

$$o_t = \tanh(Vh_t + c) \tag{21}$$

where h_t is the hidden state which is calculated based on the previous hidden state h_{t-1} and the input x_t at the current time step. o_t is the predicted output value which refers to the closing price in this study. U, W and V are the input-to-hidden, hidden-to-hidden and hidden-to-output parameters respectively.

A grid-search and cross-validation process is also employed, and the parameter grid consists of dropout rate $d = \{0.1, 0.35, 0.5\}$ and batch size $b = \{10, 100, 200, 400\}$.

4.3 ALSTM setup

In the experiment, three advanced methods for ALSTM training are applied. First, we make use of Root mean square prop (RMSprop) [76], a mini-batch version of rprop, as optimizer since it is "usually a good choice for recurrent neural networks" [13]. The initial learning rate is set to 0.001 as recommended in the default settings. A higher initial learning rate can reduce the time required for model optimization at an early stage, but it will bring more difficulties in achieving optimality and the model performance is restricted. Accordingly, a lower initial learning rate leads to more training epochs but a better optimum. Therefore, a decay mechanism is adopted to reduce the learning rate to half of itself when the loss rate does not decrease in 5 consecutive iterations to obtain the optimal model.

Second, early-stop mechanisms are employed to stop the training process automatically and to further reduce the overfitting risk. Max training epochs is set to 1000. When the training loss cannot be optimized after several rounds of iterations, the subsequent training becomes no longer necessary. When the loss does not decrease in 20 consecutive epochs, the model with the least loss rate is saved and is supposed to own the best generalization ability.

b=15ptThird, grid-search and cross-validation process are also employed, and the grid consists of two hyper parameters, each parameter contains several candidate hyper parameter values:

- Dropout rate = {0.1, 0.35, 0.5}: The dropout rate of dropout layers.
- Batch size = {10, 100, 200, 400}: The number of samples selected for training at a time.

4.4 BERT setup

In this study, the pre-trained language model BERT-base, which contains 12 Transformer blocks, 12 self-attention heads and the hidden size of 768, is employed as the encoder. The input sequence is output as a sequence representation through BERT. A special token [CLS] which contains the classification embedding is always placed at the sentence beginning. In sentiment classification tasks, the whole sequence is represented by the final hidden state h of the first token. A softmax layer is employed to predict the output probability of label c:

$$p(c|h) = \text{softmax}(\text{Wh}) \tag{22}$$

where W means the task-specific parameter matrix. Parameters are fine-tuned by maximizing the probability of the correct label.

The parameters are randomly initialized, most of them remaining unchanged as in pre-training, except for the batch size and learning rate. To avoid overfitting, the dropout rate was always kept at 0.1 to the dense layer. For model training, we used the Adam [36] optimizer and the number of epochs is set to 3. Max sequence length is set to 32 in the training process. The optimal parameter values are usually task-specific, and therefore we employ the grid-search process to find the optimal parameters. The following possible candidate values are found to work well across all tasks:

Batch size = {16, 32}
Learning rata = {5e-5, 3e-5, 2e-5}

In this study, 100,000 GuBa posts are selected for fine tuning of the model, 90% of them for fine-tuning to find the best parameter set and the rest of them for evaluation.

4.5 Experiment setup

We conduct a large amount of comparative experiments on 28 selected stocks based on the ALSTM networks to evaluate the predicting performance, SVR and RNN are used as baseline models. The time span of the dataset is within the range from 18 November 2016 to 18

November 2019. The data form 18 November 2016 to 1 June 2019 (about 85% of the data) is used for training to conduct cross-validation to select optimal hyper parameters, and the data from 1 June 2019 to 18 November 2019 (last 15% of the data) is used for testing to evaluate the out-of-sample performances.

Following Ratto, et al. [65], we also adopt the "walk forward testing" method in cross-validation process. To maximally utilize the available data, an increasing-window was designed to run a 5-fold time split cross-validation. The first k folds of the time series data is used for training and the k + 1th fold for validation. The cross-validation process is shown in Fig. 3.

For analysing the performance of each model, RMSE, MAE and accuracy are used as evaluation metrics. The RMSE and MAE, which provide an excellent error metric, are widely used in model valuation. The accuracy is employed to evaluate the consistency of the price movement in directions between the real and predicted values.

Given a set of time series observation values and the corresponding predictions, RMSE and MAE are defined as follows,

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(r_{t+1}-\widehat{r}_{t+1}\right)^2} \tag{23}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|r_{t+1}-\widehat{r}_{t+1}\right| \tag{24}$$

where r_{t+1} and $\widehat{r}_{t+1}$ denotes the actual closing price and the predicted one at time t + 1 respectively. RMSE is used as the evaluation metric to find the best parameter set for each model. Each transaction date was marked a label (up, down) through comparing the closing

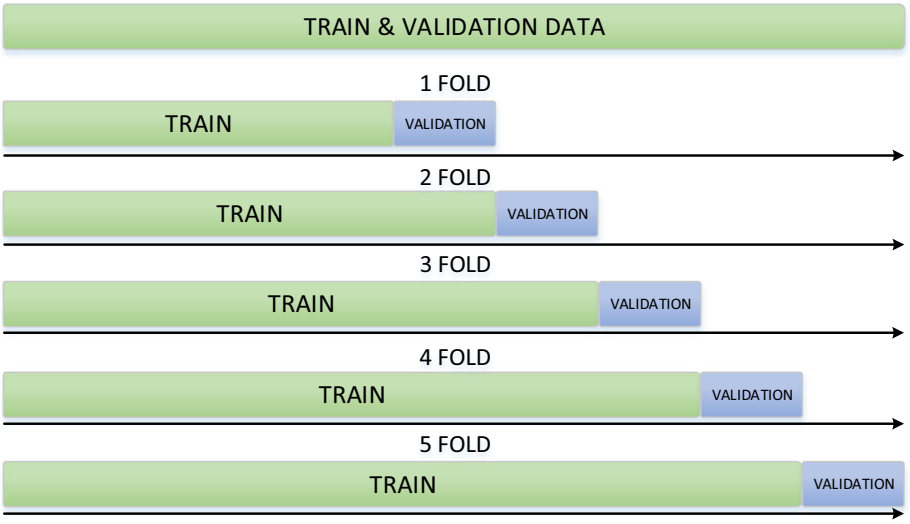

Fig. 3 Cross-validation process of 5-fold time series splitting method

price of two consecutive days. Accuracy is calculated by comparing the real trend with the predicted trend, which is defined as follows,

$$accuracy = \frac{tu + td}{tu + td + fu + fd} \tag{25}$$

where:

tu: the number of samples correctly classified as uptrend.
td: the number of samples correctly classified as downtrend.
fu: the number of samples incorrectly classified as uptrend.
fd: the number of samples incorrectly classified as downtrend.

The purpose of this study is to employ stock prices, technical indicators and GuBa sentiments of day t to predict the closing price of day $t + 1$. For the RNN and the ALSTM models, we also combine the past N days' features for training where N represents 3, 5, 7, 10, 15 and 30. This series of comparative experiments were designed to learn the sequential information and discover the best input window length for stock price prediction. We use the form of matrix and space vector to represent the input data, which is defined as:

$$X = \begin{pmatrix} X_1 = (X_{1,1}, X_{1,2}, \ldots\ldots\ X_{1,n}) \\ \vdots \\ X_N = (X_{N,1}, X_{N,2}, \ldots\ldots\ X_{N,n}) \end{pmatrix} \tag{26}$$

The meaning of this matrix is that there are N days' stock data for each training input, and each day consists n features. The timing information of the historical N trading days' sequence data are modelled, and is used for input as a vector. As shown in Fig. 4, a sliding time window is applied to get the features and labels. This window moves forward by one step until the end of the time series. Finally, by learning the historical data of the previous N days, the closing price of the $N + 1$ day is predicted.

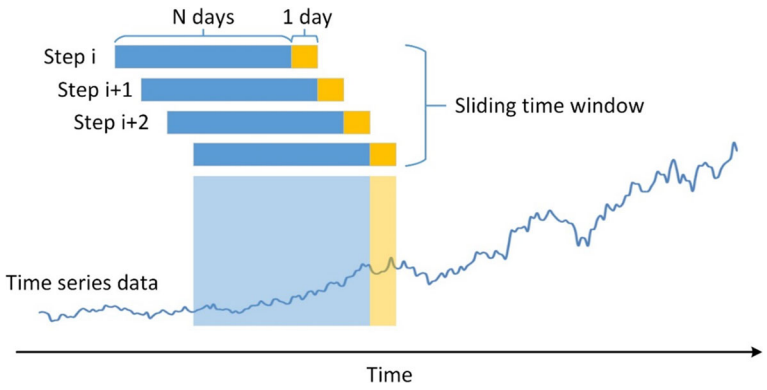

Fig. 4 Structure of one-step-ahead sliding time windows

Table 7 Accuracy of sentiment classification of GuBa posts on test set

model	Acc	Acc_{pos}	Acc_{neg}	Acc_{neu}
Sentiment lexicon	0.704	0.687	0.637	0.787
Fine-tuned BERT	0.859	0.809	0.792	0.975

4.6 Experiment results

The comparison of sentiment classification accuracy between sentiment lexicon and fine-tuned BERT is shown in Table 7 where we calculate the overall accuracy and the accuracy in predicting positive, negative and neutral posts. It can be observed that our BERT based sentiment classification method achieved better performance in predicting all three sentiment tendency on test set. The accuracy in sentiment classification reaches 85.9% on test set, 22.0% higher than sentiment lexicon method.

Table 8 tabulates the cross-validation results for based on different sentiment classification methods. The smallest RMSE score is marked in bold font. The result shows that the ALSTM model using different sentiment classification methods has the best performance in most cases. Among all 28 stocks, RNN obtains the best fitting results for only 1 stock using fine-tuned

Table 8 RMSE of models on validation sets

Stock	Fine-tuned BERT			Sentiment lexicon		
	SVR	RNN	ALSTM	SVR	RNN	ALSTM
DEEJ	0.0392	0.0246	**0.0218**	0.0678	**0.0198**	0.0243
YUNNAN BAIYAO	0.0441	0.0379	**0.0378**	0.1043	0.0507	**0.0399**
CHANGCHUN HIGH-TECH	0.0421	0.0161	**0.0157**	0.0785	0.0574	**0.0170**
HUADONG MEDICINE	0.0518	0.0349	**0.0348**	0.1009	**0.0103**	0.0345
NHU	0.0552	0.0404	**0.0333**	0.1205	**0.0159**	0.0325
HUALAN BIO	0.0459	0.0359	**0.0334**	0.1287	0.0733	**0.0356**
HEALTH 100	0.0458	0.0260	**0.0219**	0.0870	0.0290	**0.0273**
SHANGHAI RAAS	0.0510	0.0314	**0.0167**	0.1545	0.0165	**0.0160**
SALUBRIS	0.0442	0.0272	**0.0268**	0.0847	0.0260	**0.0251**
BICON	0.0409	0.0363	**0.0289**	0.0912	**0.0161**	0.0280
INDUSTRY GROUP	0.0496	0.0317	**0.0305**	0.0789	0.0330	**0.0302**
KANGHONG PHARMACEUTICAL	0.0423	0.0350	**0.0333**	0.1257	**0.0206**	0.0340
LEPU MEDICAL	0.0509	0.0318	**0.0246**	0.0897	0.0300	**0.0289**
AIER EYE HOSPITAL	0.0476	0.0374	**0.0355**	0.1330	0.0408	**0.0379**
ZFSW	0.0627	0.0645	**0.0438**	0.1841	0.0538	**0.0513**
WALVAX	0.0466	**0.0239**	0.0251	0.1163	0.0444	**0.0332**
Tongrentang Chinese Medicine	0.0467	0.0463	**0.0179**	0.1274	**0.0188**	0.0367
FOSUN PHARMA	0.0424	0.0335	**0.0198**	0.0808	0.0222	**0.0209**
HENGRUI MEDICINE	0.0440	0.0316	**0.0288**	0.1163	0.0297	**0.0276**
BYS	0.0473	0.0413	**0.0296**	0.1004	0.0670	**0.0364**
PIEN TZE HUANG	0.0468	0.0256	**0.0229**	0.0919	0.0246	**0.0224**
TASLY HOLDING GROUP	0.0446	0.0272	**0.0222**	0.1466	**0.0147**	0.0273
JUMPCAN	0.0429	0.0359	**0.0319**	0.0636	**0.0204**	0.0400
TONGHUA DONGBAO PHARMACEUTICAL	0.0432	0.0290	**0.0189**	0.0863	0.0218	**0.0210**
JOINTOWN PHARMACEUTICAL GROUP	0.0376	0.0358	**0.0208**	0.0885	0.0252	**0.0237**
SHANGHAI PHARMA	0.0397	0.0316	**0.0189**	0.1198	0.0253	**0.0237**
WuXi AppTec	0.0452	0.0362	**0.0303**	0.1145	0.0387	**0.0327**
BUCHANG PHARMA	0.0441	0.0256	**0.0105**	0.1470	**0.0103**	0.0129

The smallest RMSE score is marked in bold font

BERT sentiment classification and 9 stocks using sentiment lexicon. SVR obtains the worst performance among all three models.

The results of the test set are shown in Table 9 for fine-tuned BERT and Table 10 for sentiment lexicon respectively. The smallest MAE and RMSE scores for each stock are marked in bold font, and the highest accuracy score is underlined.

Based on fine-tuned BERT (Table 9), the ALSTM model outperforms the baselines on 21 stocks under the MAE, 20 stocks under the RMSE and 23 stocks under the accuracy. The RNN has the best performance on 7 stocks under the MAE, 8 stocks under the RMSE and 4 stocks under the accuracy. The SVR has the best performance on 1 stocks under the accuracy. It is clear that the ALSTM model outperforms the RNN and the SVR (64:15:1). Based on sentiment lexicon (Table 10), the ALSTM outperforms the baselines on 20 stocks under the MAE, 19 stocks under the RMSE and 21 stocks under the accuracy. The RNN has the best performance on 8 stocks under the MAE, 9 stocks under the RMSE and 1 stocks under the accuracy. The SVR has the best performance on 6 stocks under the accuracy. In summary, the ALSTM outperforms the RNN and the SVR (60:18:6). By comparing the results based on different sentiment classification methods, it is clear that the ALSTM obtains the best performance, the RNN obtains the second best results, while SVR has the worst results.

The average accuracy of 28 stocks using different input window length is calculated in Table 11 for easy comparison. It can be concluded that when setting the input window length to 5-day, the ALSTM model using fine-tuned BERT sentiment classification method achieves the highest accuracy. The average accuracy of 28 stocks reaches 61.24%.

4.7 Discussions on experimental results

4.7.1 The effectiveness of integrating sentiments

We use Δ_s to represent changes in accuracy between the results with and without sentiment feature to assess the effectiveness of integrating sentiments into stock predictions. Δ_s is calculated by,

$$\Delta_s = \frac{Acc_{all} - Acc_p}{Acc_p} \tag{27}$$

where Acc_{all} represents the accuracy of the ALSTM model using both price and sentiments and Acc_p the accuracy using price data only. The improvements between two sentiment classification methods are shown in Fig. 5. It is clear that combining price data and sentiments for stock predicting outperforms using exclusively price data for most stocks. Through further comparison, most of the improvements brought by sentiment lexicon are under 15%. The fine-tuned BERT method significantly improves the prediction accuracy to a greater extent, with some of the improvements exceeding 15%.

4.7.2 The effectiveness of using multiple information sources

To verify whether multiple information sources can improve predicting performance or the sentiment information is enough for prediction and other additional statistical measures are

Table 9 MAE, RMSE and Accuracy of BERT based models on test sets

Stocks	SVR			RNN			ALSTM		
	MAE	RMSE	Accuracy	MAE	RMSE	Accuracy	MAE	RMSE	Accuracy
DEEJ	0.0507	0.0668	0.5688	**0.0120**	**0.0169**	0.4771	0.0189	0.0311	0.5872
YUNNAN BAIYAO	0.0655	0.0854	0.4948	0.0404	0.0525	0.5052	**0.0312**	**0.0385**	0.5464
CHANGCHUN HIGH-TECH	0.0581	0.0581	0.5556	0.0314	0.0428	0.5093	**0.0195**	**0.0263**	0.5741
HUADONG MEDICINE	0.0722	0.1042	0.5229	**0.0108**	**0.0136**	0.5138	0.0237	0.0353	0.5321
NHU	0.1135	0.1335	0.5185	0.0194	**0.0236**	0.5185	**0.0189**	0.0253	0.5741
HUALAN BIO	0.0899	0.1317	0.5596	0.0376	0.0743	0.4679	**0.0342**	**0.0741**	0.5963
HEALTH 100	0.0635	0.0847	0.4815	**0.0136**	0.0217	0.5093	0.0147	**0.0208**	0.5278
SHANGHAI RAAS	0.1209	0.1521	0.5556	**0.0157**	0.0263	0.4921	0.0237	**0.0241**	0.5238
SALUBRIS	0.0618	0.0850	0.5046	0.0143	0.0205	0.4771	**0.0120**	**0.0202**	0.5229
BICON	0.0759	0.1050	0.5341	0.0117	**0.0172**	0.5682	**0.0112**	0.0295	0.5568
INDUSTRY GROUP	0.0546	0.0706	0.4312	0.0213	**0.0288**	0.5138	**0.0208**	0.0332	0.5229
KANGHONG PHARMACEUTICAL	0.0824	0.1133	0.5596	**0.0134**	**0.0208**	0.5688	0.0355	0.0378	0.6055
LEPU MEDICAL	0.0882	0.1094	0.5185	0.0208	0.0272	0.5278	**0.0199**	**0.0266**	0.5648
AIER EYE HOSPITAL	0.0705	0.1212	0.5514	0.0196	0.0389	0.5607	**0.0191**	**0.0395**	0.5888
ZFSW	0.1387	0.1794	0.5000	0.0301	0.0458	0.5000	**0.0295**	**0.0446**	0.5313
WALVAX	0.0630	0.0835	0.5046	0.0379	0.0467	0.5229	**0.0285**	**0.0309**	0.5963
Tongrentang Chinese Medicine	0.0982	0.1334	0.5424	0.0169	0.0216	0.5085	**0.0137**	**0.0178**	0.5932
FOSUN PHARMA	0.0625	0.0798	0.5321	0.0155	0.0234	0.5138	**0.0139**	**0.0201**	0.5413
HENGRUI MEDICINE	0.0720	0.1013	0.4862	0.0235	0.0334	0.5505	**0.0217**	**0.0308**	0.6514
BYS	0.0783	0.0958	0.5588	0.0265	0.0343	0.5098	**0.0254**	**0.0328**	0.5872
PIEN TZE HUANG	0.0711	0.0922	0.5138	0.0209	0.0262	0.5046	**0.0178**	**0.0234**	0.5780
TASLY HOLDING GROUP	0.1028	0.1348	0.5046	**0.0186**	**0.0224**	0.5138	0.0235	0.0284	0.5229
JUMPCAN	0.0339	0.0591	0.4954	**0.0153**	**0.0203**	0.4954	0.0289	0.0325	0.5229
TONGHUA DONGBAO PHARMACEUTICAL	0.0602	0.0819	0.5421	0.0165	0.0236	0.5701	**0.0148**	**0.0208**	0.6355
JOINTOWN PHARMACEUTICAL GROUP	0.0658	0.0886	0.5278	0.0183	0.0245	0.5556	**0.0165**	**0.0228**	0.4907
SHANGHAI PHARMA	0.0943	0.1205	0.5596	0.0162	0.0217	0.5780	**0.0148**	**0.0198**	0.6055
WuXi AppTec	0.0856	0.1071	0.4821	0.0337	0.0408	0.5536	**0.0188**	**0.0265**	0.5357
BUCHANG PHARMA	0.1213	0.1514	0.4679	0.0091	0.0116	0.5872	**0.0055**	**0.0109**	0.5046

The smallest MAE and RMSE scores are marked in bold font, and the highest accuracy score is underlined

Table 10 MAE, RMSE and Accuracy of sentiment lexicon based models on test sets

Stocks	SVR			RNN			ALSTM		
	MAE	RMSE	Accuracy	MAE	RMSE	Accuracy	MAE	RMSE	Accuracy
DEEJ	0.0518	0.0678	0.5780	**0.0147**	**0.0198**	0.4954	0.0218	0.0256	0.5413
YUNNAN BAIYAO	0.0818	0.1043	0.4742	0.0438	0.0507	0.4742	**0.0293**	**0.0371**	0.4948
CHANGCHUN HIGH-TECH	0.0624	0.0785	0.5093	0.0428	0.0574	0.4630	**0.0185**	**0.0214**	0.5185
HUADONG MEDICINE	0.0696	0.1009	0.5046	**0.0064**	**0.0103**	0.5138	0.0098	0.0127	0.5229
NHU	0.1024	0.1205	0.5278	0.0130	0.0159	0.5000	**0.0128**	**0.0150**	0.5463
HUALAN BIO	0.0861	0.1287	0.5321	0.0323	0.0733	0.4771	**0.0305**	**0.0446**	0.5596
HEALTH 100	0.0636	0.0870	0.4815	0.0234	0.0290	0.5093	**0.0218**	**0.0287**	0.5370
SHANGHAI RAAS	0.1226	0.1545	0.5397	0.0187	0.0198	0.5238	**0.0115**	**0.0163**	0.5556
SALUBRIS	0.0619	0.0847	0.4954	0.0221	0.0260	0.4771	**0.0198**	**0.0256**	0.5138
BICON	0.0618	0.0912	0.6023	**0.0110**	**0.0161**	0.5227	0.0246	0.0287	0.5000
INDUSTRY GROUP	0.0598	0.0789	0.4404	0.0246	0.0330	0.4862	**0.0245**	**0.0322**	0.5138
KANGHONG PHARMACEUTICAL	0.0978	0.1257	0.5688	**0.0161**	**0.0206**	0.4771	0.0335	0.0403	0.5780
LEPU MEDICAL	0.0723	0.0897	0.5000	0.0237	0.0267	0.5370	**0.0185**	**0.0244**	0.6019
AIER EYE HOSPITAL	0.0747	0.1330	0.5421	0.0205	0.0408	0.5421	**0.0202**	**0.0403**	0.5888
ZFSW	0.1402	0.1841	0.5000	0.0438	0.0538	0.5000	**0.0326**	**0.0421**	0.5313
WALVAX	0.0892	0.1163	0.5872	0.0390	0.0444	0.5046	**0.0362**	**0.0388**	0.5321
Tongrentang Chinese Medicine	0.0944	0.1274	0.5763	0.0148	0.0188	0.5085	**0.0146**	**0.0185**	0.6102
FOSUN PHARMA	0.0628	0.0808	0.5046	0.0173	0.0222	0.5138	**0.0160**	**0.0206**	0.5505
HENGRUI MEDICINE	0.0766	0.1163	0.4587	**0.0213**	**0.0297**	0.4771	0.0233	0.0324	0.5780
BYS	0.0807	0.1004	0.5098	0.0594	0.0670	0.5098	**0.0318**	**0.0382**	0.5196
PIEN TZE HUANG	0.0689	0.0919	0.4954	0.0193	0.0246	0.4771	**0.0170**	**0.0221**	0.5229
TASLY HOLDING GROUP	0.1153	0.1466	0.4954	**0.0114**	**0.0147**	0.4954	0.0341	0.0404	0.5321
JUMPCAN	0.0480	0.0636	0.5505	**0.0151**	**0.0204**	0.5413	0.0749	0.0951	0.5138
TONGHUA DONGBAO PHARMACEUTICAL	0.0640	0.0863	0.5234	0.0154	0.0218	0.5140	**0.0145**	**0.0215**	0.5421
JOINTOWN PHARMACEUTICAL GROUP	0.0659	0.0885	0.5370	0.0199	0.0252	0.5185	**0.0185**	**0.0240**	0.5000
SHANGHAI PHARMA	0.0935	0.1198	0.5872	0.0206	**0.0253**	0.4771	**0.0205**	0.0258	0.4954
WuXi AppTec	0.0901	0.1145	0.5000	0.0322	0.0387	0.5439	**0.0197**	**0.0267**	0.5179
BUCHANG PHARMA	0.1179	0.1470	0.4404	**0.0067**	**0.0103**	0.4679	0.0202	0.0244	0.4771

The smallest MAE and RMSE scores are marked in bold font, and the highest accuracy score is underlined

Table 11 Average accuracy of different models using different input window length

Model	Input window length						
	N=1	N=3	N=5	N=7	N=10	N=15	N=30
RNN	0.5062	0.5012	0.5223	0.5208	0.5018	0.4956	0.5106
RNN+lexicon	0.5017	0.5234	0.5428	0.5186	0.5062	0.5145	0.5118
RNN+BERT	0.5241	0.5315	0.5486	0.5360	0.5108	0.5338	0.5266
ALSTM	0.5084	0.5216	0.5388	0.5410	0.5122	0.5081	0.5122
ALSTM+lexicon	0.5355	0.5288	0.5520	0.5326	0.5102	0.5208	0.5189
ALSTM+BERT	0.5614	0.5772	0.6124	0.5810	0.5436	0.5425	0.5365

unnecessary. Thus, we use Δ_p to evaluate the difference in accuracy between the ALSTM models with and without price data. Δ_p is calculated by,

$$\Delta_p = \frac{Acc_{all} - Acc_s}{Acc_s} \tag{28}$$

where Acc_s represents the predicting accuracy based on sentiment feature only. The results of Δ_p are shown in Table 12. It is clear that using multiple information sources outperforms using sentiment source only in all cases.

4.7.3 The effectiveness of using long input window length

To investigate whether the increase of the input window length can help the models to extract more time series information and improve the predicting performance, we employ Δ_T to represent the changes of the accuracy between N time steps and 1 time step where N represents 3, 5, 7, 10, 15 and 30. Δ_T is calculated by,

$$\Delta_T = \frac{Acc_N - Acc_1}{Acc_1} \tag{29}$$

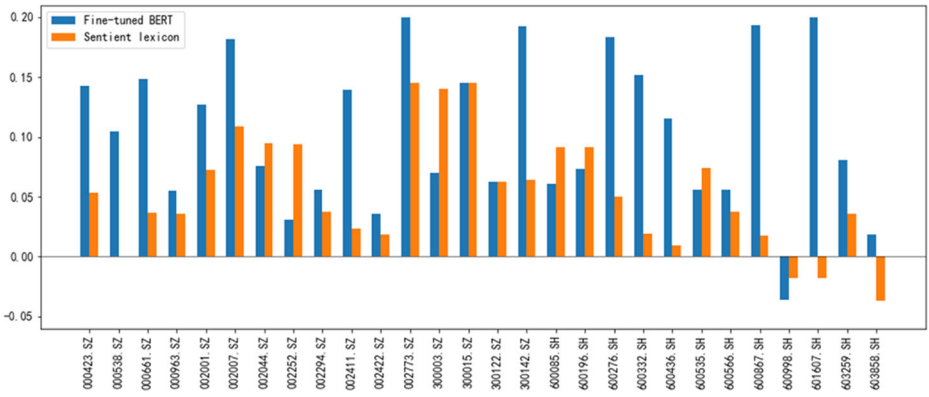

Fig. 5 The Δ_s of each stock, where x axis represents stock codes

Table 12 The prediction of the ALSTM model using different data sources

Stocks	Fine-tuned BERT			Sentiment lexicon		
	Acc_s	Acc_{all}	Δ_p	Acc_s	Acc_{all}	Δ_p
DEEJ	0.5046	0.5872	0.1637	0.4954	0.5413	0.0927
YUNNAN BAIYAO	0.4639	0.5464	0.1778	0.4433	0.4948	0.1162
CHANGCHUN HIGH-TECH	0.5093	0.5741	0.1272	0.4907	0.5185	0.0567
HUADONG MEDICINE	0.4954	0.5321	0.0741	0.5046	0.5229	0.0363
NHU	0.4907	0.5741	0.1700	0.4722	0.5463	0.1569
HUALAN BIO	0.5138	0.5963	0.1606	0.5046	0.5596	0.1090
HEALTH 100	0.5000	0.5278	0.0556	0.4722	0.5370	0.1372
SHANGHAI RAAS	0.4762	0.5238	0.1000	0.4444	0.5556	0.2502
SALUBRIS	0.4954	0.5229	0.0555	0.4587	0.5138	0.1201
BICON	0.4318	0.5568	0.2895	0.4091	0.5000	0.2222
INDUSTRY GROUP	0.4862	0.5229	0.0755	0.4587	0.5138	0.1201
KANGHONG PHARMACEUTICAL	0.4404	0.6055	0.3749	0.4220	0.5780	0.3697
LEPU MEDICAL	0.4630	0.5648	0.2199	0.4907	0.6019	0.2266
AIER EYE HOSPITAL	0.5047	0.5888	0.1666	0.4953	0.5888	0.1888
ZFSW	0.5000	0.5313	0.0626	0.4375	0.5313	0.2144
WALVAX	0.5321	0.5963	0.1207	0.4954	0.5321	0.0741
Tongrentang Chinese Medicine	0.4746	0.5932	0.2499	0.4746	0.6102	0.2857
FOSUN PHARMA	0.5046	0.5413	0.0727	0.4679	0.5505	0.1765
HENGRUI MEDICINE	0.4679	0.6514	0.3922	0.4587	0.5780	0.2601
BYS	0.4902	0.5872	0.1979	0.4804	0.5196	0.0816
PIEN TZE HUANG	0.4954	0.5780	0.1667	0.4954	0.5229	0.0555
TASLY HOLDING GROUP	0.4679	0.5229	0.1175	0.4862	0.5321	0.0944
JUMPCAN	0.4404	0.5229	0.1873	0.4312	0.5138	0.1916
TONGHUA DONGBAO PHARMACEUTICAL	0.5421	0.6355	0.1723	0.5047	0.5421	0.0741
JOINTOWN PHARMACEUTICAL GROUP	0.4537	0.4907	0.0816	0.4907	0.5000	0.0190
SHANGHAI PHARMA	0.5229	0.6055	0.1580	0.4587	0.4954	0.0800
WuXi AppTec	0.5000	0.5357	0.0714	0.4643	0.5179	0.1154
BUCHANG PHARMA	0.5000	0.5046	0.0092	0.4273	0.4771	0.1165

where Acc_N is the average accuracy of 28 stocks when the input window length is set to N and Acc_1 is the average accuracy when $N = 1$. The changes are shown in Table 13. It can be observed that using the 5-day time series data as model input can improve the performance for all proposed models in average accuracy.

Table 13 The Δ_T of each model based on different input window length

Model	Input window length					
	N=3	N=5	N=7	N=10	N=15	N=30
RNN	−0.0099	0.0318	0.0288	−0.0087	−0.0209	0.0087
RNN+Lexicon	0.0433	0.0819	0.0337	0.0090	0.0255	0.0201
RNN+BERT	0.0141	0.0467	0.0227	−0.0254	0.0185	0.0048
ALSTM	0.0260	0.0598	0.0641	0.0075	−0.0006	0.0075
ALSTM + Lexicon	−0.0125	0.0308	−0.0054	−0.0472	−0.0275	−0.0310
ALSTM + BERT	0.0281	0.0908	0.0349	−0.0317	−0.0337	−0.0444

5 Conclusions and future work

Stock price prediction is an important aspect of formulating a low-risk and high-return investment. This study focuses on an increasingly significant aspect of financial market research, namely: how to integrate investor sentiments from social media, and make model more qualified to learn time series information. To address the problem, we take the GuBa dataset of 28 stocks from November 18, 2016 to November 18, 2019 for efficient stock price movement prediction using SVR, RNN and ALSTM models. In this work, we propose a fine-tuned BERT sentiment classification model for sentiment analysis and a sentiment lexicon based on NTUSD for comparison. MAE, RMSE and accuracy are employed to evaluate the predictive accuracy. Furthermore, we evaluate the improvements bring by using different input window length. Results show that,

1. Based on multiple information sources, the ALSTM model performs better than the SVR and the RNN under the MAE, RMSE and accuracy.
2. Based on ALSTM, using multiple information sources improves the prediction accuracy than using either stock price data or sentiments.
3. The fine-tuned BERT model achieves higher accuracy in sentiment classification task, and the exploitation of the sentiment feature computed by the fine-tuned BERT model also led to better predicting performance.
4. Combining the 5-day features as a long time series sequential input for models to learn achieves the best predicting accuracy.

Furthermore, there are several future avenues available for this study. Sentiments from social media are the only sentiment resource considered in this study. However, the news data is also widely used in stock price predictions, as it is an important information source about the situation of the country. Moreover, only the historical prices, technical indicators and social media sentiments are employed in this study. Considering the complex and volatile stock market environment, we can further design another prediction model to extract information from other useful sources to make more comprehensive prediction. For example, the company's financial conditions, which can be concluded from the company's financial statements and balance sheet. Finally, a more advanced hyper-parameters selection scheme can also be employed in future experiments.

Acknowledgements This paper was supported by the National Natural Science Foundation of China (project numbers are 72274096, 72174087, 71774084 and 71874082), the National Social Science Fund of China (project number is 17ZDA291), program for Jiangsu Excellent Scientific and Technological Innovation Team (project number is [2020]10).

Authors contributions **Zhongtian Ji**: Conceptualization, Methodology, Investigation, Writing - original draft. **Peng Wu**: Project administration, Supervision, Writing - review & editing, Funding acquisition. **Chen Ling**: Formal analysis, Writing - review & editing, Data curation. **Peng Zhu**: Writing - review & editing.

Data availability The datasets analysed during the current study are not publicly available due to data privacy policy but are available from the corresponding author on reasonable request.

Declarations

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

References

1. Anjaria M, Guddeti RMR (2014) A novel sentiment analysis of social networks using supervised learning. Soc Netw Anal Min 4(1):181. https://doi.org/10.1007/s13278-014-0181-9
2. Antweiler W, Frank M (2004) Is all that talk just noise? The information content of internet stock message boards. J Financ 59:1259–1294. https://doi.org/10.2139/ssrn.282320
3. Baek Y, Kim HY (2018) ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module. Expert Syst Appl 113:457–480. https://doi.org/10.1016/j.eswa.2018.07.019
4. Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. J Financ 61(4):1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x
5. Ballings M, Van den Poel D, Hespeels N, Gryp R (2015) Evaluating multiple classifiers for stock price direction prediction. Expert Syst Appl 42(20):7046–7056. https://doi.org/10.1016/j.eswa.2015.05.013
6. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8. https://doi.org/10.1016/j.jocs.2010.12.007
7. Cambria E, Fu J, Bisio F, Poria S (2015) AffectiveSpace 2: enabling affective intuition for concept-level sentiment analysis. Proc AAAI 29:508–514
8. Cavalcante RC, Brasileiro RC, Souza VLF, Nobrega JP, Oliveira ALI (2016) Computational intelligence and financial markets: A survey and future directions. Expert Syst Appl 55:194–211. https://doi.org/10.1016/j.eswa.2016.02.006
9. Chandra R, Chand S (2016) Evaluation of co-evolutionary neural network architectures for time series prediction with mobile application in finance. Appl Soft Comput 49:462–473. https://doi.org/10.1016/j.asoc.2016.08.029
10. Checkley MS, Higón DA, Alles H (2017) The hasty wisdom of the mob: how market sentiment predicts stock market behavior. Expert Syst Appl 77:256–263. https://doi.org/10.1016/j.eswa.2017.01.029
11. Chen W, Yeo CK, Lau CT, Lee BS (2018) Leveraging social media news to predict stock index movement using RNN-boost. Data Knowl Eng 118:14–24. https://doi.org/10.1016/j.datak.2018.08.003
12. Chen M-Y, Liao C-H, Hsieh R-P (2019) Modeling public mood and emotion: stock market trend prediction with anticipatory computing approach. Comput Hum Behav 101:402–408. https://doi.org/10.1016/j.chb.2019.03.021
13. Chollet F (2016) Keras. https://github.com/keras-team/keras. Accessed 13 Feb 2023
14. Cortes C, Vapnik V (1995) Support vector network. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
15. Oliveira FA, Zárate LE, de Azebedo Reis M, Nobre CN (2011) The use of artificial neural networks in the analysis and prediction of stock prices. 2011 IEEE international conference on systems, man, and cybernetics, pp 2151–215., https://doi.org/10.1109/ICSMC.2011.6083990
16. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805
17. Eapen J, Bein D, Verma A (2019) Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction. *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, 0264-0270, https://doi.org/10.1109/CCWC.2019.8666592
18. Fama EF (1991) Efficient capital markets: II. J Financ 46(5):1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x
19. Faraji-Rad A, Pham M (2016) Uncertainty increases the reliance on affect in decisions. SSRN Electron J 44. https://doi.org/10.2139/ssrn.2715333
20. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270(2):654–669. https://doi.org/10.1016/j.ejor.2017.11.054
21. Gerlein EA, McGinnity M, Belatreche A, Coleman S (2016) Evaluating machine learning classification for financial trading: an empirical approach. Expert Syst Appl 54:193–207. https://doi.org/10.1016/j.eswa.2016.01.018
22. Giles C, Lawrence S (2001) Noisy time series prediction using recurrent neural networks and grammatical inference. Mach Learn 44:161–183. https://doi.org/10.1023/A:1010884214864
23. Gradojevic N, Lento C, Wright C (2007) Investment information content in Bollinger bands? Appl Financ Econ Lett 3:263–267. https://doi.org/10.1080/17446540701206576
24. Gunasekarage A, Power DM (2001) The profitability of moving average trading rules in south Asian stock markets. Emerg Mark Rev 2(1):17–33. https://doi.org/10.1016/S1566-0141(00)00017-0
25. Güreşen E, Kayakutlu G, Daim T (2011) Using artificial neural network models in stock market index prediction. Expert Syst Appl 38:10389–10397. https://doi.org/10.1016/j.eswa.2011.02.068
26. Harb JGD, Ebeling R, Becker K (2020) A framework to analyze the emotional reactions to mass violent events on twitter and influential factors. Inf Process Manag 57(6):102372. https://doi.org/10.1016/j.ipm.2020.102372

27. Henrique BM, Sobreiro VA, Kimura H (2018) Stock price prediction using support vector regression on daily and up to the minute prices. J Finance Data Sci 4(3):183–201. https://doi.org/10.1016/j.jfds.2018.04.003
28. Hiransha M, Gopalakrishnan EA, Menon VK, Soman KP (2018) NSE stock market prediction using deep-learning models. Procedia Comput Sci 132:1351–1362. https://doi.org/10.1016/j.procs.2018.05.050
29. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
30. Hung C (2017) Word of mouth quality classification based on contextual sentiment lexicons. Inf Process Manag 53(4):751–763. https://doi.org/10.1016/j.ipm.2017.02.007
31. Junqué de Fortuny E, De Smedt T, Martens D, Daelemans W (2014) Evaluating and understanding text-based stock price prediction models. Inf Process Manag 50(2):426–441. https://doi.org/10.1016/j.ipm.2013.12.002
32. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, pp 137–146. https://doi.org/10.1145/956750.956769
33. Kim S-H, Kim D (2014) Investor sentiment from internet message postings and the predictability of stock returns. J Econ Behav Organ 107:708–729. https://doi.org/10.1016/j.jebo.2014.04.015
34. Kim T, Kim H (2019) Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. PLoS One 14:e0212320. https://doi.org/10.1371/journal.pone.0212320
35. Kim HY, Won CH (2018) Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Syst Appl 103:25–37. https://doi.org/10.1016/j.eswa.2018.03.002
36. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. *International Conference on Learning Representations*
37. Klinker F (2011) Exponential moving average versus moving exponential average. Math Semesterber 58: 97–107. https://doi.org/10.1007/s00591-010-0080-8
38. Kumar S, Kumar K (2018) IRSC: integrated automated review mining system using virtual Machines in Cloud environment. 2018 Conference on Information and Communication Technology (CICT), 1–6, https://doi.org/10.1109/INFOCOMTECH.2018.8722387
39. Kumar K, Kurhekar M (2017) Sentimentalizer: Docker container utility over cloud. 2017 ninth international conference on advances in pattern recognition (ICAPR), 1–6, https://doi.org/10.1109/ICAPR.2017.8593104
40. Kumar K, Bamrara R, Gupta P, Singh N (2020) M2P2: Movie's trailer reviews based movie popularity prediction system. In: Soft Computing: Theories and Applications, pp 671–681. https://doi.org/10.1007/978-981-15-0751-9_62
41. Kumar A, Purohit K, Kumar K (2021) Stock Price prediction using recurrent neural network and Long short-term memory. *Conference proceedings of ICDLAIR2019*, 153-160
42. Lee C, Soo V (2017) Predict stock Price with financial news based on recurrent convolutional neural networks. 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI), 160–165, https://doi.org/10.1109/TAAI.2017.27
43. Lee C, Swaminathan B (1999) Price momentum and trading volume. J Financ 55. https://doi.org/10.2139/ssrn.92589
44. Lee S, Ha T, Lee D, Kim JH (2018) Understanding the majority opinion formation process in online environments: an exploratory approach to Facebook. Inf Process Manag 54(6):1115–1128. https://doi.org/10.1016/j.ipm.2018.08.002
45. Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl-Based Syst 69:14–23. https://doi.org/10.1016/j.knosys.2014.04.022
46. Li B, Chan KCC, Ou C, Ruifeng S (2017) Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Inf Syst 69:81–92. https://doi.org/10.1016/j.is.2016.10.001
47. Li X, Wu P, Wang W (2020) Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Inf Process Manag 57(5):102212. https://doi.org/10.1016/j.ipm.2020.102212
48. Li Y, Bu H, Li J, Wu J (2020) The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. Int J Forecast 36(4):1541–1562. https://doi.org/10.1016/j.ijforecast.2020.05.001
49. Long J, Chen Z, He W, Wu T, Ren J (2020) An integrated framework of deep learning and knowledge graph for prediction of stock price trend: an application in Chinese stock exchange market. Appl Soft Comput 91:106205. https://doi.org/10.1016/j.asoc.2020.106205
50. Loughran TIM, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Financ 66(1):35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
51. Mandic D, Chambers J (2001) Recurrent neural networks for Prediction: Learning Algorithms,Architectures and Stability. https://doi.org/10.1002/047084535X
52. Maqsood H, Mehmood I, Maqsood M, Yasir M, Afzal S, Aadil F, Selim MM, Muhammad K (2020) A local and global event sentiment based efficient stock exchange forecasting using deep learning. Int J Inf Manag 50:432–451. https://doi.org/10.1016/j.ijinfomgt.2019.07.011

53. Mourelatos M, Alexakos C, Amorgianiotis T, Likothanassis S (2018) Financial indices modelling and trading utilizing deep learning techniques: the ATHENS SE FTSE/ASE large cap use case. 2018 Innovations in Intelligent Systems and Applications (INISTA), 1–7, https://doi.org/10.1109/INISTA.2018.8466286
54. Nelson DMQ, Pereira ACM, Oliveira RAD (2017) Stock market's price movement prediction with LSTM neural networks. *International Joint Conference on Neural Networks*.
55. Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42(24):9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052
56. Oh C, Sheng O (2011) Investigating predictive Power of stock Micro blog sentiment in forecasting future stock Price directional movement. Proceedings of the international conference on information systems, ICIS 2011, Shanghai, China, December 4–7, 2011
57. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2:1–135. https://doi.org/10.1561/1500000011
58. Pang X, Zhou Y, Wang P, Lin W, Chang V (2018) An innovative neural network approach for stock market prediction. J Supercomput 76:2098–2118. https://doi.org/10.1007/s11227-017-2228-y
59. Patel J, Shah S, Thakkar P, Kotecha K (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Syst Appl 42(1):259–268. https://doi.org/10.1016/j.eswa.2014.07.040
60. Peng Y, Jiang H (2015) Leverage financial news to predict stock Price movements using word Embeddings and deep neural networks
61. Picasso A, Merello S, Ma Y, Oneto L, Cambria E (2019) Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl 135:60–70. https://doi.org/10.1016/j.eswa.2019.06.014
62. Qian B, Rasheed K (2007) Stock market prediction with multiple classifiers. Appl Intell 26:25–33. https://doi.org/10.1007/s10489-006-0001-7
63. Qian Y, Li Z, Yuan H (2020) On exploring the impact of users' bullish-bearish tendencies in online community on the stock market. Inf Process Manag 57(5):102209. https://doi.org/10.1016/j.ipm.2020.102209
64. Qu H, Zhang Y (2016) A new kernel of support vector regression for forecasting high-frequency stock returns. Math Probl Eng 2016:1–9. https://doi.org/10.1155/2016/4907654
65. Ratto AP, Merello S, Oneto L, Ma Y, Cambria E (2018) Ensemble of Technical Analysis and Machine Learning for market trend prediction. *2018 IEEE symposium series on computational intelligence (SSCI)*
66. Rezaei H, Faaljou H, Mansourfar G (2020) Stock price prediction using deep learning and frequency decomposition. Expert Syst Appl 114332:114332. https://doi.org/10.1016/j.eswa.2020.114332
67. Schumaker RP, Chen H (2009) A quantitative stock prediction system based on financial news. Inf Process Manag 45(5):571–583. https://doi.org/10.1016/j.ipm.2009.05.001
68. Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM J Trans Inf Syst 27(2):12. https://doi.org/10.1145/1462198.1462204
69. Sehgal V, Song C (2007) SOPS: stock prediction using web sentiment. Seventh IEEE international conference on data mining workshops (ICDMW 2007), 21–26, https://doi.org/10.1109/ICDMW.2007.100
70. Sharma S, Kumar P, Kumar K (2017) LEXER: LEXicon based emotion AnalyzeR. Pattern recognition and machine intelligence, pp 373–379. https://doi.org/10.1007/978-3-319-69900-4_47
71. Sharpe M, Walczak S (2001) An empirical analysis of data requirements for financial forecasting with neural networks. J Manag Inf Syst 17
72. Shynkevich Y, McGinnity TM, Coleman S, Belatreche A, Li Y (2017) Forecasting Price movements using technical Indicators: Investigating the Impact of Varying Input Window Length. *Neurocomputing* 264:71–88. https://doi.org/10.1016/j.neucom.2016.11.095
73. Si J, Mukherjee A, Liu B, Li Q, Deng X (2013) Exploiting Topic based Twitter Sentiment for Stock Prediction. *ACL* 2013
74. Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune BERT for text classification? *China National Conference on Chinese Computational Linguistics*
75. Taylor M, Allen H (1992) The use of technical analysis in the foreign exchange market. J Int Money Financ 11:304–314. https://doi.org/10.1016/0261-5606(92)90048-3
76. Tieleman T, Hinton GE, Srivastava N, Swersky K (2012) Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. Neural Networks for Machine Learning, COURSERA
77. Tsibouris G, Zeidenberg M (1995) Testing the efficient market hypothesis with gradient descent algorithms. Neural Netw Capital Markets 8:127–136
78. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Proceedings of the 31st international conference on neural information processing systems, 6000–6010
79. Verma I, Dey L, Meisheri H (2017) Detecting, Quantifying and Accessing Impact of News Events on Indian Stock Indices. https://doi.org/10.1145/3106426.3106482

80. Vijayvergia A, Kumar K (2018) STAR: rating of reviewS by exploiting variation in emoTions using trAnsfer leaRning framework. 2018 Conference on Information and Communication Technology (CICT), 1–6, https://doi.org/10.1109/INFOCOMTECH.2018.8722356
81. Vijayvergia A, Kumar K (2021) Selective shallow models strength integration for emotion detection using GloVe and LSTM. Multimed Tools Appl 80(18):28349–28363. https://doi.org/10.1007/s11042-021-10997-8
82. Vu TIT, Chang S (2012) An experiment in integrating sentiment features for tech stock prediction in twitter. *Workshop on Information Extraction & Entity Analytics on Social Media Data*
83. 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. https://doi.org/10.1016/j.neucom.2018.02.095
84. Xing F, Cambria E, Welsch R (2018) Intelligent asset allocation via market sentiment views. IEEE Comput Intell Mag 13:25–34. https://doi.org/10.1109/MCI.2018.2866727
85. Yeh C-Y, Huang C-W, Lee S-J (2011) A multiple-kernel support vector regression approach for stock market price forecasting. Expert Syst Appl 38(3):2177–2186. https://doi.org/10.1016/j.eswa.2010.08.004
86. Yong BX, Abdul Rahim MR, Abdullah AS (2017) A stock market trading system using deep neural network. In: Modeling, Design and Simulation of Systems, pp 356–364. https://doi.org/10.1007/978-981-10-6463-0_31
87. Yu JH, Kang J, Park S (2019) Information availability and return volatility in the bitcoin market: analyzing differences of user opinion and interest. Inf Process Manag 56(3):721–732. https://doi.org/10.1016/j.ipm.2018.12.002
88. Zhang X, Tan Y (2018) Deep stock ranker: A LSTM neural network model for stock selection. In (pp. 614-623). https://doi.org/10.1007/978-3-319-93803-5_58
89. Zhang L, Aggarwal C, Qi G-J (2017) Stock Price prediction via discovering multi-frequency trading patterns. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2141–2149. https://doi.org/10.1145/3097983.3098117
90. Zhang X, Zhang Y, Wang S, Yao Y, Fang B, Yu PS (2018) Improving stock market prediction via heterogeneous information fusion. Knowl-Based Syst 143:236–247. https://doi.org/10.1016/j.knosys.2017.12.025
91. Zhang Y, Chu G, Shen D (2020) The role of investor attention in predicting stock prices: the long short-term memory networks perspective. Financ Res Lett 101484. https://doi.org/10.1016/j.frl.2020.101484
92. Zhang YA, Yan B, Aasma M (2020) A novel deep learning framework: prediction and analysis of financial time series using CEEMD and LSTM. Expert Syst Appl 159:113609. https://doi.org/10.1016/j.eswa.2020.113609
93. Zuo Y, Kita E (2012) Stock price forecast using Bayesian network. Expert Syst Appl 39(8):6729–6737. https://doi.org/10.1016/j.eswa.2011.12.035
94. Zuo Y, Kita E (2012) Up/down analysis of stock index by using Bayesian network. Eng Manag Res 1. https://doi.org/10.5539/emr.v1n2p46

Zhongtian Ji is a Ph.D. student of the School of Economics and Management, Nanjing University of Science and Technology. His research work mainly involves online users' sentiment analysis and deep learning.

Peng Wu is a Professor of the School of Intelligent Manufacturing, Nanjing University of Science and Technology. His research work mainly involves online users' behavior analysis and sentimental analysis.

Chen Ling is an Associate Professor of the School of Economics and Management, Nanjing University of Science and Technology. His research work mainly involves crowd simulation and sentimental analysis.

Peng Zhu is an Associate Professor of the School of Economics and Management, Nanjing University of Science and Technology. His research work mainly involves user behavior, blockchain, data mining and business intelligence.