2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

Stock Trend Prediction using Historical Data and Financial Online News

Yuemei Xu, Weihang Lin, Yiran Hu
School of Information Science and Technology,
Beijing Foreign Studies University, 100089, Beijing, China
Email: xuyuemei@bfsu.edu.cn, oscar.lam.bfsu@gmail.com, yiranbfsu@gmail.com

***Abstract*—Financial online news on social networks has been proven to be a crucial factor that causes fluctuations in stock market. Regarding the impact of financial online news, this paper introduces Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) to exploit the relation between financial online news and the fluctuations of stock price. A stock trends prediction model is proposed by deep combining the historical financial data feature, the news event feature and the sentiment orientation feature. News events and the corresponding sentiment orientations of financial news are introduced to help improving the accuracy of stock trends prediction. In order to verify the applicability of this model on different industries to predict the trend of individual stocks, two stocks are selected as the experimental objects, i.e., GREE Electric Appliances in the household appliance industry, and ZTE in the electronic appliance industry. The experiment results conducted on the past ten years data show that the proposed model improves the prediction accuracy about 10% and 20% in the household appliance industry and the electronic appliance industry, respectively, compared with the baseline algorithms.**

***Index Terms*—Stock Forecasting, Bi-Long Short Term Memory, Financial Online News**

I. Introduction

Stock market prediction has always been a popular and challenging task in financial time-series forecasting. It has been proven that the price of the stock market does not follow a random work and thus can be predicted to some extent [1]. However, as the stock market is volatile and influenced by many factors [2], such as global economy, political situations and other unexpected events, it is difficult to analyze all the related factors and make decisions according to a huge volume of data.

Recently, many Artificial Intelligence (AI) approaches have been investigated to predict stock market trends. These existing approaches can be mainly divided into two main categories: the pure historical data analysis and the combination analysis. In the pure historical data analysis, mathematics, such as stepwise regression analysis, is used to analyze the historical stock price patterns and predict stock prices in the near future [3]–[5]. Although these work [3]–[5] can achieve good results, it is hard to predict the stock prices accurately by using only the historical prices because unexpected events expressed on social media and financial news can also affect the stock price. Financial news has been proven to be a critical factor which causes the price fluctuation in stock market.

Therefore, the combination analysis approach has been proposed. It uses social media posts or news data along with historical stock market data to increase the accuracy of stock market prediction. Most of the previous work attempt to detect the sentiment orientations (positive or negative) of social media or financial news, and then include the sentiment factors to help stock market prediction. Results in work [6]–[8] have illustrated the effectiveness of combination analysis approach. However, some issues still need to be carefully addressed in the combination analysis approach.

First, how to combine different features from numerical and textual information is lack of research. To the best of our knowledge, many of the previous works often concatenate different features to form a high-dimensional feature matrix [7], [9]. It may face the curse of dimensionality because data turn out to be sparse in space of high dimension. Moreover, simplify concatenation of numerical and textual features sometimes may perform worse than the method of using only the numerical features [10]. Second, sentiment analysis is used to study the effect of financial news on stock market movements and generate the features of textual information. However, the sentiment orientation of financial news is not so obvious than that in the fields of product reviews, which brings challenges in detecting sentiment orientations of financial news. Moreover, as many financial word vocabularies are not included in the existing sentiment dictionary, which further decreases the effectiveness of sentiment detection of financial news [11].

To address the issues mentioned above, this paper proposes a stock price prediction method in a combination research approach with the aid of news event detection and sentiment orientation analysis. The event detection extracts the news events (what is it about?) and the sentiment orientation analysis deals with sentiment classification (what is the underlying sentiment towards the news event?). At first, the news events are extracted from the online financial news and then the sentiment orientations of the news events are detected. The news event feature and the sentiment orientation feature are combined with the historical price data in a concatenation way into the Long Short-Term Memory (LSTM) to predict the stock price.

The remainder of the paper is organized as follows. Section II reviews the related work. Section III gives the details of

This work is supported by the project of Double Top-Class Foundation of BFSU (No.YY19ZZA012).

978-1-6654-1485-2/20/$31.00 ©2020 IEEE
DOI 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00227

the proposed model for stock market prediction. Section V presents the experimental results and shows the comparisons with other methods. Section VI concludes the whole paper.

II. Related Work

The existing researches on stock trends prediction can be mainly classified into two categories: (1) pure historical models based on financial numerical data; (2) combination models based on financial numerical data and text information.

The traditional quantitative investment analysis [12] was the first attempt to adopt pure historical models. Financial numerical data, such as historical transaction data, company financial data and macroeconomic data, were analyzed by time series forecasting algorithms and then used to predict the stock price patterns. Chen et al. verified that historical stock price patterns can be used to predict the future trend of individual stocks [3]. Many traditional machine learning algorithms such as Support Vector Machine (SVM), decision tree and Native Bayes have been widely used to analyze the patterns of large volume of numerical data. Zhang et al. used SVM to predict the trends of individual stocks. It was shown in their empirical analysis that the prediction accuracy rate of individual stocks was greater than 60% [13]. However, in addition to financial numerical data, financial news will cause fluctuation in stock prices and thus should be taken consideration for stock volatility prediction.

Recently, some related works [14]–[16] have proposed combination models to integrate both historical stock price data and text information from financial news to improve the prediction performance. Manuel R. et al. used financial news titles and some technical indicators as input of deep learning models, such as convolutional neural networks (CNN) and recurrent neural network (RNN), to predict daily directional movements of stock price. The experimental results showed that news titles are more beneficial than news content to improve the prediction accuracy [14]. Chen et al. used a RNN model with Gate Recurrent Units to predict the stock volatility through news on Sina Weibo [15]. Cen et al. studied the impact of the sentiment orientations of online news and forum posts on stock market. It was shown that investors reacted more immediately and strongly to positive sentiment than those of negative information [16].

These existing stock prediction models mainly supplemented the financial numerical features of stock price patterns from the perspective of sentiment orientation analysis of financial news. The accuracy of sentiment orientation classification on financial news greatly affects the accuracy of stock trend prediction. However, the sentiment polarity of financial news is not so obvious as financial news mostly describes the object facts. Moreover, sentiment orientations of news highly relies on news events and news events can better explain the reason of stock price fluctuation than the sentiment orientation of news. Therefore, different from the existing works, this paper proposes to incorporate the news events detection and sentiment orientation of financial news to further improve the accuracy of stock prediction.

III. Stock Prediction Model

A. Model Architecture

Stock trend prediction is a binary classification problem, e.g., going up or going down. Fig. 1 depicts the flowchart of the proposed stock prediction model. Stock-related financial numerical data and news are crawled and preprocessed to build stock numerical databases and news databases. Then the features of financial numerical data and the features of news events and sentiment orientation of news are extracted as input to train the stock prediction model.

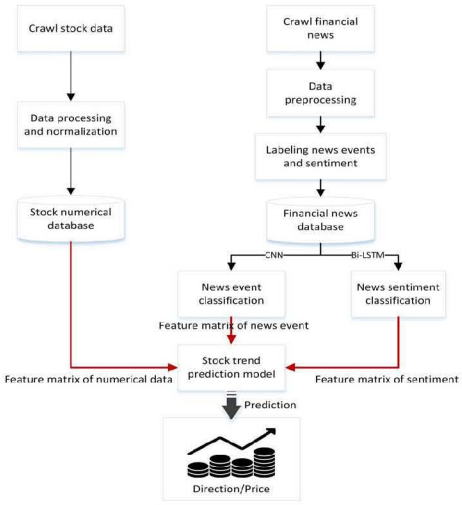

Fig. 1. Flowchart of stock prediction model based on the combination features of news event and sentiment orientation.

B. Extract features of financial numerical data

Relevant researches show that technical indicators such as company's price-earning ratio, price-to-book ratio and net inflow are related to the trend of individual stocks [17]. Therefore, this paper selects 4 kinds of technical indicators to generate a financial numerical dataset: (1) company's financial indicators, such as price-earning ratio and price-to-book ratio. (2) company's capital flow indicators, such as net flows and turnover rate. (3) stock price, such as opening price and closing price. (4) market index and sector index that reflect the price fluctuation and performance of 300 stocks traded in the Shanghai and Shenzhen stock exchanges.

The financial numerical data is collected from a large-scale financial data warehouse in China, namely Wind economic database [18]. Wind economic database relies on the scientific verification means and advanced management methods to ensure the accuracy of its stored data over 99.95%. Data preprocess is conducted on the collected financial numerical data in the following two steps. The first step is removing the records with missing data. If the data of a certain day is missing some indicators (e.g., for the reason of trading suspension), it will be removed. The second step is Z-score normalization. If the predicted model uses the original values

of the indicators, those indicators with larger orders of magnitude will play a more dominant role in the model, weakening the impact of the indicators with smaller orders of magnitude. Therefore, let d_{ij} be the *j-th* financial numerical indicator in the *i-th* day. Then the *j-th* indicator over T_n days will form a vector $D_j = \{d_{1j}, d_{2j}, ..., d_{nj}\}$. Each element in D_j is normalized as:

$$d_{ij} = \frac{d_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where μ_j and σ_j are the mean and standard deviation of D_j, respectively. Table I illustrates a sample of financial numerical matrix generated by p financial indicators collected over T_n days.

TABLE I
FEATURE MATRIX OF FINANCIAL NUMERICAL DATA

	D_1	D_2	...	D_j	...	D_p
T_1	d_{11}	d_{12}	...	d_{1j}	...	d_{1p}
T_2	d_{21}	d_{22}	...	d_{2j}	...	d_{2p}
...	...	...	...	...	...	...
T_i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{ip}
...	...	...	...	...	...	...
T_n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{np}

C. Extract features of financial news events

TABLE II
FEATURE MATRIX OF NEWS EVENT

Category	**News Event**
Transaction cluster	Resumption, Delisted, Capital inflow, Capital outflow, Block transaction
Equity cluster	Floatation, Listed, Buyout, Back-door List, Banner acquisition
Investment cluster	Bonus, Investment and construction, Issue bonds, Issue shares, Dividend-distribution
Corporate affair cluster	Rapid development, Change of registered capital, Unfavorable development, Expand business, Periodic report
External event cluster	Downgrade, Good rating, Exchange penalties, Monetary policy, Favorable policy

Financial news is collected from various online websites. The attributes collected in news data are title, publication date and publisher. Text pre-processing is done on the collected news data. Jieba, a Chinese text segmentation tool, is used to segment each piece of news into a sequence of tokens. Then all stop words are removed based on a Chinese stop word dictionary. To improve the accuracy of word segmentation, some financial-domain vocabularies are supplemented, including commonly used financial words, stock code, abbreviation of A-share listed company and etc.

The financial news data are analyzed and categorized to different clusters of objective news events. According to the keywords of news entries in CSMAR economic database [19], we define 82 kinds of objective news events that may cause fluctuations in stock market. Table II lists part of the defined news events.

As CNN model shows superiority in text semantic feature extraction and short text classification, it is adopted to build news event classification model. The detailed process of CNN-based news text classification has been described in our previous work [20]. Here we briefly introduce the key idea. There are five layers in the CNN-based news event classification model, including input layer, convolutional layer, pooling layer, fully connected layer and output layer. The output of each layer is the input of next layer. First, word2vec is used to transform news titles to a high-dimensional word vector matrix, representing each word with a h-dimensional vector. This word vector matrix is taken as the input of convolution layer. The convolutional layer uses filters to perform a convolution operation on the word vector matrix and then generates feature maps. The pooling layer samples these feature maps to extract the most important features and then transmits the extracted features to the fully connected layer. Finally, the fully connected layer with softmax as its activation function will output the classified results of news titles.

The CNN-based news event classification model will classify each piece of news to one or several kinds of news events. Then the frequency of news events in each day can be calculated. $\{E_1, E_2, ...E_q\}$ represents q kinds of news events and q is the number of news events. Let e_{ij} represent the frequency of news event E_j occurred in the date T_i. Table III illustrates the frequency matrix of news events.

TABLE III
FEATURE MATRIX OF NEWS EVENTS

	E_1	E_2	...	E_j	...	E_q
T_1	e_{11}	e_{12}	...	e_{1j}	...	e_{1q}
T_2	e_{21}	e_{22}	...	e_{2j}	...	e_{2q}
...	...	...	...	...	...	...
T_i	e_{i1}	e_{i2}	...	e_{ij}	...	e_{iq}
...	...	...	...	...	...	...
T_n	e_{n1}	e_{n2}	...	e_{nj}	...	e_{nq}

D. Extract sentiment features of news data

News events are the objective descriptions of news and sentiment analysis further reveals the sentiment orientations towards the news events. Bi-LSTM is used to extract the sentiment feature of financial news as it can achieve more precise sentiment semantic expression by using a forward LSTM and a backward LSTM to integrate both the future context and the previous information.

Formally, given a sequence of news titles transformed by word2vec embedding vectors, each piece of news title $d = [x_1, x_2, ...x_t, ...x_N]$ will be represented as a $N \times K$ dimensional vector, where $x_t \in \mathbf{R}^K$ is the embedding vector of the *t-th* word in d and N is the number of words in d.

The key element of the LSTM is the memory cell, which is controlled by 3 different gates, forget gate f_t, input gate i_t and output gate o_t. Given the input vectors at time t is x_t and the output of previous moment is h_{t-1}, the state of previous hidden layer is C_{t-1}, the sigmoid function σ of the forget

gate determines which information of the previous cell state is remembered or forgotten. The coefficient f_t is denoted as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2}$$

Then input layer decides the new information which should be stored in the memory cell. This process consists of two steps. First, the sigmoid layer of the input gate determines which information of the input vector x_t are retained. Then the tanh layer of the input gate creates new candidate information denoted by $\overline{C}_t$. The current status of the memory cell, defined as C_t, can be obtained as follow:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$\overline{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \overline{C}_{t-1} \tag{5}$$

The output gate decides the output information by using sigmoid threshold and tanh function as follow:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \odot \tanh(C_t) \tag{7}$$

In functions (2)-(7), W_* and b_* represent weight matrix and bias term, respectively, and $\odot$ means component-wise multiplications.

Based on the above calculations, we can obtain the N-*th* output status at time $t = N$, denoted as h_N. The last stage of Bi-LSTM model is a traditional fully connected layer, which takes h_N as input and outputs the probability distribution of sentiment labels.

$$y_d = \text{softmax}(W_o \cdot h_N + b_o) \tag{8}$$

Then sentiment orientation of d can be calculated as:

$$\text{so}_d = [1, -1] \cdot y_d \tag{9}$$

where $\text{so}_d \in [1, -1]$. if $\text{so}_d > 0$, the sentiment orientation of the news title d is positive; otherwise, the sentiment orientation of d is negative.

Supposing that there are m_i pieces of news in T_i day, the sentiment orientation of news j is so_j, $j = 1, 2, ..., m_i$, $\text{so}_j \in [-1, 1]$. Then the sentiment score of news in T_i is calculated as:

$$S_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \text{so}_j \tag{10}$$

$S = [S_1, S_2, ..., S_n]'$ is the feature vector of sentiment orientation over T_n days and is added as one column of data in Table III to generate a total feature matrix of financial news.

E. Stock Trend Prediction

The stock trend prediction takes the feature matrix of financial numerical data and the feature matrix of financial news as input and outputs the prediction results, i.e., going up or going down. Suppose that the lookback size is a days, the prediction model uses the matrixes of data from the past $t-a$ to $t-1$ days to predict the stock price on the t-*th* day.

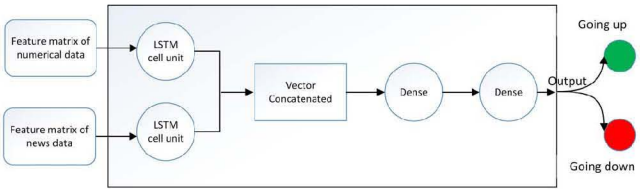

Fig. 2. The process of stock trend prediciton.

Fig. 2 illustrates the process of stock trend prediction. The feature matrix of financial numerical data and the feature matrix of news are taken as input to two separate LSTM cell units. Here we do not concatenate the two feature matrixes to form a single one, as the concatenated matrix will form a high-dimensional sparse matrix, leading to the vanishing gradient problem.

The outputs of two LSTM units are concatenated as input to a fully connected layer. Finally, the fully connected layer will output a binary class label indicating that the stock price will increase or decrease.

IV. EXPERIMENTS

A. Dataset Description and Model Settings

Two stocks in different industries are selected as the experiment subjects: the stock of GREE electronic application (000651.SZ) and the stock of ZTE corporation (000063.SZ). Historical stock price data and financial news data of each stock in the period from 2010-02-09 to 2020-03-02 are acquired from the Wind finance dataset. For each trading day, twelve stock price indicators are collected, i.e., opening price, lowest price, highest price, closing price, the net inflow of main fund, price-earnings ratio, price-to-book ratio, turnover rate, change, rise and fall, Shenzhen Stock Index A, and household appliance sector index. The experiment dataset collects a total of 29352 trading day of stock price data and 18796 news data about ZTE corporation, and a total of 29364 trading day of stock price data and 18766 news data about GREE electronic application. We then split the dataset into training part, validation part and testing part, using 75% of them to be the training dataset, 15% of them to be the training dataset, 15% of them to be the testing dataset.

We compare the proposed model, termed as financial-news model, with (1) Financial model: the proposed model with only the financial numeric matrix as input; (2) Financial-event model: the proposed model with the financial numeric matrix and the news event matrix as input; (3) GBDT model: the Gradient Boosting Decision Tree (GBDT) algorithm is used for stock prediction with the same input as the proposed model. Accuracy rate Acc is adopted as the metric to evaluate the prediction performance, which is the number of correctly prediction samples on the entire dataset.

B. Results of Stock Trends Prediction

Table IV lists the accuracy performance comparison of stock trend prediction among the 4 methods. There are three

TABLE IV
ACCURACY PERFORMANCE COMPARISONS OF STOCK TREND PREDICTION

Stock	Financial	Financial-event	Financial-news	GBDT
GREE	0.6000	0.7143	0.8571	0.5789
ZTE	0.5000	0.5283	0.5714	0.5588

interesting findings. First, the proposed model can achieve the best prediction accuracy performance compared with the rest 3 methods, It achieves 85.71% in predicting the stock price movements of GREE electronic application. Second, Table IV indicates the effectiveness of introducing news data to stock prediction. The accuracy of stock forecasting increases by changing the model input from the financial numeric matrix to a combination matrix of financial numeric data and news data. As we can see, the prediction accuracies of Financial model, Financial-event model and the proposed model are 60%, 71.43% and 85.71% in GREE stock prediction, respectively. Last, the performance of stock prediction model varies in different industries. The proposed model performs better in predicting the price movements of GREE stock (family appliance industry) compared with that of ZTE corporation (electronic appliance industry). The reason lies in that the electronic appliance industry is a defensive industry and is less affected by external economic factors.

Fig. 3 and Fig. 4 demonstrate the predicted and the real stock closing prices of GREE and ZTE stocks, respectively. As the stock trend prediction is a binary classification problem, it can not tell the stock closing price directly. The lookback value is set to 13 and let the actual closing prices of 13-th day and 14-th day be cp_{13} and cp_{14} , respectively. Therefore, if the prediction in the 14-th day is a rise, the predicted closing price of the 14-th day is set to be the sum of cp_{13} and the difference between cp_{13} and cp_{14}. If the prediction is a fall, the predicted closing price is set to be the difference between cp_{13} and $|cp_{14} - cp_{13}|$. The blue lines and the red dots in the figures denote the true values and predicted values, respectively. It is observed that the red dots fit well with the blue lines which indicates the proposed model are capable to predict the stock price tendency.

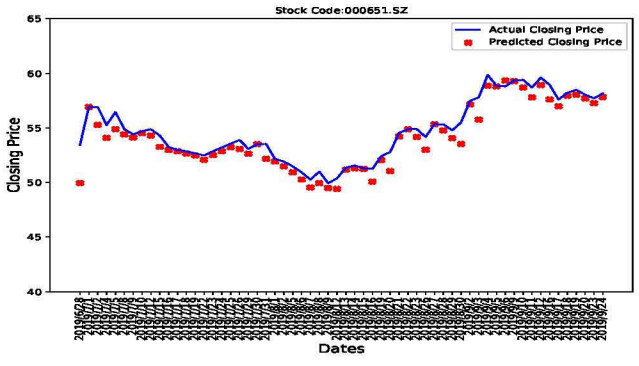

Fig. 3. The predicted and real values of stock closing prices of GREE

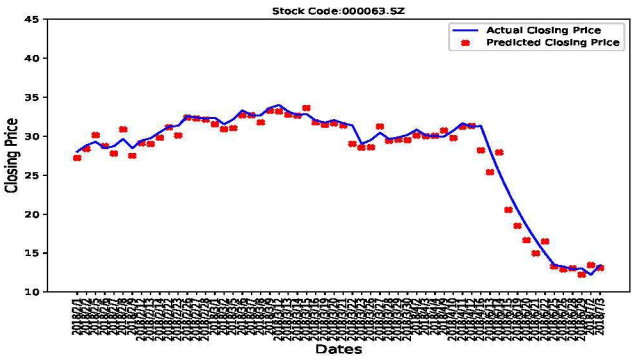

Fig. 4. The predicted and real values of stock closing prices of ZTE

TABLE V
Comparisons of lookback (a) among different models on ZTE.

a	ZTE			
	Financial	Financial-event	Financial-news	GBDT
7	58.33	**67.31**	55.78	51.67
8	40.63	**54.56**	**54.56**	42.35
9	50	**53.33**	**60**	53.5
10	42.31	**59.26**	**62.92**	55.32
11	41.67	**48.15**	**59.26**	48.39
12	36.36	**41.67**	**41.67**	39.57
13	50	**52.38**	**57.14**	47.23
14	50	42.86	**52.38**	**55.88**

C. Impact Factor on Stock Trend Prediction

Regarding to the impact factors on stock trend prediction, our work measure the parameter settings of lookback value a and threshold value λ. If lookback value a is 13 days, the prediction model will use the collected data from the past 1 to 13 days to predict the stock price on the 14-th day. The threshold value is set to eliminate the price jitter in labeling the stock going up or going down. In our experiment, if the increase rate of the stock price is greater than a preset threshold (e.g., 1%), then the price on that day is labeled as going up; otherwise, it is labeled as going down.

Table V and Table VI show the experiment results with lookback value a ranging from 7 to 14. Though the prediction accuracy fluctuates without a stabilized pattern along with switching a from 7 to 14, it shows that the stock prediction model is sensitive to the lookback value. In ZTE, the optimal value of a is 10, which improves the prediction accuracy

TABLE VI
Comparisons of lookback (a) among different models on GREE.

a	GREE			
	Financial	Financial-event	Financial-news	GBDT
7	50	**61.53**	53.85	48.44
8	43.75	**54.54**	**54.55**	48.43
9	51.74	40.0	**50**	41.36
10	53.85	**66.67**	**77.78**	49.30
11	41.67	**55.56**	**55.56**	**58.3**
12	54.54	**62.5**	**62.5**	49.20
13	60	**85.71**	**85.71**	56.52
14	60	**71.43**	**71.43**	50.0

TABLE VII
Comparison of threshold (λ) among different models on ZTE, $a = 10$.

λ	ZTE			
	Financial	Financial-event	Financial-news	GBDT
0.01	0.45	**0.5714**	**0.6190**	0.4688
0.02	0.45	**0.6190**	0.5238	0.4715
0.03	0.45	**0.5714**	**0.5714**	0.4727
0.04	0.50	**0.5238**	**0.6190**	0.4719
0.05	0.55	**0.6667**	0.5714	0.4796
0.06	0.45	**0.5714**	0.5238	0.4685
0.07	0.55	0.4238	**0.5714**	0.4712
0.08	0.50	**0.5714**	0.5238	0.4715

TABLE VIII
Comparison of threshold (λ) among different models on GREE, $a = 13$.

λ	GREE			
	Financial	Financial-event	Financial-news	GBDT
0.01	0.50	**0.8751**	0.7143	0.5670
0.02	0.60	**0.8751**	**0.8751**	0.5630
0.03	0.40	**0.8751**	**0.8751**	0.5648
0.04	0.50	**0.8751**	**0.8751**	0.5600
0.05	0.50	**0.8751**	0.7143	0.5613
0.06	0.60	**0.8751**	**0.8571**	0.5626
0.07	0.50	**0.8751**	**0.8751**	0.5661
0.08	0.50	**0.8751**	**0.8751**	0.5583

about 3% compared with the setting of $a = 7$. In GREE, the optimal value of a is 13, at least improving the prediction accuracy about 10% in comparison with setting $a = 7$. It also reveals that the Financial-event model and the Financial-news model consistently outperform the Financial model, and the Financial-news model achieves the highest accuracy of prediction.

Table VII and Table VIII show the experiment results with threshold value λ ranging from 0.01 to 0.08. The prediction accuracy doesn't flow in accordance with a specific pattern. But the experiment results illustrate that Financial-event model outperforms Financial model in most of the thresholds by about 10% on average in ZTE and about 20% on average in GREE. Moreover, we can see that Financial-news model does only the same to but no better than Financial-event model. It proves that threshold values are less important compared with a certain lookback between the two models. At last, comparison between Financial-news model and GBDT model provides us with evidence that LSTM model performs better in stock trend prediction than GBDT model.

Comparing Table VII and Table VIII, statistics of GREE are not sensitive to the ranging thresholds. We ran through the experiments again with ranging thresholds λ but under a reset lookback value a and found that the prediction accuracy on GREE was more sensitive to a rather than λ. Furthermore, GREE source data scarcely has the increasing rate ranging from 0.01 to 0.08. Therefore, switching λ from 0.01 to 0.08 has no effect on the prediction performance in Table VIII and they mostly remain the same.

V. Conclusion

This paper exploits the impact of financial online news on the fluctuations of stock prices. The stock historical numeric data as well as the features of news event and sentiment orientation are incorporated into a two-unit LSTM model to predict the stock trends movement. To evaluate the effectiveness of the proposed model, two individual stocks on different industries are selected as experiment objects. The experiments conducted on the past ten years data show that the proposed model greatly improves the prediction accuracy compared with the historical numeric-data-only method. It also indicates that financial online news is a crucial factor causing fluctuations on stock market.

References

[1] X. Q. SUN, H. W. SHEN, and X. Q. CHENG, "Trading network predicts stock price," *Scientific reports*, vol. 4, pp. 1–6, 2015.

[2] A. Bastianin and M. Manera, "How does stock market volatility react to oil price shocks?" *Macroeconomic Dynamic*, vol. 22, no. 3, pp. 384–412, 2018.

[3] K. Chen, Y. Zhou, and F. Dai, "A lstm-based method for stock returns prediction: A case study of china stock market," *Proceeding of IEEE International Conference on Big Data*, pp. 32 823–2824, 2015.

[4] I. I. Hassan, "Exploiting noisy data normalization for stock market prediction," *International Conference on Engineering and Technology*, vol. 12, no. 1.

[5] Seungwoo, Jeon, Bonghee, and et al., "Pattern graph tracking-based stock price prediction using big data," *Future Generations Computer Systems Fgcs*, vol. 80, pp. 171–187, 2014.

[6] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Systems with Application*, vol. 83, pp. 187–205, 2017.

[7] R. Akita, A. Yoshihara, T. Matsubara, and et al., "Deep learning for stock prediction using numerical and textual information," in *Proceeding of IEEE/ACIS International Conference on Computer and Information Science*. IEEE, 2016, pp. 978–984.

[8] M. Usmani, S. H. Adil, K. Raza, and et al., "Stock market prediction using machine learning classifiers and social media, news," in *Proceeding of International Conference on Computer and Information Sciences*. IEEE, 2016, pp. 322–327.

[9] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and et al., "Predicting stock returns by classifier ensembles," *Applied Soft Computing*, vol. 11, no. 2, pp. 2452–2459, 2011.

[10] M. ZHANG, W. DU, and N. ZHENG, "Predicting stock trends based on news event," *Data analysis and knowledge discover*, vol. 3, no. 5, pp. 11–17, 2019.

[11] T. Loughran and B. Mcdonald, "When is a liability not a liability? textual analysis, dictionaries, and 10ks," *Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[12] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of arima and articial neural networks models for stock price prediction," *Journal of Applied Mathematics*, pp. 1–7, 2014.

[13] Y. C. ZHANG, Z. C. ZHANG, and Z. HUANG, "Application of support vector machine in selecting high quality stock," *Statistics and decision*, vol. 4, pp. 163–165, 2008.

[14] M. R. Vargas, C. E. M. dos Anjos, G. L. G. Bichara, and et al., "Deep learning for stock market prediction using technical indicators and financial news articles," *Proceeding of International Joint Conference on Neural Networks*, pp. 1–8, 2018.

[15] W. Chen, Z. Yan, K. Y. Chai, and et al., "Stock market prediction using neural networks through news on online social networks," *Proceeding of International Smart Cities Conference*, pp. 23–29, 2017.

[16] Y. Cen, Z. Tan, and C. Wu, "Impact of financial media information on stock market: an empirical study of sentiment analysis," *Data Analysis and Knowledge Discover*, vol. 3, no. 9, pp. 98–114, 2019.

[17] E. F. Fama and K. R. French, "Size and book-to-market factors in earnings and returns," *The Journal of Finance*, vol. 50, no. 1, pp. 131–155, 1995.

[18] "Wind economic database: https://www.wind.com.cn/en/edb.html."

[19] "Csmar economic database: http://cn.gtadata.com."

[20] Y. Xu, Y. Liu, and L. Cai, "Predicitng retweets of government microblogs with deep-combined features," *Data Analysis and Knowledge Discovery*, vol. 4, no. 2, pp. 18–28, 2020.