Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 174 (2020) 128–140

Procedia Computer Science

www.elsevier.com/locate/procedia

2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)

Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM Model

Yinghao Ren[a], Fangqing Liao[a], Yongjing Gong[a]

School of Computer Science, Chongqing University, Chongqing, China
1084096870@qq.com
Warran College, University of California: San Diego, Beijing, China
liaofangqing2008@163.com
Shanghai University, Shanghai, China
18800202738@163.com

These authors are contributed equally to this work

Abstract

With the in-depth study of the stock market, the impact of news media on stock prices has gradually been paid attention to. However, in previous studies, most of them used the real-time rise and fall of stock prices to reflect the impact of news on stock prices, ignoring that it takes time for investors to react to news. In this paper, considering the lag of investors' response to stock price, we choose BIAS as a measure index after news happened for a period of time to analyze the impact of news media on stock price trends. Based on the DBLSTM (Deep Bidirectional Long Short-Term Memory) model, we establish a model to predict the short-term trend of stock prices using news text data. Experiments show that the model we adopted outperforms other models in terms of prediction accuracy.

Keywords: DBLSTM, BIAS, Stock Market

1. INTRODUCTION

The efficient market hypothesis (EMH) has laid the foundation for modern stock investment, which demonstrates the capability of the price of assets in reflecting global information. Based on which, the investors can reasonably evaluate the value of assets [1]. However, realizing the global financial crisis in 2008 and the turmoil in China's stock market in 2019, rational decision-makers may find that EMH is not always practical. Meanwhile, affected by the risk of decision- making and other complex factors, investors cannot make fully rational decisions. For example,

10.1016/j.procs.2020.06.068

stock investors in China are mostly composed of retail investors. The inadequacy of relevant professional knowledge and skill training make them more vulnerable due to the uncertainties caused by information asymmetry. More importantly, they are often influenced by emotions and market environment, resulting in large fluctuations in the stock market [2]. Hence, we should consider the critical roles of behavioral finance in the field of modern economics, the impact of investors' psychological factors and market sentiment on the stock market when investigating the stock investment.

News has a direct impact on short-term stock price movements according to [3]. Detailedly, news dissemination through the media can play an essential role in influencing stock prices, leading to emotional investors easily to be affected by the news [4]. For example, a piece of news revealing the scandal of a company may induce negative attitudes among the investors, leading to the decrease of stock price consequently because it is conceivable for the investors to sell the stocks. Additionally, investors are apt to take similar investment actions as others when facing a new event based on the herd effect theory, which makes the change of stock price more traceable. News on stock market is continuously updated every day and the types of news exposed to the investors are also varied. This news tends to spread emotions. Investors are vulnerable to these market emotions when making investment decisions. Therefore, how to analyze the impact of news on the stock price has become a hot research topic. In this paper, we intend to establish a model to study the effects of news texts on stock prices.
Two critical factors are taken into consideration when evaluating the response of stock price to the news: the reaction time and the degree of reaction. We choose BIAS as a technical index in this paper, which will be explicitly elaborated in section III.

According to the facts that the Chinese investors are accustomed to getting stock-related information from the media, we crawl a lot of news data about the medical industry from mainstream financial websites to predict the stock's reaction to the news. In order to better process data and achieve prediction results, we choose LSTM, which is an improved model based on RNN model. This machine learning model will be described in detail in section III.

2.RELATED WORK

The issue of the stock price has received considerable critical attention because it can reflect the financial information behind it, arousing academic frenzy since the 1980s. In the pioneer studies, news reports on stocks have become the object of study as important information [5].

First of all, studies have shown that the news has an impact on stock prices. In detail, through comparing different effects of different media on stock prices, Mao et al. in [6] found that news on the Internet media would play a more significant role in investors' decision-making process. Besides, Tetloc et al. in [4] pointed out that when Wall Street's pessimism rose, the overall market rewards would fall in the next day.

Secondly, relevant studies describe that news impacts investors' decision-making process by conveying information to them. Recently, due to the popularity of social media, researches focusing on the impact of Internet public opinion on the stock market have increased dramatically [7]. As for online public opinion, studies are more concerned with emotions, such as investor sentiment on financial forum websites, social networking sites (Twitter, Weibo) and other platforms on the impact of local (individual) stock prices or overall stock market volatility and so on [7][8]. Different indicators and methods have been employed in different studies to measure the investor sentiment [9][10][11], such as the range of rising and falling, turnover rate, PSY psychological line, closed- end fund discount and so on [12]. However, none of these indicators can effectively demonstrate the timely response of prices to information, that is, the phenomenon of stock price delay. Thus, we choose a new indicator, namely BIAS to evaluate the impact of news on stock prices. In terms of research methods, many scholars adopt news text processing to measure investor sentiment. For example, Jin et al. utilized KNN algorithm to divide relevant news posts into three categories: bullish, neutral and noisy, thus constructing investors' bullish index and opinion convergence index [Jin et al.'s citation]. Other prevailing mechanisms include Bayesian algorithm, SVM model and so on.

Thirdly, many scholars have attempted to study the influence of sentiment on stock prices. Detailedly, Wenlong et al. in [6] put forward a new kernel function, called semantic kernel function and structural kernel function to predict the

stock price based on support vector machine (SVM). In [14], Nam et al. designed a program to predict stock price movements in light of financial news which happens causally. Both the relationship between news and stocks and the causal relationship between companies are taken into consideration in [14].

3.METHODOLOGY

RNN is a time-series neural network. The interconnection structure between the hidden layers reflects the interaction between time series. However, there are vital problems exist in RNN: the fast gradient descent problem and non-convergent problem [citation]. Fortunately, the bi-directional LSTM model can solve the gradient problem of RNN network by adding gates and using the context relation of forward and backward time directions in time series, improving the prediction accuracy subsequently [15].

A. LSTM Neural Network Model

In 1997, Hochreiter and Schmidhuber proposed LSTM, which had achieved surprising performance in the NLP field[16]. LSTM aims at resolving long-term dependence problem based on improved RNN (Annotation) neural network. Keeping information in mind for a long time is an inherent characteristic of LSTM [17]. All RNN models have a chain form of repetitive neural network modules. As shown in Fig. 1 about the standard RNN models, this repetitive module is usually straightforward in structure, such as a tanh layer.

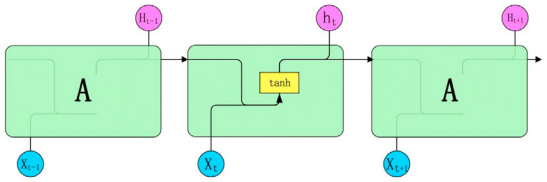

Fig. 1. Standard RNN Model Structure

Similarly, as a variant of RNN, LSTM also has this chain module structure, shown in Fig.2, but with different repetitive modules and layers. As Fig.2 shown, LSTM has three more gates than RNN with only the tanh layer.

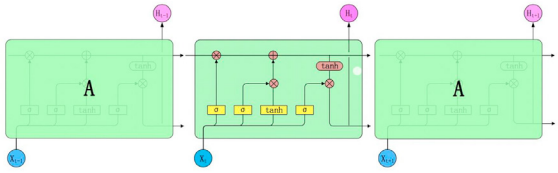

Fig. 2. LSTM Model Structure

The key to LSTM is the cell state, which is the horizontal line that runs through the top of the chart. LSTM deletes or adds information to the cell state. This alteration changes the structure of cell state information, which are gates. Conclusively, LSTM has three gates.

3.1 .Forgotten Gate

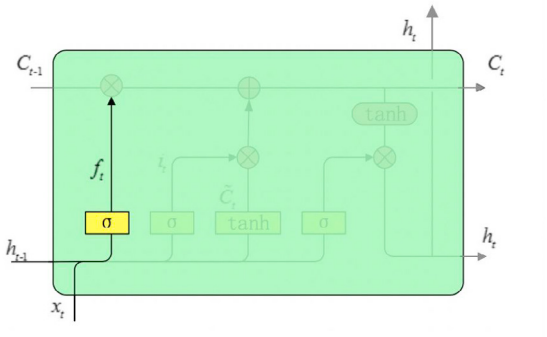

Fig. 3. Forgotten Gate

LSTM processes sequential data from left to right. In the face of a large amount of miscellaneous information, the forgetting gate decides which information of cell status is lost. According to h_{t-1} and x_t, the forgetting gate calculates $f_t \in 0,1$ as the input of state C_{t-1}. $f_t = 0$ is for "total discarding" meaning all discarded and $f_t = 1$ for "total acceptance" meaning all accepted. f_t evolves as follows:

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

where W_f and b_f represents the weight and error of forgetting gate, respectively.

3.2 .Input Gate

The input gate consists of two parts: the first part is a sigmoid function, the output(it)is to decide which value to update; the second part is a tanh activation function, the output is C t. The multiplied results of it and C t are used to update the cell state. The formulas are as follows:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i), \quad (2)$$

$$C_t = \tanh(W_c * [h_{t-1}, x_t] + b_c), \quad (3)$$

where W_i and W_c represent the corresponding weight and b_i, b_c present the error.

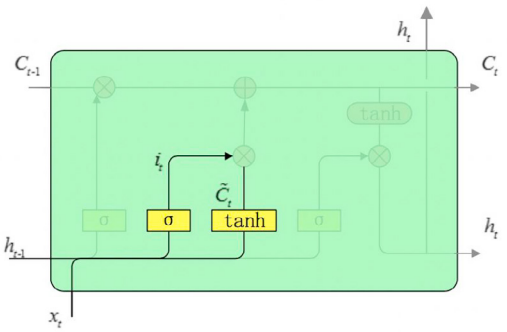

Fig. 4. inputGate

Combining the forgotten gate and the input gate, the cell status is updated to Fig. 5:

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (4)$$

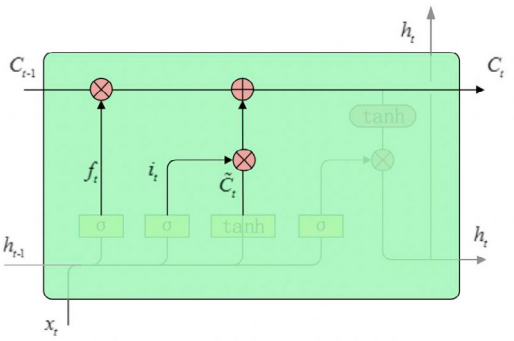

Fig. 5. Updated cell status

3. 3 Output Gate

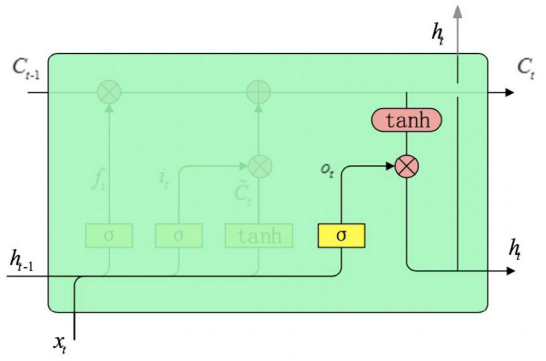

Fig. 6. OutputGate

After forgetting the gate and the input gate and completing the cell status update, the output gate determines the output information. The first layer is the sigmoid layer, which determines which parts of the output cell state, and the second layer is a tanh function, which deals with the results of the first layer. Then, the outputs of the two layers o_t and tanh are multiplied to determine the final results.

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o), \quad (5)$$

$$h_t = o_t * \tanh(C_t), \quad (6)$$

where W_o and b_o represent the weight and deviation of the output gate respectively, and h_t is the final output value.

Dropout

Dropout is a regularization technique of neural network model proposed by Srivastava et al[18]. Dropout is a tech- nique of randomly ignoring neurons in training process. They randomly "Dropout" means that the contribution to the activation of downstream neurons is temporarily eliminated and no weight updates will occur in these neurons. When the neural network is trained, different neurons will adjust their parameters according to different characteristics. Neurons become dependent on this particular feature, some of which are even harmful. Learning too many of these features may lead to the decline of generalization ability of the model, and can not adapt to data other than training data. The formula is as follows: The formula is as follows:

$$y = f(W * d(x)) \quad (7)$$

$$d(x) = \begin{cases} mask * x, training \\ (1 - p)x, else \end{cases} \quad (8)$$

Among them, p is the dropout library, and mask $1 - p$ is the binary vector produced by Bayesian effort distribution with probability.

Dropout does not modify the cost function, but the deep network itself. Dropout randomly "deletes" some hidden neurons in the network, keeping the input and output neurons unchanged. In this way, for a network, dropout is like training several different neural networks with the same data, resulting in a number of different degrees of fitting state. But these networks have a common loss function, which is equivalent with optimizing the neural network itself and getting the average value of all states. At the same time, it reduces the synergistic relationship between the neural units and increases the robustness of the network.[19]

Deep Bidirectional LSTM

Deep Bidirectional Long Short-Term Memory (DBLSTM) is divided into an input layer, a bidirectional LSTM part, a full connection layer part and an output layer. The bidirectional LSTM part is composed of multilayer bidirectional LSTM, and the whole connection part is composed of a multilayer full connection layer, shown in Fig.7.

DBLSTM has the following advantages: 1) it can avoid RNN gradient disappearance and gradient explosion in long time series; 2) it can learn time-dependent information; 3) it can make use of the context relationship of forward and

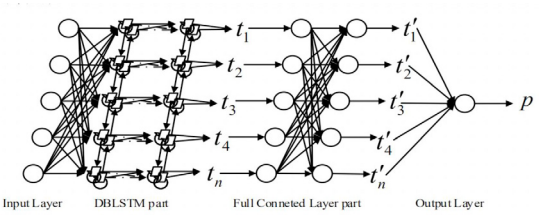

Fig. 7. Deep Bidirectional LSTM

backward time series. In order to realize deep data mining, multiple bidirectional LSTM layers are superimposed to learn the deep features of time series through multilayer neural network structure. In addition to adding a bidirectional LSTM layer, you can also add full connected layers (FC). FC has good nonlinear mapping ability,

and can weigh the nonlinear characteristics of bidirectional LSTM output, that is, to com- bine these nonlinear features. This process is essential to learn a (nonlinear) equation in a vector space, and to learn these nonlinear combinatorial features in a simple way of weight. However, with the increase of the network layer, the training difficulty of the model increases, the convergence speed slows down, and the problem of over-fitting is easy to occur. So Dropout strategy is used to solve these problems. The principle of Droupout is intuitively to stop the output of neurons with pre-set probability when training network. The ”strike” of some neurons means that only part of the data features is involved in the training of each network, so as to prevent the network from learning too much of the data features of the training set and achieve the purpose of preventing over-fitting.

BIAS

BIAS calculates the percentage difference between a market index or closing price and a moving average. Thus, the index reflecting the deviation degree between price and its moving average in a certain period of time can be obtained, and the possibility of price backing or rebound caused by deviating from the moving average trend when price fluctuates violently can be obtained, as well as the reliability of price moving within the normal range of fluctuation to form a continuing potential. The calculation formula is as follows:

$$\mathrm{BIAS} = \frac{\mathrm{P} - \mathrm{P_i}}{\mathrm{P_i}} * 100\% \quad (9)$$

where P represents the closing price of the stock on the day of news occurrence, and P (i) represents the average price of the stock after the i^{th} day of news occurrence.

4. EXPERIMENT

4.1 Model

We have established a prediction model based on BDL- STM’s stock price trend. In this model, we use news data to predict the change of BIAS. The model is shown in Fig8.

First, we designed a web crawler program to obtain news data, and then removed the symbols and garble codes in these news data. For these processed data, we need to analyze the emotions it contains, so we designed a word segmentation and vectorization process. Through “jieba” word splitter and word2vec model, we can obtain a set of vectors that can be recognized by the machine as the input layer. We need to predict the trend of stock prices in a certain period of time, so we need to obtain the corresponding BIAS data. As for the raw data of BIAS, we cleaned it and put it into the DBLSTM model with news vector for training. Finally, the prediction accuracy of the model will be output. In order to compare the prediction effect under different conditions, we will also carry out some comparative experiments under this model framework.

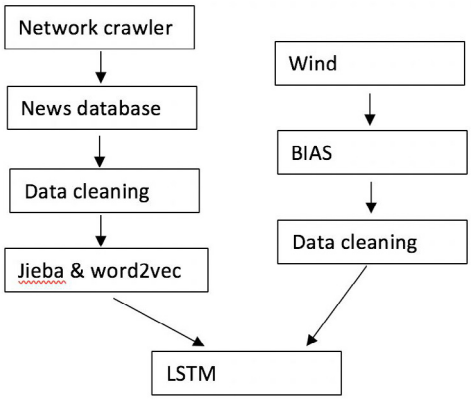

Fig. 8. Stock Trend Predicting Model

4.2 Data sources

We analyze real news data as experimental subjects. Chinese financial information media are becoming more and more mature, such as Sina Finance and Economics, Oriental Wealth Network, Cathay Tai'an Data Terminal and so on, which have become the main sources for the stock investors to obtain news information. According to the data released by Quest Mobile in 2018, Sina's financial and economic clients have covered more than 10 million people monthly. Additionally, according to the statistics of Oriental Wealth Network, more than 30 million users are accessed independently every day, and more than 100 million users are accessed independently in a single month, which is the largest financial and economic information portal in China. In order to get news that can truly reflect investor sentiment, we use the Internet crawler to get 18,000 news from these financial data media, including the title, author, source and release time. In order to ensure that the influence of news on investors is not dispersed (theory comes from: experimental study of media reports affecting investment behavior based on investors' concerns, Xi'an Jiaotong University) [20], we select a total of 18984 stock news and industry news from 24 listed companies in the same industry, from March 10, 2013 to December 30, 2017. The following table shows the distribution of news quantity:

TABLE 1.STOCK DISTRIBUTION

Stock	News	Stock	News
000423.SZ	1125	600056.SH	1075
000538.SZ	1190	600079.SH	992
000661.SZ	792	600085.SH	814
002252.SZ	1431	600196.SH	1813
002399.SZ	802	600252.SH	895
002424.SZ	531	600332.SH	1192
300015.SZ	1029	600436.SH	784
300049.SZ	741	600518.SH	1708

300142.SZ	830	600645.SH	1240
Total		18984	

After the financial data get news data, in order to reflect the specific role of these news on stocks, we need to know the market performance of these stocks after the news happened for a period of time. We selected BIAS as an evaluation index. On the sixth day, we obtained the BIAS data of 16 listed companies from March 10, 2013 to December 30, 2017 via Wind Financial Data Terminal. According to the BIAS value if BIAS > 0, the stock price has an upward trend and vice versa. In order to facilitate data analysis, we put a "1" label on stocks with an upward trend and a "1" label on stocks with a downward trend.

Instead of analyzing data immediately, we need to preprocess text first. News data processing is a significant part of our experiment. Because the original data cannot be directly used for modeling, and news data is Chinese text, as well as some numbers and symbols, cannot be directly recognized by the machine. We use data processing to obtain continuous, machine-recognizable data. Firstly, some unnecessary information in news content text should be removed, such as website address, scrambling code, punctuation marks, numbers and so on. Then, the Chinese word segmentation is needed for the text. At present, the research on Chinese word segmentation is very mature. We choose to use python's Jieba Chinese word segmentation library to segment news content, and then remove the stop words.

We chose Word2vector as the model to process the news data. Word2vec is a group of related models that are used to produce word embeddings. Word2vec is a toolkit for acquiring word vectors launched by Google in 2013. It is efficient and straightforward, so it is widely used.[21][22]

4.3 Experimental Model

After processing BIAS data and news text data, we need to build a binary classification model with LSTM. We transfer the problem of stock price forecasting into a two-class problem under different changing rate thresholds: when the changing trend is less than the threshold, it is judged as a negative sample, i.e., the stock price has a downward trend and vice versa.

(1)Selection of Input Layer and Output Layer

We use the processed news data and BIAS data as the input layer, and ultimately output the news data to predict the accuracy of BIAS changes.

(2)Selection of Implicit Layer

LSTM implied layer has only one new parameter, that is, n_{hidden}. This parameter is used to remember and store the number of nodes in the past state. In this paper, we set the number of the initial hidden layer and nodes in the hidden layer as 1 and 128, respectively.

(3)Establishment of Contrast Model

The measurement index of the model in this paper is the accuracy index, that is, the correct number of classifications in the model accounts for the proportion of all classifications. The following formulas are used to express the correct rate:

$$\text{Accurary} = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i = y_i) \quad (10)$$

where $\hat{y}_i$represents the i^{th} sample's predication type, y_i represents the i^{th} sample's real type, 1(x) represents when the predicted results are consistent with the real results, the value is 1, and in other cases is 0.

In order to present the performance of LSTM mechanism, we have established several contrast models to be compared, including the model based on news data volume, the model based on different algorithms and the model based on different indicators. Firstly, based on deep bidirectional LSTM model, we use all the news data to model and get the accuracy. Then we use 50%, 75% and 90% data to model and get the accuracy. Then, we use the SVM model and the LR model to model all the news data volume, and compare the accuracy with that based on Deep Bidirectional LSTM model. Finally, we analyze the indicators to measure the trend of stock movements. Based on deep bidirectional LSTM model, we compared the accuracy between two indexes: BIAS and the ratio between the closing market price and open market price.

4.4 Comparative Analysis of Model Empirical Results

Fig.9 and Fig.11 depict the accuracy among three different models: deep bidirectional LSTM, SVM and LR based on the two different indexes (BIAS and ratio). We can find that the accuracy for the deep bidirectional LSTM is the highest among the three models whatever index we use. Deep bidirectional LSTM can avoid RNN gradient disappearance and gradient explosion in long time series, so it overcomes the shortcomings of averaging the word vector of each sentence in SVM and of losing the ordered information between sentence words and retains the semantic information between words. Obviously, we can see that the accuracy for the LSTM model is much higher when using BIAS.

Fig.10 and Fig.12 describe the accuracy with different data size using deep bidirectional LSTM model based on BIAS and ratio. Through the figures, we can obtain that when using different data size, the accuracy rate is that much different. This also infers that we can get an acceptable result by using a little data as possible.

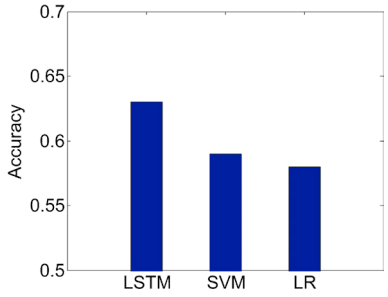

Fig. 9. Accuracy of Prediction Based on Different Models When Taking BIAS as An Indicator.

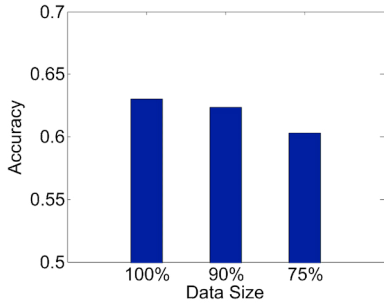

Fig. 10. Accuracy among Data Size Using BIAS

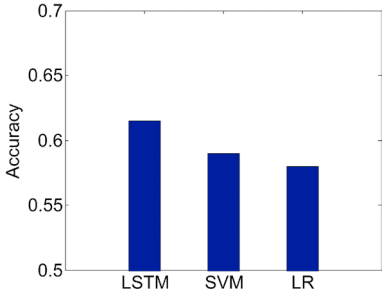

Fig. 11. Accuracy of Prediction Based on Different Models When Taking Ratio as An Indicator.

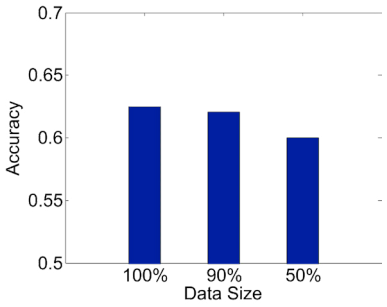

Fig. 12. Accuracy among Data Size Using Ratio

5. CONCLUSION

We focus on how news affects the trend of stock prices over a period of time, rather than the range of stock prices. For this reason, we select BIAS index based on the characteristics of investors' behavior, and build a model to predict the trend of stock price changes on the basis of DBLSTM model. Our model can provide a reference for stock investment, and predict the trend of stock price change in a period of time after news happens, so as to guide investors to make correct investment decisions.

Acknowledgements

The completion of the thesis is attributed to many people's support and encouragement. Without of the help of many people,it could never be completed, thus here we would like to express my sincere gratitude towards them.
First and foremost, we owe our heartfelt thanks to my distinguished and cordial supervisor,Professor Yingjie Tian, who influenced us with his insightful ideas and meaningful inspirations, guided us with practical academic advice and feasible instructions, and enlightened us while we was confused during the writing procedure. His thought-provoking comments and patiently encouragements are indispensable for my accomplishment of this thesis.
Ultimately, thanks go to our parents who are our mentor and guardian from the very beginning in primary school. Without their refined education and care,we could never grow up in such a joyous and cozy environment nor have the courage to confront any obstacles on our way to success.

References

[1]Fama, Eugene F. ″Market efficiency, long-term returns, and behavioral finance.″ Journal of financial economics 49.3 (1998): 283-306.

[2]The Impact of the Structure of Individual Investors on the Chinese Stock Market: the Impact of the Nature on the Trend

[3]Chan, Wesley S. ″Stock price reaction to news and no-news: drift and reversal after headlines.″ Journal of Financial Economics 70.2 (2003): 223-260.

[4] Tetlock,Paul C.″Giving content to investor sentiment: The role of media in the stock market.″ The Journal of finance 62.3 (2007): 1139-1168.

[5] Lee, Dong Wook, and Mark H. Liu. ″Does more information in stock price lead to greater or smaller idiosyncratic return volatility?.″ Journal of Banking and Finance 35.6 (2011): 1563-1580.

[6] Mao, Huina, Scott Counts, and Johan Bollen. ″Predicting financial markets: Comparing survey, news, twitter and search engine data.″ arXiv preprint arXiv:1112.1051 (2011).

[7]Bollen, Johan, Huina Mao, and Xiaojun Zeng. ″Twitter mood predicts the stock market.″ Journal of computational science2.1 (2011): 1-8.

[8]Chen,Wenhao,etal.″Atopic-based sentiment analysis model to predict stock market price movement using Weibo mood.″ Web Intelligence. Vol. 14. No. 4. IOS Press, 2016.

[9]Kumar,Alok,and CharlesMCLee.″Retail investor sentiment and return co movements.″ The Journal of Finance 61.5 (2006): 2451-2486.

[10]Baker,Malcolm,andJeffreyWurgler.″Investorsentimentandthecross- section of stock returns.″ The journal of Finance61.4 (2006): 1645-1680.

[11]Brauer, Greggory A. ″″Investor sentiment″ and the closed-end fund puzzle: A 7 percent solution.″ Journal of Financial Services Research 7.3 (1993): 199-216.

[12]Li,Yun-Fei,and Xiao-FengHui.″Evaluation in dexes selection of stocks′ investment value based on fuzzy clustering.″ 2007 International Confer- ence on Wireless Communications, Networking and Mobile Computing. IEEE, 2007.

[13] Long, Wen, Linqiu Song, and Yingjie Tian. ″A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity.″ Expert Systems with Applications 118 (2019): 411-424.

[14] Nam, KiHwan, and NohYoon Seong. ″Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market.″ Decision Support Systems 117 (2019): 100-112.

[15] Graves, Alex, Santiago Ferna ndez, and Ju rgen Schmidhuber. ”Bidi- rectional LSTM networks for improved phoneme classification and recognition.” International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2005.

[16] Hochreiter, Sepp, and Ju rgen Schmidhuber. ”Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.

[17] Zhou, Peng, et al. ″Attention-based bidirectional long short-term mem- ory networks for relation classification.″ Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.

[18] Srivastava, Nitish, et al. ″Dropout: a simple way to prevent neural networks from overfitting.″ The journal of machine learning research 15.1 (2014): 1929-1958.

[19] Pham, Vu, et al. ″Dropout improves recurrent neural networks for hand- writing recognition.″ 2014 14th International Conference on Frontiers in Handwriting Recognition. IEEE, 2014.

[20] Yahui, Z. H. A. N. G., Difang, W., and Leiming, F. (2012). An ex- perimental study on how media reports affect investment behav- ior based on investor attention. Systems Engineering, 30(10), 19-35.

[21] Le, Quoc, and Tomas Mikolov. ″Distributed representations of sen- tences and documents.″ International conference on machine learn- ing. 2014.

[22] Mikolov, Tomas, et al. ″Efficient estimation of word representa- tions in vector space.″ arXiv preprint arXiv:1301.3781 (2013).