ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 88-98
doi: 10.18178/wcse.2019.06.014

Deep Learning for Stock Market Prediction Using Social Media and Technical Information

Di Wu, Jianhua Cao[+]

School of Computer Science and Technology, Dalian University of Technology, Dalian, China
wudi@dlut.edu.cn, cjh1572318427@mail.dlut.edu.cn

Abstract. In recent years, the stock market has played an increasingly important role and attracted more and more attention. However, the complexity of the stock market makes stock prediction facing a considerable challenge. Many studies found that investor sentiment and stock technical indicators have a secure connection with the stock market movement. Also, in recent studies, deep learning has been widely used in time series forecasting and natural language processing, making it possible to predict stock markets successfully. In this paper, we apply the two-layer bidirectional long short-term Memory networks(Bi-LSTM) model based on glove word embedding and attention mechanism to extract stock sentiment indicators from social media, and use decision tree (DT) and principal component analysis (PCA) integrated model to extract stock technical indicators; then, these indicators apply to the LSTM model to forecast the US stock market movement. The experimental results show that our proposed method can significantly improve the accuracy of the stock market forecast.

Keywords: Stock market; Deep learning; Social media; Time series forecasting; Technical indicators; Attention mechanism; Glove word embedding; LSTM; Bi-LSTM

1. Introduction

The stock market is an essential part of a country's economy, which severely affects individual and national economic development. Successfully predicting the stock market movement can help investors make correct investment strategies to obtain benefits. However, the stock market is a complex and non-linear environment that is influenced by many factors such as the economy, politics, and environment, etc. Therefore, stock market forecasting is seen as one of the most meaningful and challenging tasks. In the early days, many stock prediction methods were proposed. The most common method was using traditional machine learning algorithms for stock price prediction[1]. For example,[2]-[3] used support vector machine (SVM) with good generalization capabilities and fast computing power to predict stock prices. However, early researchers only considered some simple and highly relevant stock indicators, which led to very low prediction accuracy. In fact, the stock market is a dynamic, non-linear, non-stationary, non-parametric chaotic system. The stock market is influenced by many high-relevant factors, including five aspects: (1) economic variables (2) company specific variables (3) factory-specific variables (4) political variables (5) investor psychological variables [4].

Researchers were gradually realizing that the stock market is made up of a large number of stocks rather than any individual stock. How to effectively deal with massive stock data and make predictions has become the focus of research. [1] used 128 stock technical indicators generated by TA-Lib for forecasting; [4] applied dimension reduction algorithm PCA and kernel function-based principal component analysis (KPCA) to reduce 45 stock market indicators. In general, in the era of big data, it is possible to predict the stock

[+] Corresponding author. Tel.: 18742507275
E-mail address: cjh1572318427@mail.dlut.edu.cn

market movement successfully, but using only selection algorithms or reduction algorithms did not significantly improve the prediction accuracy.

Recently, using the sentiment indicators reflected in social media to predict the stock market was a hot research topic. Therefore, many studies used Facebook, Twitter, and some other social platforms to predict the stock market, also, some scholars have extracted related information from stock-related news websites to predict stock movements. for examples,[5] filtered important twits and smart user identification to predict stock price movements with higher accuracy.

Extracting effective stock indicators from social texts has always been a huge technical problem. The emergence of deep learning has greatly improved the accuracy of natural language processing and time series forecasting. The excellent performance of deep learning makes stock market forecasting possible. In the early days, many studies used neural network frameworks to predict the stock market movement, such as [6] proposed an improved bacterial chemotaxis optimization (IBCO), which is then integrated into the back propagation (BP) artificial neural network to develop an efficient forecasting model for prediction of various stock indices. However, due to a large amount of neural network parameters, over-fitting is likely to occur, the prediction accuracy is not stable. With the introduction of dropout and other structures in neural networks, deep learning has once again entered the field of stock prediction research and get great performance.

This work develops a bidirectional long short-term Memory networks based on attention mechanism and glove algorithm (AGBi-LSTM) model to predict trend of related stocks, and extract the sentiment indicators from the predicted results. AGBi-LSTM model leverages the strengths of Bi-LSTM, glove embedding algorithm and attention mechanism, and merges them together. The experimental results show that AGBi-LSTM model performs well when categorizing stock-related social media texts. We not only use the sentiment indicators alone to predict the stock market, but also survey various technical indicators and use DT and PCA integrated model to extract technical indicators. Compared to other models mentioned above, we studied the predictive effect of combining emotional and technical indicators in LSTM networks.

The rest of the paper organizes as follows: The second part introduces the model proposed in this paper, and the third part introduces the results analysis of model experiments. The fourth part gives the summary of the paper and the prospect of the future.

2. Methodlogy

Figure 1 shows the system proposed in our paper. The system's stock indicators come from two aspects: social media texts and stock numeric technical indicators. For the social media text data, using our proposed model: the bidirectional long short-term Memory networks based on attention mechanism and glove algorithm (AGBi-LSTM) to predict the emotional trend of related stocks, and extract the sentiment indicators from the predicted results. For numeric technical indicators, we constructed a DT and PCA integrated model to extract important information. Finally, the extracted stock sentiment indicators and numeric technical indicators are applied to LSTM model to forecast the stock market movement. We apply the system to 10 different company stocks in the United States for the experiment. The detail will introduce in the following section.

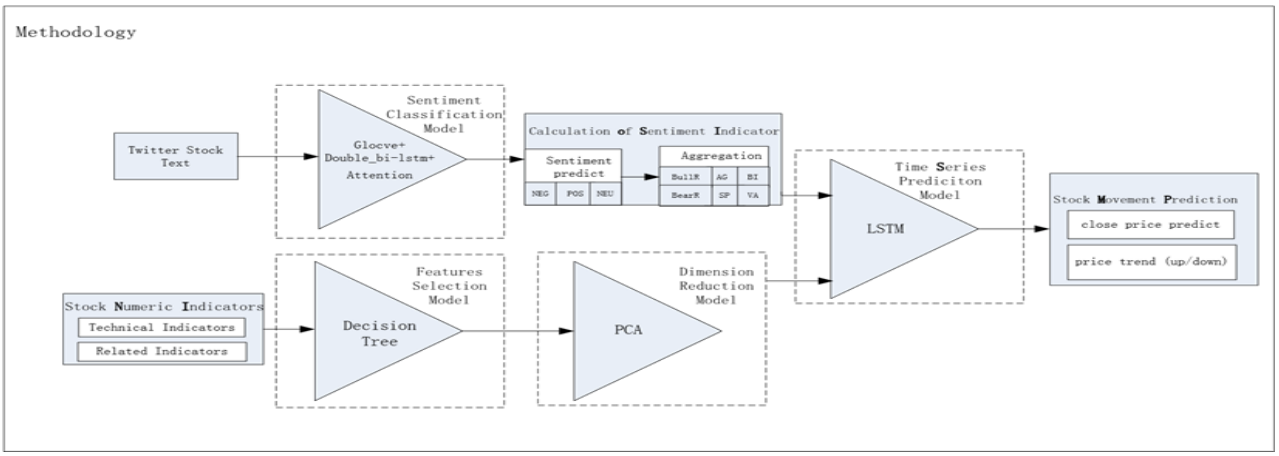


Fig. 1: Methodology

2.1. Social media classification model

Social media generates massive amounts of text every day, reflecting every aspect of personal and national life. Studies found that the price fluctuations in the stock market are positively correlated with the sentiment trends in social media. In this paper, we develop an AGBi-LSTM model to extract stock sentiment indicators, which performs well in classify social media textual sentiment. In order to successfully extracting sentiment indicators from social media textual, We follow the four steps below.

2.1.1 Craw stock related texts from social media platform

In this paper, the social media platform: Twitter, is used as a textual data source for stock sentiment analysis. It is a challenge to craw massive stock related texts from social media platform, but the study found that users often use a hashtag in stock market conversations to refer to related stocks. The hashtag consists of the "$" character and the corresponding stock symbol (e.g., $ AAPL). These hashtags mean that the message texts are related to the stock [7]. So, in this paper, using the program interface provided by twitter get all stock-related blog posts based on hashtags.

2.1.2 Texts clean up and label

Twitter texts have following features :(1) short texts (2) lots of irrelevant content (3) languages are more colloquial, which make classification more difficult. Therefore, it is important to preprocess twitter texts. We clean up irrelevant contents, and the experimental results shown that step can effectively improve prediction accuracy.

2.1.3 Construct sentiment classification model

In this paper, we combined the glove embedding technique, the attention mechanism, and the Bi-LSTM to construct a new model for the emotional classification of social media texts. Compared with [8], the accuracy rate has been improved.

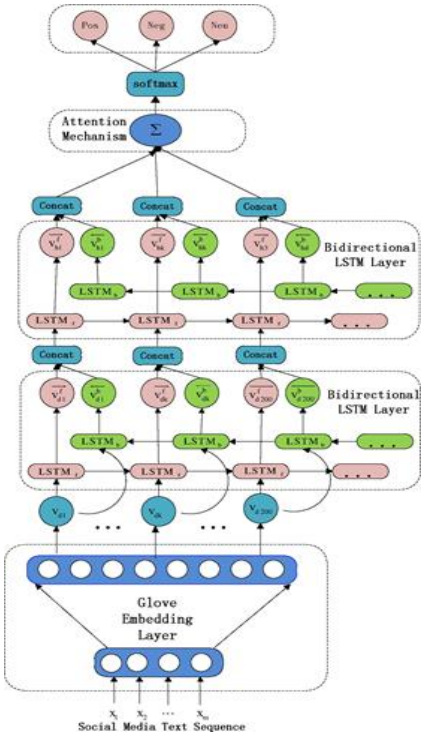


Fig. 2: Social medial sentiment predicted model

The structure of social media texts sentiment classification model is shown as Figure 2. As the first layer of AGBi-LSTM model, Glove embedding algorithm was proposed by [9]. The word vector calculated by Glove embedding algorithm can represent the relationship between words and words well. It has been widely used in natural language tasks. The Glove embedding algorithm has the characteristics of fast training and excellent scalability. The loss function of the Glove embedding algorithm expresses as:

$$J=\sum_{ik} f(X_{ik})(w_i^T w_k + b_i + b_k - \log X_{ik})^2 \quad (1)$$

where X_{ik} represents the total number of occurrences of the text k in the text i context; f represents the weight function used to measure the effect of two words, and the function formula is as (2). w_i , w_j represents the word vector. In this paper, Glove embedding layer outputs 200 dimensions to represent each word; b_i , b_k are bias terms introduced to solve symmetry.

$$f(x)=\begin{cases}(\frac{x}{x_{\max}})^{\alpha}, & if\ x<x_{\max}\\ 1\ , & otherwise\end{cases} \tag{2}$$

The quality of the word vector generated from Glove embedding layer depends on whether Glove embedding model was been well trained. Generally, well-trained glove model needs vast volume of corpus as training set, it is time consuming and difficult to achieve. Therefore, in this paper, we use the trained 200-dimensional word vector provided by Stanford directly. For a sequence of fixed length m, after processing by the glove embed layer, it transforms into a matrix of M*200 dimensions. In AGBi-LSTM model, Bi-LSTM layer will use this word embedding matrix as input.

As the second and third layers of our model, Bi-LSTM is introduced into our paper. Compared with Unidirectional LSTM, Bi-LSTM learns both past and future information, but unidirectional LSTM only preserves information of the past. Bi-LSTM shows outstanding results in natural language applications and can better understand the context. Our experiment results also shown the power of Bi-LSTM structure.

LSTM as the core of Bi-LSTM, is a kind of RNNS structures. Compared with standard RNNS, LSTM can reduce long delays between inputs and solve vanishing gradient problem, which causes the RNNS to encounter trouble when memorizing past input values after more than 10 timesteps. LSTM is ideal for tasks such as time series and natural language. Therefore, in our paper, the stock text sentiment analysis and the stock market forecast both use the LSTM structure.

As the fifth layers of our model, attention mechanism is introduced into our paper. Before attention mechanism, NLP relies on reading a complete sentence and compress all information into a fixed-length vector, as you can image, a sentence with hundreds of words represented by several words will surely lead to information loss, inadequate translation, etc. However, attention partially fixes this problem [10-12]. It allows machine to look over all the information the original sentence holds, then generate the proper result according to current word it works on and the context.

2.1.4 Extract Stock sentiment indicators

The well-trained AGBi-LSTM model can predict the daily stock market sentiment trend in real time. According to the predicted emotional classification, we created six sentiment indicators that represent the stock market, as shown in the following formulas:

$$BullR_t=\frac{M_t^{pos}}{M_t^{pos}+M_t^{pos}} \tag{3}$$

$$BearR_t=\frac{M_t^{neg}}{M_t^{pos}+M_t^{pos}} \tag{4}$$

$$BI_t=\ln\frac{M_t^{pos}+1}{M_t^{neg}+1} \tag{5}$$

$$VA_t=BullR_t-BullR_{t-1} \tag{6}$$

$$AG_t=1-\sqrt{1-\left(\frac{M_t^{pos}-M_t^{neg}}{M_t^{pos}+M_t^{neg}}\right)} \tag{7}$$

$$SP_t=\frac{M_t^{pos}-M_t^{neg}}{M_t^{pos}+M_t^{neg}} \tag{8}$$

where M_t^{pos} and M_t^{neg} represent the number of tweets that have a positive and negative attitude towards the stock on day t. In our study, the above sentiment indicators came from other papers, indicators (3)-(7) came from [13], and these indicators have a specific effect on forecasting stock returns, trading volume, and

volatility. The indicator (8) came from [14], which combined with basic technical indicators to improve the accuracy of the stock's movement.

2.2. Stock technical indicators model

In addition to sentiment indicators affecting stock prediction, many technical indicators also influence the accuracy of stock prediction. In this section, we will introduce the technical indicators included and establish a technical indicators dimension reduction model.

In this paper, technical indicators have two types. The first type is internal technical indicators. These technical indicators are all calculated from basic stock indicators (open, close, low, high, volume). A total of 55 internal indicators were introduced and some of them from [4]. The second type is external technical indicators, which indirectly affect the stock market movement. In this paper, 27 external indicators were introduced. Some of them came [14]. Some have been considered for the first time, such as the Employment rate, Producer Price Index for All Commodities. External technical indicators are influenced by the country's politics, economy and other aspects, so they also affect the changes in the stock market.

As described above, except for the six stock sentiment indicators extracted from social media, there are 82 technical indicators of stocks. These indicators obey not only different probability distributions but also have different influences on forecasting stock market changes, so we construct a technical indicators dimension reduction model to improve the accuracy of stock market forecasting.

As shown in Figure 1, using DT algorithm selects the most relevant technical indicators at first. Then, using PCA algorithm reduces the dimension of selected technical indicators. The experimental results show that the DT and PCA integrated model can significantly improve the accuracy of stock prediction.

In this paper, the DT algorithm was used to select relevant stock technical indicators. Feature selection in the decision tree construction process is a crucial step, because this step decides which features will be used to divide the feature space and select useful features. Chosen features can improve the learning efficiency of the decision tree. Indicators selected by DT algorithm will use as inputs of PCA algorithm and aim to improve stock prediction accuracy. In machine learning, there are many linear or nonlinear algorithms for data reduction. Among these algorithms, PCA is the most widely used unsupervised linear algorithm [4]. The idea of PCA is to map n-dimensional features to k-dimensional, and this k-dimensional vector is orthogonal and is called the main element. PCA aims to reduce the dimensions of data without losing data implied information.

3. Data

In this paper, sentiment and technical indicators are used as data sources to predict the stock market. Therefore, how to collect sentiment and technical indicators data becomes the key. As for sentiment indicators, they came from social media texts, and social media texts come from social media platform with more than 3 million users: Twitter. As for technical indicators, some are directly from Yahoo! Finance website and https://fred.stlouisfed.org website, and some are calculated indirectly according to economic formulas. Below we will explain these data sources in detail from the sentiment indicators and technical indicators parts.

3.1. Sentiment indicators data

As mentioned above, sentiment indicators were extracted indirectly from social media texts through our proposed sentiment classification model. In this paper, we use Twitter's open interface for programming and craw twitter posts through hashtag '$'. In order to generalize in real-world prediction, we craw all twitter stock posts from 2014-1-1 to 2017-12-31.

In the experiment, labeled texts are divided to three sentiment classes: positive, negative and neutral. In order to ensure the balance of classification model parameters tuning, making sure that the amount of texts of each type is similar in manual labeling is vital.

Labeled social media texts will be used to extract six sentiment indicators (See detail in section 2.1). In order to ensure the effectiveness and reliable of these emotional indicators, there must be an amount of

stock-related twitters per day. In this paper, each stock has 500 tweets per day. After well-trained AGBi-LSTM model labeled, sentiment indicators will be generated.

Since we downloaded all social media texts from 2014-1-1 to 2017-12-31 for AGBi-LSTM model training and real-world classification generalization, in order to facilitate statistical results, only chosen 10 stocks in the stock prediction section.

3.2. Stock technical indicators data

Technical indicators data can be divided into internal and external. Internal indicators are indicators that have been calculated by experts over the years and have a significant impact on stock forecasts. For examples, Moving Average Convergence Divergence (MACD) and Simple Moving Average (SMA) have been proved by many papers that have a significant effect on the prediction of the stock market. In this paper, a total of 55 internal indicators were introduced. External indicators collect indicators that affect stock prediction indirectly, which mainly came from https://fred.stlouisfed.org website. According to many studies, external indicators also have a significant role in stock forecasting, such as the Employment rate, Producer Price Index for All Commodities. the result why external technical indicators are so important, mainly because them are influenced by the country's politics, economy and other aspects, and indirectly affect the changes in the stock market.

4. Results

4.1. Social media text classification

As described in 3.1 section, in order to generalize in real-world prediction, we randomly select some twitter posts for labeling, and made sure that each sentiment category is similar in number. In this experiment, 12670 Twitter posts are labeled, 4753 in the positive category, 3703 in the negative category, and 4215 in the neutral category. So, there is no need to consider the problem of data imbalance and make it easy to tune AGBi-LSTM model parameters.

In order to train AGBi-LSTM model and evaluate the performance of the model, we randomly select 70% labeled twitter texts as the training set and 30% as the test set. To evaluate the performance of the model, evaluation metrics, combing by precision, recall, accuracy and F-measure are selected.

Table 1: Social media text prediction

Method	Precision	Recall	F	Accuracy
Bi-LSTM	0.7174	0.6576	0.6857	0.6923
CNN+Bi-LSTM	0.7258	0.6414	0.6803	0.6892
glove+Bi-LSTM	0.7419	0.6753	0.7065	0.7168
AGBi-LSTM	0.7659	0.7282	0.7463	0.75

AGBi-LSTM model training results are shown in Table 1. Compared with other baselines in Table 1, proposed model AGBi-LSTM performs best in test set (with a precision of 0.7659, a recall of 0.7282, a F-score of 0.7663, an accuracy of 0.75). Compared with model Bi-LSTM and glove+Bi-LSTM, classification model AGBi-LSTM fully demonstrates the power of glove algorithm and attention mechanism in natural language processing; the performance of model CNN+Ni-LSTM is 7% lower than the model AGBi-LSTM, the may reason is that social media texts are shorter and more irregular, and CNN structure performs better in long processing texts.

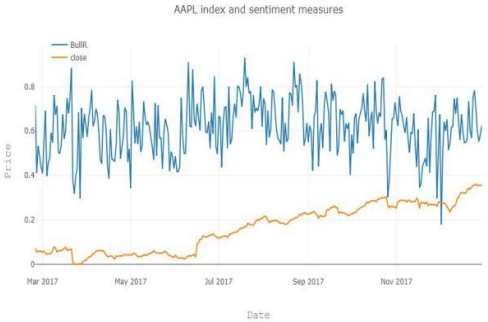


Fig. 3: AAPL stock price and sentiment measures

Well-trained AGBi-LSTM model is used to predict sentiment types of massive stock twitter texts, and indirectly extract six sentiment indicators of stocks daily (See 2.1 section for details). As shown in Figure 3, it is easy to see that the daily close price of AAPL stock has a strong correlation with the indicator BullR. At a period after July 2017, the stock price showed a significant upward trend. At this time, the corresponding BullR sentiment indicator was substantially higher than 50% during this period.

4.2. Close price prediction

Close price prediction is critical for investors to make investment strategies. Therefore, this part of the experiment mainly studies the effects of sentiment indicators and technical indicators on the close price prediction in the LSTM structure.

As mentioned above, there are a total of 82 technical indicators, these indicators obey not only different probability distributions but also have different influences on forecasting stock market, so proposed dimension reduction model at section 2.2 will be used to select most significant translated technical indicators. In this part of experiment, a total of 22 translated technical indicators are selected, and selected translated technical indicators will be used as input of the LSTM structure to predict close price. To evaluate the performance of the model, choosing statistical methods RMSE, MAE, MAPE, MSE and R estimate the model results.

Table 2: Prediction performance for all stocks

Method	Average performance measure in test data						Rank
	RMSE	MAE	MAPE (%)	R	R^2	MSE	
LSTM	0.407	0.259	1.394	0.509	0.259	0.165	1
Ridge	0.505	0.353	1.903	0.594	0.353	0.255	2
KNeighbors	1.67	1.204	6.273	1.097	1.204	2.79	3
DecisionTree	3.109	2.679	15.745	1.637	2.679	9.669	4
SVM	3.439	3.312	18.761	1.820	3.312	11.827	5

Fig. 4: Samples of close price prediction

Table 2 and Figure 4 show experimental results. Among all results, LSTM model performs best for the close price prediction (with an RMSE of 0.407, MAE of 0.257, MAPE of 1.397, R of 0.509, R2 of 0.259, and MSE of 0.165). The above results clearly reveal that the close price accuracy of all stocks predicted by the LSTM model is significantly higher than baselines. Over all, proposed dimension reduction model and LSTM merged method yielded better prediction for stock dataset, which also sufficiently proves the validity and robustness of LSTM for predicting time series tasks. In future work, LSTM can be used as a research model for long-term investment strategy of stocks.

To verify the strengths of the LSTM model in time series prediction, we introduced the Ridge, KNeighbors and Decision Tree algorithms as baselines. In order to facilitate the statistical results of the test data, we selected 10 different fields stocks for experiment. The 10 stocks include AMZN, AAPL, FB, GOOG, MSFT, NFLX, QQQ, SPY, TWTR and TSLA. In addition, the rest of experiment, those 10 stocks are also used as experimental subject. In addition to above, Since the LSTM model needs to tune many parameters, to

ensure the comparability of the entire experiment process, the fixed LSTM parameters need to be set in the entire experiment process. We set the timesteps of LSTM to 22, the output unit to 128, and the dropout to 0.2.

For all baseline methods and LSTM model, experiment use 1461 trading day data from 2014-1-1 to 2017-12-31 as original data, and select 70% of the previous contiguous part as training set and the rest as test set.

4.3. Stock trend prediction

In the stock market, it is vital to predict the trend of stock price movement. It can provide users with suggestions for buying or selling. The prediction results of the above experiment can predict stock price fluctuates within the actual price range, but it cannot accurately provide short-term users with suggestions for buying or selling. Therefore, in this section, making predictions about stock price movement trend and giving the users strategies for buying or selling become the main task.

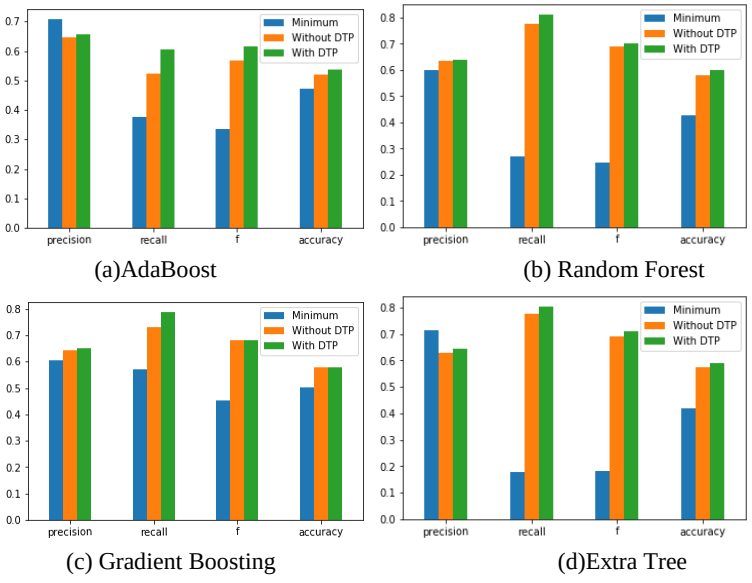
As many studies defined, buying and selling signal was defined as formula (9). when the close price $close_t$ at day t is greater than close price $close_{t-1}$ at day $t-1$, it indicates that the stock price rises at day t. Users can choose to buy to get more profit at day t. On the contrary, when the stock price drops, the user can choose to sell to reduce losses.

$$y = \begin{cases} 1, & close_t - close_{t-1} >= 0 \\ 0, & others \end{cases} \tag{9}$$

In this part of the experiment, it is necessary to prove that the validity and robustness of technical indicators extraction model and sentiment indicators for predicting stock trends. Therefore, we set up several sets of inputs for comparison.

$$\begin{aligned} I_1 &= \{M_{t-5}, M_{t-4}, \cdots, M_t\} \\ I_2 &= \{T_{t-5}, T_{t-4}, \cdots, T_t\} \\ I_3 &= \{M_{t-5}, M_{t-4}, \cdots, M_t, I_{t-1}, I_{t-4}, \\ &\cdots, I_t, O_{t-5}, O_{t-4}, \cdots, O_t, T_{t-5}, T_{t-4}, \cdots, T_t\}\} \\ I_4 &= \{S_{t-5}, S_{t-4}, \cdots S_t, T_{t-5}, T_{t-4}, \cdots, T_t\} \end{aligned}$$

As indicated by the formatting formula above, represents the basic technical indicators (See detail in section 3.2); represents internal indicators, 55 in total; represents external indicators, 27 in total; represents the sentiment indicators, 6 in total; T represents the indicators generated by DT and PCA integrated model, 22 in total. In the experiment, s generated by following process: Firstly, Decision Tree model select 35 technical indicators whose importance is higher than 0.01. Then, applying these indicators to the PCA model selects indicators with cumulative sums greater than 0.99, totaling 22.



(a)AdaBoost (b) Random Forest

(c) Gradient Boosting (d)Extra Tree

Fig. 5: Different methods to verify the true and predicted values of several sample stock.

In the experiment of stock trend forecasting, in addition to studying the effectiveness of sentiment indicators and DT and PCA merged model, it is also necessary to study the strengths of LSTM structure on stock trend prediction. Therefore, in this section, Extra Trees, RF(Random Fores)t, AdaBoost, and GB(Gradient Boosting) algorithms are constructed as baselines.

Table 3: Stock movement prediction

Method	Input	Precision	Recall	F	Accuracy
ExtraTrees	I_1	0.712	0.179	0.182	0.420
RF	I_1	0.599	0.270	0.248	0.429
AdaBoost	I_1	0.708	0.377	0.334	0.470
GB	I_1	0.606	0.572	0.453	0.503
LSTM	I_1	*	*	*	*
ExtraTrees	I_2	0.644	0.745	0.690	0.58
RF	I_2	0.6356	0.741	0.683	0.575
AdaBoost	I_2	0.637	0.706	0.669	0.567
GB	I_2	0.640	0.778	0.702	0.589
LSTM	I_2	0.652	0.896	0.748	0.634
ExtraTrees	I_3	0.629	0.777	0.692	0.575
RF	I_3	0.634	0.776	0.691	0.578
AdaBoost	I_3	0.646	0.524	0.569	0.518
GB	I_3	0.6410	0.731	0.680	0.5773
LSTM	I_3	*	*	*	*
ExtraTrees	I_4	0.642	0.804	0.711	0.590
RF	I_4	0.640	0.812	0.700	0.599
AdaBoost	I_4	0.655	0.606	0.617	0.538
GB	I_4	0.649	0.787	0.682	0.579
LSTM	I_4	*	*	*	*

Figure 5 shows the results of different inputs using different models. Where “Minimum” means input . “Without DTP”means input and “With DTP” means input . From the experimental results, we can see that the indicators introduced in this paper can greatly improve the accuracy of stock trend forecasting.

There also has an interesting finding, as shown in Table 3, it is clear that the best prediction result is to use the sentiment indicators as the input LSTM model, with 0.89623 recall, 0.74795 f-values, and 0.63368 accuracy. This result fully illustrates the importance of sentiment indicators and LSTM model for forecasting the time series stock market, but for other inputs, the LSTM model shows the convergence of the training set and the test set fails to converge, so we did not give numerical values in the table. The main reasons for this situation are as follows: (1) The data set used in this experiment is relatively small, with only 1461 transaction days. (2) The experiment uses the same simple LSTM model. All the parameters are the same, and not consider different inputs and different probability distributions. (3) The sentiment indicators can successfully apply to the LSTM model, but the simple combination of the sentiment indicators and the technical indicators used as an input fails. It indicates that the simple integration of emotional indicators and technical indicators is not predictive in the LSTM model. Therefore, in the future work, we will study how to integrate the emotional indicators with technical indicators to be applied to the LSTM model.

5. Conclusion

Predicting stock market trends and close price are very complicated since many factors can affect stock market. this work presents a novel approach, based on deep learning structure, sentiment indicators and technical indicators, to constructing a stock forecasting expert system, with the aim of improving forecasting accuracy. The system generates indicators that affect stock forecasts in all aspects, making prediction results more reasonable and precise. The advantages of deep learning structure in time series task prediction and natural language processing are also fully reflected.

In this paper, we study indicators that affect stock forecasting in all aspects, mainly divided into sentiment indicators and technical indicators. Sentiment indicators are extracted from labeled social media texts daily, and labeled social media texts generated from proposed classification model AGBi-LSTM. AGBi-LSTM model is based on Bi-LSTM structure, glove algorithm and attention mechanism, which performs well in social media texts classification task because of metaheuristic optimization characteristic. For technical indicators, we investigate technical indicators that directly and indirectly affect the stock market and derive the technical indicators using model integrated with DT and PCA. Finally, the sentiment indicators and technical indicators are combined as input and used to predict the close price and movement trend of stocks in the LSTM model.

To evaluate the proposed model, we applied the model to 10 different US company stocks. The experimental results were evaluated using statistical methods, and introduce various algorithms as evaluation baselines for our model. Social media text sentiment prediction achieved a 75% accuracy rate, and the stock market price forecast reached a low SME with an average value of 0.165. The bullish and bearish trend of the stock foretasted f-value reached 0.74795.

This paper focuses on the US stock market. To promote the proposed system, we will apply the method to other stock markets in the future. Besides, the paper's experiment used only 1461 trading days, and more experimental data may perform better for the deep learning model LSTM. Finally, the experimental results of this system show that the simple combination of technical indicators and sentiment indicators cannot be successfully predicted in the LSTM model. In the future, we need to focus on this problem.

6. References

[1] Nelson, D. M., Pereira, A. C., & de Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 1419-1426). IEEE.

[2] Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319.

[3] ZHANG, C. X., ZHANG, Y. P., ZHANG, Y. C., CHEN, J., & WAN, Z. (2006). Stock Prediction Based on Support Vector Machine [J]. Computer Technology and Development, 6.

[4] Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications, 67, 126-139.

[5] Coyne, S., Madiraju, P., & Coelho, J. (2017, November). Forecasting Stock Prices Using Social Media Analysis. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) (pp. 1031-1038). IEEE.

[6] Zhang, Y., & Wu, L. (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. Expert systems with applications, 36(5), 8849-8854.

[7] Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73, 125-144.

[8] Vargas M R, Lima B S L P D, Evsukoff A G. Deep learning for stock market prediction from financial news articles[C]// IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications. IEEE, 2017:60-65.

[9] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[10] Cho, K., Courville, A., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. IEEE Transactions on Multimedia,17(11), 1875-1886.

[11] Show, A. (2015). Tell: Neural image caption generation with visual attention. Kelvin Xu et. al.. arXiv Pre-Print,23.

[12] Raffel, C., & Ellis, D. P. (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756.

[13] Si J, Mukherjee A, Liu B, et al. Exploiting topic based twitter sentiment for stock prediction[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013, 2: 24-29.

[14] Gken M, zalc M, Boru A, et al. Integrating metaheuristics and artificial neural networks for improved stock price prediction[J]. Expert Systems with Applications, 2016, 44: 320-331.