
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
 Volume 25, Issues 3, pp. 2062–2075, DOI 10.12694/scpe.v25i3.2717

ARIMA-BPNN BASED STOCK PRICE PREDICTION MODEL BASED ON FUSION NEWS SENTIMENT ANALYSIS

XIAOZHE GONG*

Abstract. In recent years, the prediction of stocks has mainly focused on improving and combining stock prediction algorithms, or analyzing news sentiment tendencies to simulate subjective investor consciousness. However, both methods have shortcomings in practicality and comprehensiveness. Therefore, based on the use of stock data, the sentiment propensity of vocabulary in the article was processed, and a new algorithm model was obtained by combining the differential integration moving average autoregressive model and backpropagation feedforward neural network model. Finally, sentiment propensity was integrated into the combination model to obtain an algorithm model that integrates sentiment analysis. After optimizing the sentiment vocabulary of news. The algorithm has improved its ability to recognize emotional tendency words, while traditional algorithms have been improved to improve the accuracy of stock prediction, further verifying the relationship curve between emotional tendency and stock prediction fluctuations. The experimental results show that the combined model of sentiment analysis is close to the true value in predicting stock results, with an error of less than 1.5%. The accuracy and stability of the model's prediction results are significantly better than the uncombined model and traditional prediction models. The new combination model provides better judgment basis for investors through experimental prediction results, creating conditions for investors to avoid stock market risks and improve investment value.

Key words: Stock prediction; News information; Emotional tendencies; Combination model; accuracy

1. Introduction. As the country's economic strength improves, stock investment has gradually penetrated into ordinary people's hearts. But the variability and uncontrollability of stock market have left many investors at a loss. Therefore, timely prediction of stocks is an important part of achieving control over stock market. At present, the most commonly used methods for stock analysis are basic analysis and technical analysis. The basic analysis method is to use one's own knowledge of economics, finance, and other aspects to analyze stock market through understanding. However, the relative basic analysis is influenced by stock market fluctuations, resulting in a significant deviation in analysis accuracy [2]. Technical analysis is mainly conducted through methods such as historical data and mathematical analysis, which can achieve most predicted results. However, technical analysis also has significant technical limitations and cannot predict stock data from multiple dimensions. Therefore, on the basis of traditional analysis methods, the sentiment analysis section is designed, which combines the differential integration moving average autoregressive model and backpropagation feedforward neural network model to obtain a new combination algorithm model for sentiment analysis fusion, thereby improving the accuracy of stock prediction by the network model. At the same time, the neural network model after sentiment analysis is fused, can further verify the direct relationship between stock market changes and emotional tendencies [4]. There are many current studies on stock prediction, and most of them only improve the prediction model or perform sentiment analysis prediction from the perspective of sentiment tendency, and very few studies that combine the two exist. For this reason, this study adds the new concept of positional weights and punctuation weights, and integrates the sentiment value obtained from sentiment analysis as a feature into the combination model, realising a prediction method that combines numbers and words. This method not only improves the existing theory, but also proposes new directions and research ideas for the research in this field. This study consists of four parts. Firstly, it mainly introduces the research achievements of various experts and scholars at home and abroad. Secondly, the method and structure of building the entire model were introduced. Next, comparative experiments were conducted on the accuracy and feasibility of the model using stock data. Finally, there is a summary of this study and prospects for future research directions.

*Advertising Institute, Communication University of China, Beijing, 100024, China (Corresponding author, gongxiaozhe1996@126.com)

Table 1.1: Comparative Analysis of Related Studies

Author Name	Research Title	Research Areas	References
Somesh Yadav et al.	Stock price forecasting and news sentiment analysis model using artificial neural network	Data network, finance, news analysis	References [5]
Zitnik, Slavko et al.	Target-level sentiment analysis for news articles	News sentiment, Internet of Things	References [6]
Mohan B R et al.	Hybrid ARIMA-deep belief network model using PSO for stock price prediction	Internet of Things, Computers, Data Forecasting	References [7]
Colasanto F et al.	AlBERTino for stock price prediction: a Gibbs sampling approach	Computers, Internet of Things, Finance	References [8]
Vara P V et al.	Sree R M. Sruthi, Nishanthi, K.Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis	Statistics, Computers, Stocks	References [9]
Chen Y et al.	Stock Price Forecast Based on CNN-BiLSTM-ECA Model	Statistics, Computers, Stocks	References [10]
Shapiro A H et al.	Measuring news sentiment	Statistics, stocks, news	References [11]

2. Related works. The ever-changing stock market has attracted many investors, and the prediction of stock price fluctuations has also deeply attracted domestic and foreign scholars to explore and research in this field. Somesh Yadav et al. used artificial neural networks (ANNs) to predict stock prices, and studied the impact of past stock trends, daily opening prices, and news sentiment on investors' investment directions. They proposed a model based on ANNs. Artificial neural networks for predicting stock prices is feasible and effective [5]. Zitnik, Slavko, and others created a new article dataset in their research on news and social media information. The dataset was analyzed and evaluated using deep NN algorithms and machine deep learning algorithms. The new dataset performed better than traditional datasets and had significantly higher accuracy [6]. After analyzing the operational planning and sales of enterprises, Shaikh Sahil Ahmed et al. found that the accuracy of stock prediction (SP) for listed companies is an important factor. Therefore, to improve SP accuracy on the foundation of deep NN, the sequence model was enhanced and particle swarm optimization algorithm was used. The improved new model outperforms other models in terms of market prediction accuracy [7]. Colasanto, Francesco and others believe that traditional encoders have some uneven classification problems when classifying language emotions and words. Therefore, a new emotion classification framework is proposed to join the traditional encoder. The new encoder can use emotional orientation to improve SP, and use Monte Carlo method to generate a series of feasible paths. The new encoder can solve classification inequality and SP [8].

Prasad, Venkata Vara et al. compared traditional SP methods and found that among three main algorithms currently used. Kalman filter can consider market fluctuations and greatly improve the accuracy of predictions. XGBoost can collect data and provide a dataset, effectively capturing time series. The main function of Autoregressive Integrated Moving Average Model (ARIMA) is to eliminate stationarity and improve prediction performance. In combination with the advantages of three algorithms, a new algorithm Mixture model is proposed. The new hybrid algorithm's performance is significantly better than other three algorithms, and its clarity is higher [9]. Chen et al. found in their study that SP is a time series problem. However, due to the instability and variability of SP, successfully predicting stock prices has become a highly challenging issue. Therefore, to improve prediction accuracy, they proposed a new model based on convolutional NN and long short-term memory network. The new model utilizes NN characteristics to reduce noise impact in SP. This new model is significantly superior to other traditional models in predicting stock prices [10]. Shapiro A H et al. believe that a new sentiment analysis method has been discovered for stock analysis and prediction. By using this method, some stock data can be predicted to improve the accuracy of stock prediction. The experimental

results indicate that this model has better accuracy than existing models and can predict news emotions [11].

In summary, in many experts and scholars' research, SP has low accuracy and precision in prediction algorithms due to its nonlinearity and high volatility. Many traditional algorithms currently in use can only solve a single problem. Therefore, combining multiple NN algorithms on top of traditional algorithms can greatly improve SP accuracy. This experiment uses a combination of ARIMA algorithm and Back Propagation (BP) algorithm to improve the algorithm using sentiment analysis methods, resulting in improved accuracy and prediction accuracy.

3. Method of Stock Price Prediction Based on ARIMA-BPNN Fusion News Emotional Tendency Analysis. This chapter mainly introduces the theoretical methods used, such as web crawler technology, data pre-processing technology, ARIMA model and BP model, to classify and analyze the news words with news emotional tendency. The experiment introduced how to combine multiple methods and apply them to stock price prediction.

3.1. Introduction to the Method of Combining Models Based on Fusion News Emotional Tendency Analysis. Web crawler is a means that can freely grab information data from the World Wide Web, and is a commonly used means to grab network information [12]. Network crawlers first apply for data crawling on the network, then analyze the captured data to extract useful data information, and finally store the obtained data to achieve the purpose of data crawling. When crawling news information, web crawler technology is an indispensable means. Data preprocessing technology is a means of processing and analyzing the obtained data, usually used in large datasets, mainly including filtering and cleaning the data, and deleting pauses in sentences. Since the current study is really a study conducted by investors, it is unrealistic for investors to access many aspects such as weather, economy, company data and many other factors, so in the study only the time series related data predictive analytical model ARIMA model is selected for analysis and prediction. ARIMA is a model that can perform differential classification processing on autoregressive models, with the core content of establishing, testing, and validating methods. By utilizing data that can be crawled, a random sequence is formed and arranged in chronological order. Equation 3.1 is its main parameter.

$$[p, d, q, AR(P), MA(q)] \tag{3.1}$$

In equation 3.1, $AR(P)$ represents the autoregressive model, and $MA(q)$ represents the moving average model. p represents autoregression order, d represents the difference number during data processing, and q represents the moving average order. Equation 3.2 is an explanation of [13].

$$x_t = \sum_{i=1}^{p} \mu_i x_{t-i} + \varepsilon_t \tag{3.2}$$

In equation 3.2, μ_i represents the autoregressive coefficient, ε_t represents noise sequence, and x_t represents the autoregressive process at p order. Equation 3.3 represents the moving average model.

$$x_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{3.3}$$

In equation 3.3, x_t represents that the interference at different periods is random, θ_i represents the autoregressive coefficient, and represents the noise sequence. When ARIMA model's time series is stationary, then equation 3.4 is its expression formula.

$$x_t = \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{i=1}^{p} \mu_i x_{t-i} + \varepsilon_t \tag{3.4}$$

In equation 3.4, when $q = 0$, its model is represented as an autoregressive model. When $p = 0$, its model is represented as a moving average model. ARIMA model does not need to choose time development due to its own time series. Moreover, its plasticity is strong and can be modified until the model is satisfactory [15].

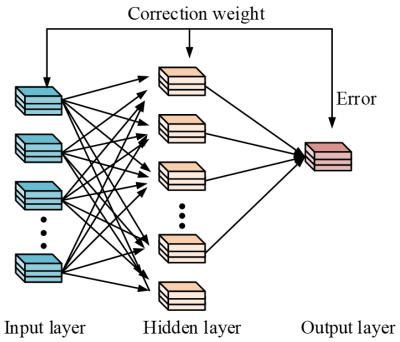


Fig. 3.1: BP Neural Network Backpropagation

Table 3.1: News Information Retrieval Content

/	1	2	3	4	5
Content captured	Financial news broadcast	Individual stock information review	Headline	News hour	Content

And this model can also compare past and present time data, thus possessing better data accuracy. However, the same ARIMA model can only deal with the original data of stable topics, and cannot analyze the data set of nonlinear relations. BPNN is currently a widely used NN, mainly composed of input layer, output layer, and hidden layer. When trained through the algorithm, it mainly propagates through forward and backward in Figure 3.1.

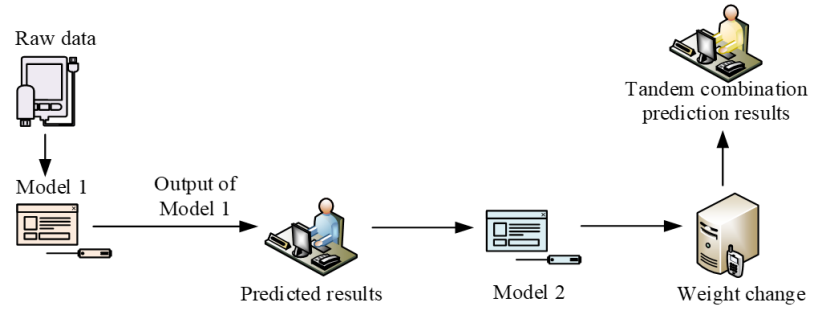
In Figure 3.1, BPNN's backpropagation process is mainly to compare the obtained data results with the real values. And activation function is used to compare the neuron parameters. The neurons gradient is reduced by using loss function, and then the neurons are updated [3]. BPNN is a mapping network for parameter data, with excellent processing ability for nonlinear relationships. BPNN can also process and classify data in many different states. At the same time, it also has the ability to apply the resulting data to another new knowledge. However, due to BPNN structure's large size, there is a tendency for non-convergence. Moreover, when the weight value is too large, BPNN will fall into local extremum, leading to training failure. A combination prediction model is a model that combines two or more models, including series combination and parallel combination. Figure 3.2 shows the series combination and parallel combination diagrams.

In Figure 3.2, the series combination is to transfer the previous model's results to the next model, and the calculated results' transfer is used from the next model to the previous one to calculate the model weight and obtain the results. The parallel combination model predicts data values through multiple models and compares them with actual values. The error is larger, the weight is smaller [4].

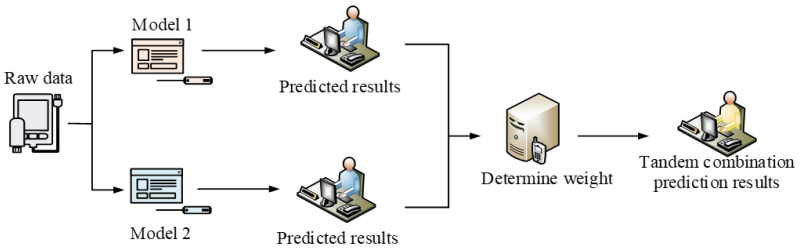
3.2. Emotional Analysis Based on News Text Information. The stock market volatility is not only related to the listed company's operation, but also influenced by domestic and foreign policies, news, and other factors. The research on news text information's emotional analysis will greatly improve the model accuracy and stability. Table 3.1 shows the content crawled through web crawlers in news and finance networks.

In Table 3.1, the latest financial content is obtained through the broadcast of financial news content. The current individual stocks' latest data information is obtained through the individual stock's comments. The data is filtered by Headline. And the feelings are analyzed by news time and content. When preprocessing the algorithm, it is necessary to annotate the vocabulary in news and filter out useful information. Some data belongs to junk data, such as advertising, copyright, personalized information, etc., which needs to be filtered.

When analyzing emotions in news information texts, it is necessary to analyze and construct the quality of each individual word [18]. Such as basic emotional vocabulary, financial direction vocabulary, modified vocabulary, etc. Basic emotive words include some daily vocabulary with positive and negative emotions.



(a) Series Structure Combination



(b) Combination of Parallel Results

Fig. 3.2: Two Combination Structures

Financial vocabulary is mainly commonly used in financial field, and modifying vocabulary includes some degree adverbs, negative words, affirmative words, turning words, and so on. When analyzing emotional vocabulary, it is necessary to divide some paragraphs in news, compare and match them with vocabulary in dictionary, and finally determine emotional inclination degree through vocabulary use in equation 3.5.

$$T = \frac{p - N}{p + N} \tag{3.5}$$

In equation 3.5, p represents the positive emotional vocabulary frequency, N represents the negative emotional vocabulary frequency, and T represents the emotional value brought by emotional vocabulary. But traditional calculation formulas cannot calculate article expression's words emotional weight. Therefore, on the basis of the basic formula algorithm, add vocabulary's basic weight value itself, vocabulary location information's weight value, and punctuation position's weight value. By increasing weight, each news vocabulary's emotional orientation can be calculated, and then the article's order and paragraphs can be quantified through emotional orientation. Equation 3.6 is the emotional paragraphs tendency.

$$\text{para} = \frac{\sum_{i=1}^{M} \text{Sent}_i}{M} \tag{3.6}$$

In equation 3.6, M represents the total sentences with emotional tendencies, and Sent_i represents sentences' emotional tendency values. Equation 3.7 is the average emotional tendency value.

$$\text{Con} = \frac{\sum_{i=1}^{Z} \text{para}_i}{Z} \tag{3.7}$$

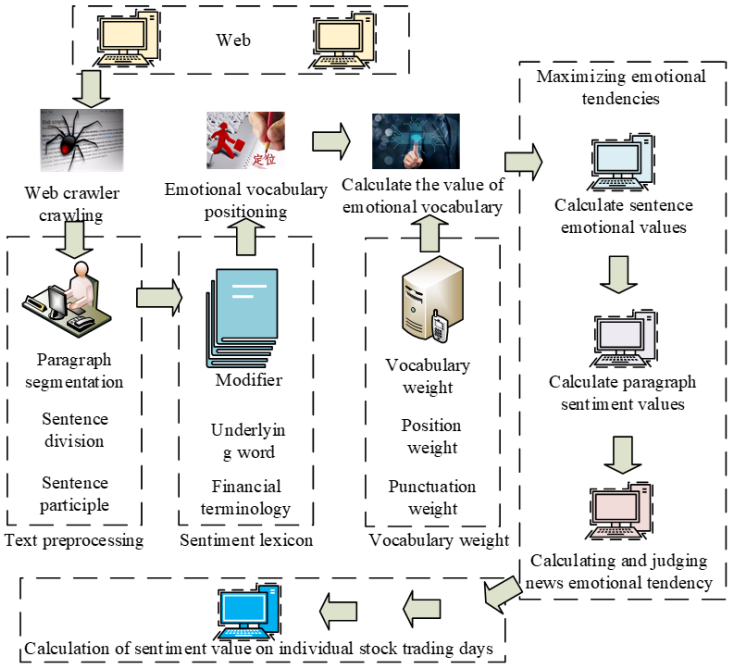


Fig. 3.3: Flow Chart of Emotional Tendency Analysis

In equation 3.7, Z represents the sum of all paragraphs. to meet the calculation requirements, it is also necessary to calculate and divide each trading day and trading time in equation 3.8.

$$\text{Every-stock} = \frac{\sum_{i=1}^{S} \text{Con}_i}{S} \tag{3.8}$$

In equation 3.8, S represents the total news that appears on each trading day. Every-stock represents the average cumulative increase in financial news on the trading day. The financial news sentiment tendency analysis chart in Figure 3.3 was obtained through the above calculation formula.

In Figure 3.3, a web crawler is first used to crawl news data, and the captured data is preprocessed for text. A vocabulary of emotional tendencies was constructed and the emotional tendency vocabulary in each financial news was determined. The vocabulary weight has been enhanced, and the emotional vocabulary's emotional value has been calculated. By quantifying emotional vocabulary's tendency value, the financial news's emotional tendency value on the trading day was calculated.

3.3. Methods Based on ARIMA Model and BP Model. During the process of capturing real-time stock data, there may be data loss or garbled code. Therefore, it is necessary to preprocess the data first. Firstly, the captured data should be cleaned up to exclude information such as advertisements and stocks. Secondly, the missing data should be supplemented [21, 22, 23]. Finally, the captured information should be formatted to make it easier for computer recognition. In the stock market, many investors analyze individual stocks through various indicators. RSI is a commonly used indicator for analyzing stocks strength, and its formula is expressed as equation 3.9.

$$\text{RSI}(t) = 100 - \frac{100}{1 + \frac{a}{b}} \tag{3.9}$$

In equation 3.9, a represents an average increase in the closing price on t day, b represents an average decrease in the closing price on t day, and t represents a time cycle. The change rate was calculated by

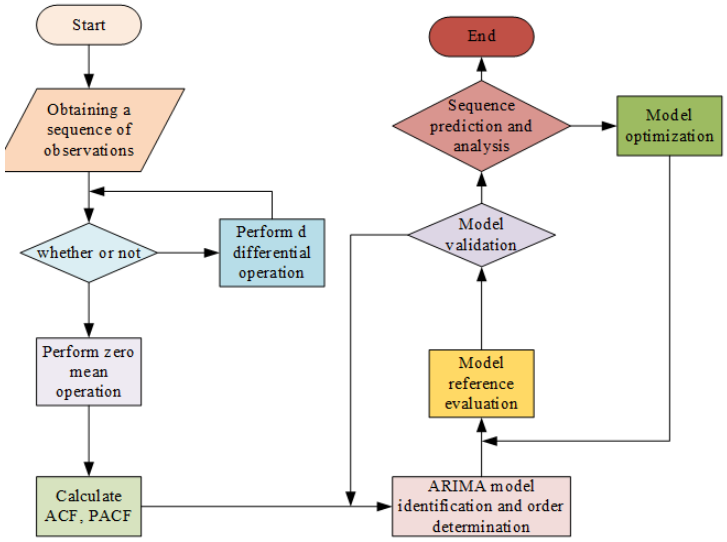


Fig. 3.4: ARIMA flowchart

comparing the current closing price with the past t weeks' price within a given time period in equation 3.10.

$$\mathrm{ROC}(t) = \frac{u-k}{k \cdot c} \tag{3.10}$$

In equation 3.10, u represents the current closing price. k represents the closing price before t days. c represents a fixed value of 100. t represents the calculation parameter. The common stock market will analyze low data characteristics based on the above two indicators, so as to reflect the daily fluctuations of stocks well. ARIMA model in Figure 3.4 was constructed based on stocks characteristics.

In Figure 3.4, stock prices stability was obtained by capturing data from stocks. Through unity value's root test, it can distinguish whether price is stable or not. Mean difference operations were performed on non-stationary time series until data was stable. The stationary time series was subjected to zero mean operation and then fitted to the data. Finally, the model feasibility is tested until it is determined that model parameters are feasible. If it is not feasible, it needs repeat step four [18]. The combination model in series can achieve two models' combination in terms of feasibility. However, the concatenated results cannot separate the linear and nonlinear parts of NN, and cannot achieve mutual superposition effect. Therefore, a parallel combination model was used in this experiment to predict and analyze data in equation 3.11.

$$f(x) = \omega_1 G_1 + \omega_2 G_2 \tag{3.11}$$

In equation 3.11, ω_1 represents a single model weight, G_1 represents ARIMA model's predicted value, and G_2 represents BP model's predicted value. In parallel combination structures, this model is mainly influenced by weight value, which determines its prediction results. Only through reasonable weight calculation can the model be more stable and reasonable. Equation 3.12 represents the average weight.

$$\omega_i = \frac{1}{n} \tag{3.12}$$

In equation 3.12, ω_i represents the weight, and n represents the weight calculating number. According to the error variance in equation 3.13, the weight value can be calculated.

$$\omega_i = \frac{i}{\sum_{i=1}^{n} i} = \frac{2i}{n(n+1)} \tag{3.13}$$

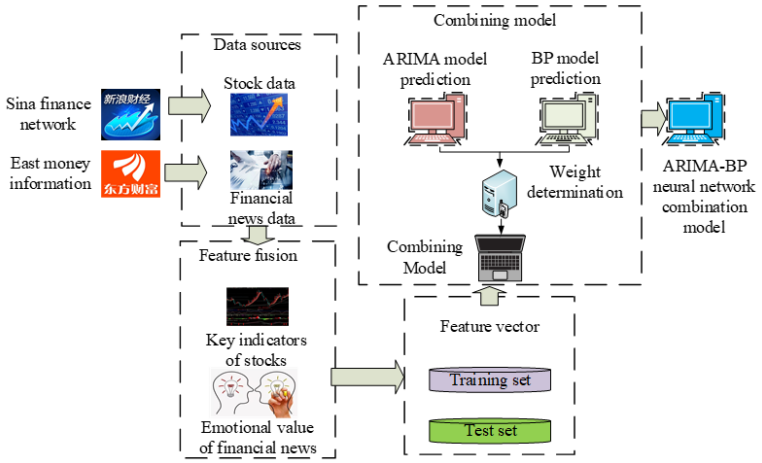


Fig. 3.5: Framework of Fusion Emotional Analysis Combination Model

In equation 3.13, $\sum_{i=1}^{n} \omega_i = 1$, $\omega_i > 0$, $i = 1, 2, 3, \ldots, n$. Calculating the weight value through relative error can reflect model's predictive accuracy. The weight is larger, the error is smaller in equation 3.14.

$$\omega_i = \frac{E_i^{-1}}{\sum_{i=1}^{n} E_i^{-1}} \tag{3.14}$$

In equation 3.14, E_i^{-1} represents the i-th model's relative error, and other parameters are the same as equation 3.13. The mean square error formula compares the true and predicted values to reflect model accuracy in equation 3.15 [16].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.15}$$

In equation 3.15, n represents the sample number, y_i represents the actual value, and $\hat{y}_i$) represents the predicted value. By calculating model's accuracy stability and error size, the model feasibility for news sentiment analysis and stock price prediction can be determined. Figure 3.5 shows the model prediction diagram fused with ARIMA-BPNN.

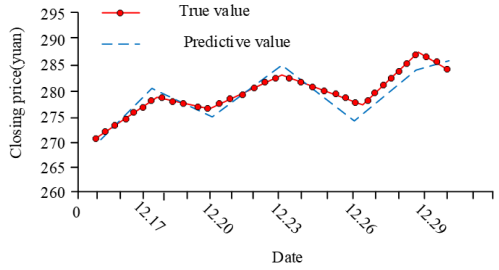
In Figure 3.5, the combined model first preprocesses the collected and captured data. Key indicators that can reflect current day's stock information were selected, such as trading volume, opening information, closing information, etc. Emotional tendencies were judged based on news' emotional information. Two types of information data were subjected to normalization analysis and processing. The data was further divided into training and testing sets based on time. Finally, the weights of the predicted results of the two models are determined to obtain the ARIMA-BPNN combination model.

4. Stock Price Prediction Results Analysis Based on ARIMA-BPNN Fusion News Sentiment Tendency Analysis. The programming language for this experiment is Python 3.8.6, the compiler is PyCharm, the MongoDB database is used for storage, and TXT and CSV documents are used for data preprocessing. Table 4.1 presents an emotional orientation analysis of news data from three listed companies and a comparative experiment on the emotional orientation of the model.

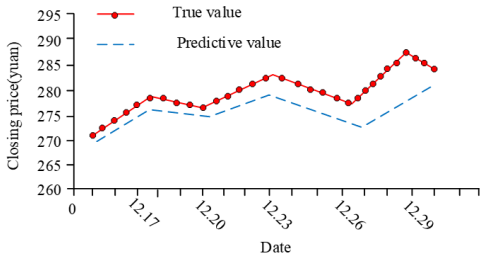
In Table 4.1, the tendency values of traditional and improved vocabulary for Boss Electric are -2.43 and -1.75, respectively. The reason may be that according to news reports, the current boss of Electric Appliances' earnings are not ideal. Hualan Biology and Jiangsu Hengrui Medicine both have positive increases in the

Table 4.1: Results of the Emotional Tendency Comparison Experiment

Stock Name	**Robam**	**Hualan Biology**	**Hengrui Medicine**
Stock code	002508	002007	600276
Headline	Stable profit margin	Complete vaccine release	Approved anti liver cancer treatment
Content	Slowing growth rate	Blood bank restore balance	Benefiting from the approval of new drugs
Improve emotional orientation values	-2.43	3.03	2.93
Traditional emotional tendency value	-1.75	2.67	2.13
/	Emotional tendency value		Basic-exp
R_active	79.24%		73.14%
A_negative	85.23%		81.24%
R_negative	82.14%		74.25%
A_negative	76.21%		69.27%



(a) BP neural network predicted value and true value

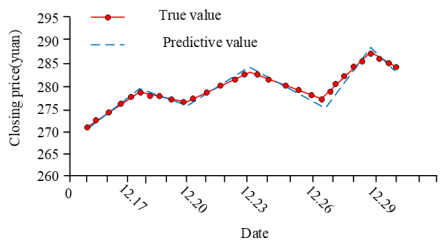


(b) ARIMA neural network predicted value and true value

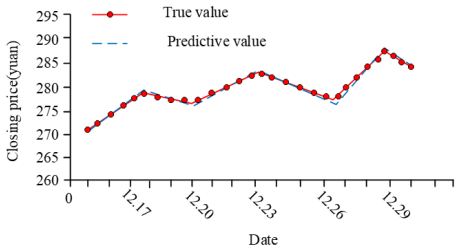
Fig. 4.1: Prediction Results of Wuliangye Yibin

improved emotional orientation value and the traditional emotional orientation value. It may be that the news shows that two listed companies' earnings are relatively good, resulting in a higher emotional tendency. Traditional emotion dictionary's calculation value is better than the improved emotion dictionary. A represents accuracy, R represents news recall, and $BAISC_{exp}$ represents emotional tendency comparison. Table 3.1 shows that the recall rate and accuracy rate of traditional emotional orientation comparison are 73.14% and 81.24%, respectively. The combined model's recall rate is 79.24%, its accuracy rate is 85.23%. Two models' difference is 6.1% and 3.99%. The difference in negative recall rate and accuracy rate is more significant, with a difference of 7.86% and 6.94%, respectively. When expressing and conveying different emotions in news, emotional inclination degree varies. It is not comprehensive to only consider vocabulary in the emotional analysis of the news itself. Therefore, optimizing the weight of vocabulary is a necessary condition to increase the accuracy and feasibility of the model. When forecasting the stock data of BPNN and ARIMA models, Wuliangye Yibin Group was selected as SP data from December 15, 2020 to December 30, 2020. Figure 4.1 is the predicted results.

In Figure 4.1, ARIMA's predicted results are roughly similar to the true curve trend. But the error results fluctuate between $\pm 7.4 \sim 2.5$, with significant fluctuations in the upper and lower errors. This suggests that when forecasting stock prices, the use of a single model for price forecasting can result in large deviations from the true value. BPNN's prediction curve results are significantly better than ARIMA, with errors fluctuating between $\pm 2.2 \sim 1.8$ yuan. The fluctuation trend of ARIMA model curve is significantly smaller than the true



(a) ARIMA-BP neural network predicted value and true value



(b) Fusion of emotional analysis ARIMA-BP neural network predicted values and real values

Fig. 4.2: Fusion sentiment analysis ARIMA-BP and ARIMA-BP model prediction results

Table 4.2: Four Model Error Indicators

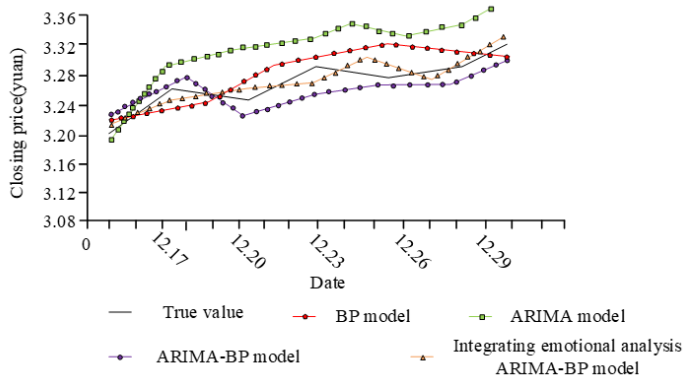
Time	12.15	12.16	12.17	12.18	12.21
True value	270.8	274.33	277.81	276.86	281.91
ARIMA error	-2.56	1.42	2.66	2.98	5.05
BP error	1.2	-1	-0.4	1.14	-1.54
ARIMA-BP error	0.56	-0.59	0.12	1.45	-0.42
Integrating Emotional AnalysisARIMA-BP error	-0.31	1.03	1.18	0.75	-0.12

value. BP not only greatly improves prediction accuracy, but also has a significant improvement in predicting trend changes, Also the BP model has a significant advantage in the predictive effect of the data. After performing algorithmic predictions on both models, a single weight value was reassigned and the ARIMA-BPNN parallel combination model was used to perform algorithmic predictions on data. The combined algorithm model integrated with sentiment analysis was used for data prediction in Figure 4.2.

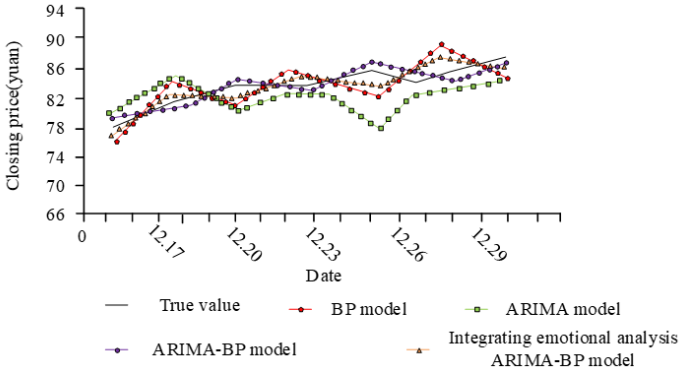
In Figure 4.2, the combined algorithm model error fluctuates between $\pm 2.1 \sim 0.5$ yuan. Most data prediction errors are between $\pm 1.1 \sim 0.9$ yuan, almost overlapping with true value's fluctuation curve. Its prediction results are significantly better than BPNN and ARIMA models. When predicting the model after fused sentiment analysis, the error value significantly decreased from error $\pm 2.1 \sim 1.5$ to $\pm 2.0 \sim 0.5$ yuan, which is the smallest in four prediction results. The predicted curve is almost consistent with the actual curve, and there are more overlapping parts compared to models that do not integrate sentiment analysis. The model that integrates sentiment analysis outperforms other three models in terms of predictive performance. This suggests that after combining the two models, the model is able to combine the strengths of the two models to improve the processing of the time series, and thus improve the prediction of the data. Table 4.2 compares four models' evaluation indicators.

In Table 4.2, the combination model in which some of Wuliangye Yibin's values are mixed with feelings has the lowest error under relative conditions, and the lowest error was -0.31% on December 15. Each model has the lowest error moment, so it is necessary to further judge whether each model can predict the stock price but is relatively optimal. After combining the two algorithmic models, the new algorithmic model appeared to have fewer prediction errors, which suggests that by combining the two algorithmic models, the prediction accuracy of the model can be improved and the errors appearing in the prediction can be reduced, which may be due to the fact that the model combines the advantages of the two separate models. To test model's ability to deal with different data, the stock data of Bank of China and Juewei Duck Neck are compared and analyzed in Figure 4.3.

In Figure 4.3, regarding the prediction results of Bank of China, among four models, ARIMA model has the largest deviation from the actual curve trend, while other three models have consistent trends with the



(a) Bank of china forecast chart

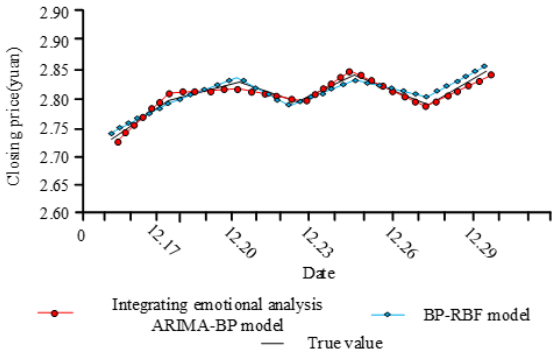


(b) Prediction diagram of juewei duck neck

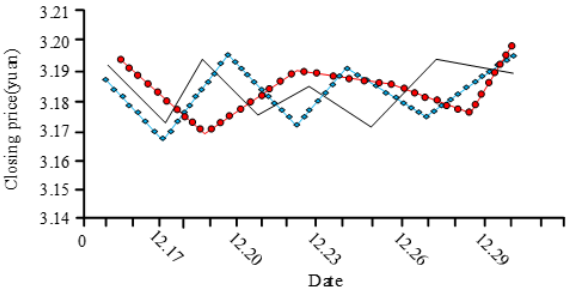
Fig. 4.3: Forecast of Bank of China and Juewei Duck Neck

actual values. The change curve of the combined model that integrates emotional analysis almost overlaps with the true value change curve. When analyzing Juewei Duck Neck's prediction curve, compared with the change trend of Bank of China, four models' prediction curves are almost consistent with the real curve. This shows that four models perform very well in Juewei Duck Neck's data predicting. However, the combination model that integrates emotional analysis still has the highest overlap degree with the true values, and its error change is also the smallest. This suggests that only fusing the models does not lead to better predictions, and with the incorporation of sentiment analysis it is possible to analyse the situation from more perspectives and subjective realities, which can enhance the predictions of the current study. Io further verify the combined model's superiority for integrating sentiment analysis, it was compared with the traditional commonly used BP-RBF combined model in Figure 4.4.

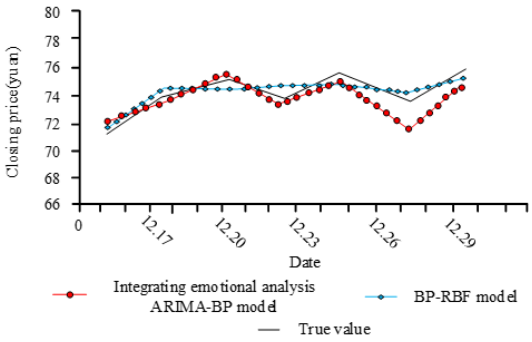
In Figure 4.4, in Wuliangye's stock data forecast curves, two models are consistent with the real results in terms of change trend. But the combination model that integrates emotional analysis has slightly higher overlap than traditional models. When comparing the data of Juewei Duck Neck and Bank of China, it shows a clear gap. The combination model that integrates emotional analysis is more in line with the true value curve, and its change trend is also consistent with the true curve. Therefore, the combination model that integrates sentiment analysis has more advantages in predicting data values, and its performance is also significantly better than traditional algorithm models. In order to compare the predictive effectiveness of different models on the same stock price, Autoregressive Moving Average (ARMA), Generalised Autoregressive Conditional Heteroskedastic-



(a) Wuliangye Yibin forecast chart



(b) Bank of china forecast chart



(c) Prediction diagram of juewei duck neck

Fig. 4.4: Prediction charts of three companies using ARIMA-BP and BP-RBF models fused with sentiment analysis

ity (GARCH), Generalised Autoregressive Conditional Heteroskedasticity (EGARCH), and Graphical Neural Network Models (GNNM) were compared with the models used in the study, and were obtained as shown in Table 4.3.

As can be seen from Table 4.3, in the price prediction of Wuliangye in the first half of December, the price of the ARIMA-BP model used by the study is closer to the true value, and the price prediction results of several of these months are consistent with the true closing price, which indicates that the study uses the algorithmic model with a higher prediction accuracy. At the same time, when several other models were analysed, it was found that the predictions in the other models deviated more from the true results, but there were also a few days where the predictions were consistent with the true results, but the number of occurrences was less. This indicates that the other algorithmic models are also capable of predicting stock prices, but their accuracy is lower.

5. Conclusion. At present, SP has become an important research direction in financial issues, and how to improve SP accuracy and precision is the current research focus. This experiment uses ARIMA algorithm and BP algorithm model to combine, and then uses news information for emotional orientation analysis to conduct stock data analysis. The study optimized the weight of news vocabulary by adding positional and punctuation weights, demonstrating a significant improvement in the accuracy of the emotional orientation values of the current optimized vocabulary weight. Simultaneously, an ARIMA-BP neural network combination model was constructed to predict non-stationary stock data. Finally, the sentiment analysis of financial news texts was integrated into the ARIMA-BP neural network combination model, further confirming that the sentiment

Table 4.3: Comparison of Five Different Models for Predicting Wuliangye Prices

/	Algorithm model	True value	ARMA	GARCH	EGARCH	GNNM	Research Use Model
Price	12.01	262	254	253	254	253	261
	12.02	268	256	255	255	257	267
	12.03	270	268	266	263	267	271
	12.04	276	268	273	270	270	275
	12.05	279	270	271	272	273	279
	12.06	276	272	274	276	271	277
	12.07	275	270	264	268	271	275
	12.08	281	273	270	271	272	280
	12.09	283	276	273	271	278	283
	12.10	286	281	281	280	279	286
	12.11	288	282	281	279	278	287
	12.12	285	276	275	279	274	286
	12.13	291	284	286	287	284	290
	12.14	290	290	276	286	284	290
	12.15	284	276	284	274	284	283

tendency of news texts is positively correlated with stock market volatility. The research results indicate that, When analyzing data of Wuliangye Yibin Group, the improved model's minimum root mean square error is 0.879%, and ARIMA model's maximum variance error is 3.342%. The combined model's average percentage error value for sentiment analysis is the lowest at 0.27%, while the ARIMA model's average percentage error value is the highest at 1.04%. The combination model that integrates sentiment analysis has the lowest error result of only 1.5%. And after analyzing the data of Juewei Duck Neck and Bank of China, the combined algorithm of emotion analysis is the same as the real value in change trend and coincidence degree. Its error is less than 1.5 yuan and 0.05 yuan, which is closer to true value compared to other models. And when compared with the traditional combination model, its performance prediction ability is also better than the traditional combination model. The errors of Wuliangye Yibin, Juewei Duck Neck and Bank of China are less than 0.03 yuan, 3.01 yuan and 0.02 yuan respectively. This study optimized the vocabulary use by designing modifications to the emotional vocabulary of the news, so that the algorithm's recognition ability for the emotionally inclined vocabulary was improved, while the improved algorithm was able to enhance the accuracy of the stock prediction, and further verified the relationship curve between the emotionally inclined and the fluctuation of the stock prediction.The algorithm studied in this study can improve the prediction ability of stocks, but in other aspects, the algorithm is limited by the ARIMA model, which can lead to missing data time series in data processing, thereby reducing the model's effectiveness. The data used in the study is only a part of the stock data, and in subsequent research, it is necessary to analyze the larger stock data. This study only used two models for combination, and more models will be combined in the future. At the same time, the models used in the study can also improve the prediction error of stocks, reducing the prediction error can more accurately predict the current stock price. At the same time, in the research, the model should also consider more advantages of the model and combine more stock factors for model improvement and analysis.

REFERENCES

[1] Ribeiro, G. André Alves Portela Santos, Mariani V C, LDS Coelho. Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Xpert Systems With Applications*. **184**, 2-14 (2021)

[2] Shapiro, A., Sudhof, M. & Sentiment, W. Journal of Econometrics. (2022)

[3] Gurrib, I. & Kamalov, F. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Tudies In Economics And Finance*. **39**, 347-364 (2022)

[4] Fedorova, E., Druchok, S. & Drogovoz, P. Impact of news sentiment and topics on IPO underpricing: US evidence. *Nternational Journal Of Accounting And Information Management*. **30**, 73-94 (2022)

[5] Yadav, S., Suhag, R. & Sriram, K. Stock price forecasting and news sentiment analysis model using artificial neural network. *Nternational Journal Of Business Intelligence And Data Mining*. **19**, 113-133 (2021)

[6] Zitnik, S. & Blagus, N. Baje cM. *Target-level Sentiment Analysis For News Articles*. **249**, 2-15 (2022)
[7] Mohan, B., Ahmed, S. & Kankar, M. Biju R. *Mohan.Hybrid ARIMA-deep Belief Network Model Using PSO For Stock Price Prediction*. **71**, 66-81 (2021)
[8] Colasanto, F., Grilli, L. & Santoro, D. Villani, Giovanni. *AlBERTino For Stock Price Prediction: A Gibbs Sampling Approach*. **597**, 341-357 (2022)
[9] Vara, P., Srinivas, G., Venkataramana, L., Srinethe, S., Sruthi, S. & Nishanthi, K. of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis. *He Computer Journal*. **5**, 1338-1351 (2021)
[10] Chen, Y., Fang, R., Liang, T., Sha, Z., Li, S., Zhou, Y. & Song, H. Stock Price Forecast Based on CNN-BiLSTM-ECA Model. *Cientific Programming*. **2021**, 2-21 (2021)
[11] Shapiro, A., Sudhof, M. & Wilson, D. Measuring news sentiment. *Ournal Of Econometrics*. **228**, 221-243 (2022)
[12] Kumar, C. Hybrid models for intraday stock price forecasting based on artificial neural networks and metaheuristic algorithms. *Attern Recognition Letters*. **147**, 124-133 (2021)
[13] Hanif, R., Mustafa, S., Iqbal, S. & Piracha, S. A study of time series forecasting enrollments using fuzzy interval partitioning method. *Journal Of Computational And Cognitive Engineering*. **2**, 143-149 (2023)
[14] Ribeiro, G. André Alves Portela Santos, Mariani V C, LDS Coelho. Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Expert Systems With Applications*. **184**, 2-14 (2021)
[15] Rezaei, H., Faaljou, H. & Mansourfar, G. Stock price prediction using deep learning and frequency decomposition. *Expert Systems With Applications*. **2020**, 5 (0)
[16] Gurrib, I. & Kamalov, F. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies In Economics And Finance*. **39**, 347-364 (2022)
[17] Fedorova, E., Druchok, S. & Drogovoz, P. Impact of news sentiment and topics on IPO underpricing: US evidence. *Nternational Journal Of Accounting And Information Management*. **30**, 73-94 (2022)
[18] Banerjee, A., Dionisio, A., Pradhan, H. & Mahapatra, B. Hunting the quicksilver: Using textual news and causality analysis to predict market volatility. *Nternational Review Of Financial Analysis*. **77**, 2-13 (2021)
[19] Zhang, Y. & Hamori, S. Do news sentiment and the economic uncertainty caused by public health events impact macroeconomic indicators?. *Evidence From A TVP-VAR Decomposition Approach*. **82**, 145-162 (2021)
[20] Ray, P., Ganguli, B. & Chakrabarti, A. Hybrid Approach of Bayesian Structural Time Series with LSTM to Identify the Influence of News Sentiment on Short-Term Forecasting of Stock Price. *IEEE Transactions On Computational Social Systems*. **8**, 1153-1162 (2021)
[21] Vijayalakshmi, B., Ramar, K., Jhanjhi, N., Verma, S., Kaliappan, M., Vijayalakshmi, K., Vimal, S., Kavita & Ghosh, U. An attention-based deep learning model for traffic flow prediction using spatiotemporal features towards sustainable smart city. *International Journal Of Communication Systems*. **34**, e4609 (2021)
[22] Nanglia, S., Ahmad, M., Khan, F. & Jhanjhi, N. An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing And Control*. **72** pp. 103279 (2022)
[23] Lim, M., Abdullah, A., Jhanjhi, N., Khan, M. & Supramaniam, M. Link prediction in time-evolving criminal network with deep reinforcement learning technique. *IEEE Access*. **7** pp. 184797-184807 (2019)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 20, 2023
Accepted: Feb 3, 2024