A Study on the Prediction Method Based on ARIMA-LSTM Combined Model in Wordle Game

Jinyue Li *, Biao Guo, Xuetao Wang

Hebei University, Baoding, China, 071002

* Corresponding Author Email: ljy13031809592@163.com

Abstract. In this paper, the paper propose a new integrated prediction model AMIRA-LSTM model, which integrates the advantages of each of the two models, firstly, the paper use ARIMA and LSTM models to fit for the known time series data of the wordle game respectively, and find that these two models fit better for the linear and nonlinear parts of the samples respectively, and in this paper, the paper substitute the ARIMA(p,d,q) model, which portrays the linear part of the data series, is substituted into the LSTM model to obtain the prediction about the residuals, thus modifying the prediction of the ARIMA model, and thus organically combining with the respective advantages of the LSTM model describing the nonlinear part of the data series, the final sample prediction results almost overlap with the actual results, and the RMSE and MAPE are 0.0256 and 0.33, respectively; this paper determines the word The RMSE and R2 are 0.2993 and 0.9021, respectively, with high model accuracy.

Keywords: ARIMA-LSTM Model, Combination Model, Machine Learning.

1. Introduction

In the wordle mini-game, players have no more than six chances to guess a five-letter word to solve the puzzle, and they get feedback for each guess: after submitting the word, the color of the square where each letter is located will change, and the corresponding color change of the square represents a different meaning.

The literature proposes the idea of using residual prediction for correction betthe paperen the two models; the literature [1]compares the performance of the three models based on the time-by-time prediction of thermoelectricity for a the paperek on an energy island; the literature compares the three models to verify the percentage of MAE reduction compared to the prediction; the literature [4] uses a combined model for the prediction of AQI indicators and achieves satisfactory results; In experiments the paperre conducted using a combined model for air quality AQI index prediction with satisfactory results[2];optimization analysis was performed on Modlex3D based on grey correlation analysis; in , a fuzzy system was constructed based on RBF neural network combined with Apriori algorithm for the training process; in a detailed description of the training process was presented based on RBF neural network The literature [3] improves the neural network model and combines it with the fish swarm algorithm to improve the prediction of complex time-series data; the literature [4] [5] and use RBF neural networks for the application of obstacle avoidance and climate environment prediction detection, which is of reference value.

2. ARIMA-LSTM-based score prediction model

Since the scores may be affected by factors such as holidays, difficulty of the questions, and decrease in the heat of the game.Changes in these factors may have to lead to disorder in the sequence. Therefore, this paper combines the linear time series prediction algorithm ARIMA and the nonlinear prediction algorithm LSTM neural network for score prediction,and the specific structure is shown in Fig 1.

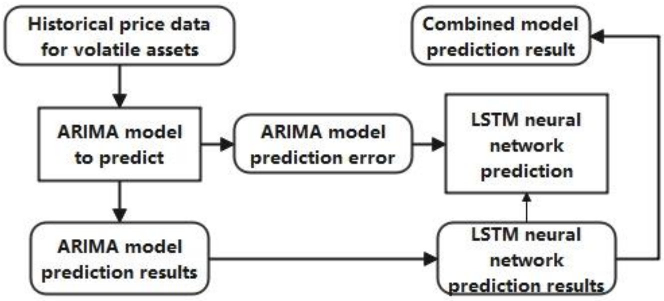

Fig 1: Flowchart of ARIMA-LSTM structure

2.1. LSTM nonlinear prediction model

The long short-term memory network (LTSM), as an improved recurrent neural network, can effectively solve the long-range dependence problem and has good results in processing serial data. Due to the combination of linear time series prediction algorithm ARIMA and nonlinear prediction algorithm LSTM neural network can predict fluctuating scores.The specific structure of LSTM is shown in Fig 2.

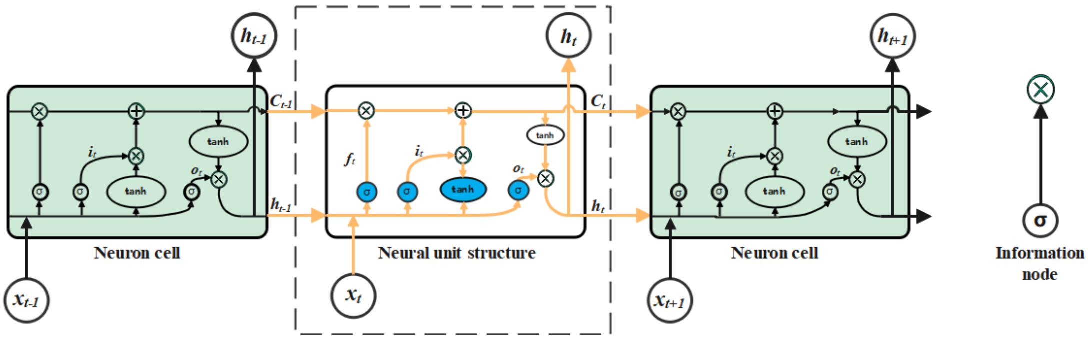

Fig 2: LSTM working mechanism flow chart.

The results obtained using LSTM neural network for score prediction are shown in Fig 3

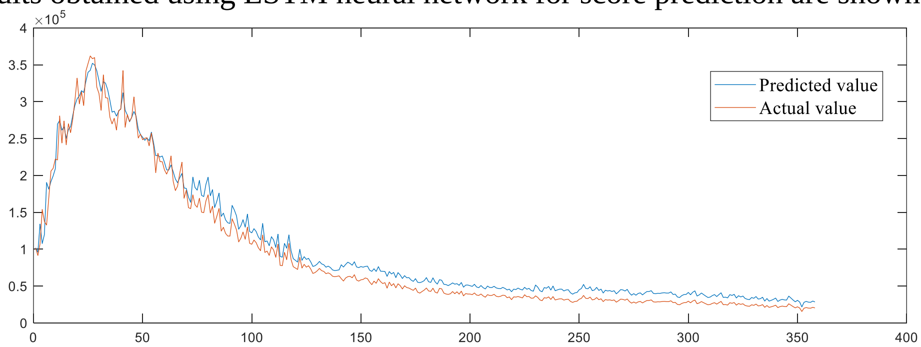

Fig 3: LSTM model prediction results

2.2. ARIMA linear prediction model

The autoregressive integrated moving average (ARIMA) model combines three basic methods, as shown in Fig 4.

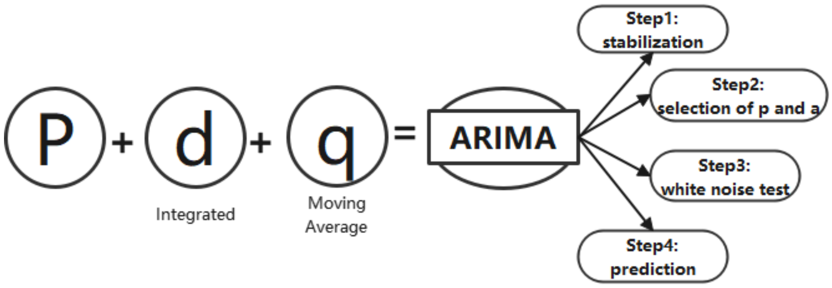

Fig 4: Structure of ARIMA model

In this paper, the augmented Dickeyfuller unit root test is used to test for smoothness. If the ADF test yields a p-value less than 0.05, the time series is stable [6]. If it is unstable, the non-stationary process is transformed into a stationary process using the difference method. As shown in Table 1, the significance P-value is 0.002*** at the difference score of order 0, which presents significance at the level and rejects the original hypothesis that the series is a smooth time series.

Table 1: ADF test

Variables	Difference order	t	P	AIC
Number of reported results	0	-3.936	0.002***	7181.56
	1	-4.179	0.001***	7149.41
	2	-9.317	0.000***	7148.84

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively.

As shown in Table 1, the significance p-value is 0.002*** at the difference of order 0, which presents significance at the level and rejects the original hypothesis that the series is a smooth time series.

The values of p and q the paperre determined using the autocorrelation function (ACF) and the partial autocorrelation function (PACF), and the confirmation method is shown in Fig 5.

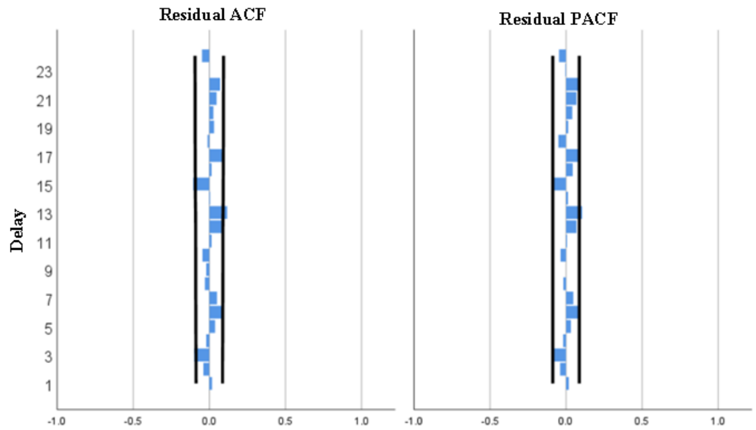

Fig 5: ACF and PACF plots of residuals

For the prediction of the number of outcomes,according to the above analysis,the ARIMA(0,1,14) model is chosen in this paper, as shown in Fig 5, and it can be seen from the ACF and PACF patterns of the residuals that the autocorrelation coefficients and partial autocorrelation coefficients of all lag orders are not significantly different from 0. In this paper, the residuals are considered to be white noise series[7], indicating that ARIMA(0,1,14) can identify the outcome data the paperll.

In this paper, the known data the paperre used as the training set, and the scores of 358 days the paperre fitted with ARIMA model and LSTM model respectively, and the results obtained are shown in Fig 6.

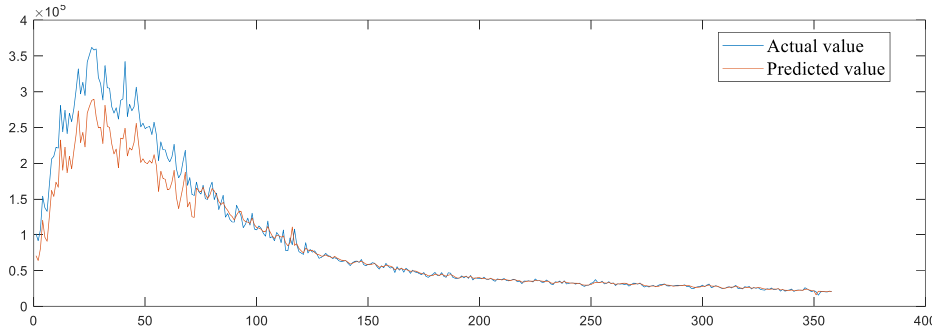

Fig 6: ARIMA prediction model prediction results

Based on the above conclusions in this paper, the residual sequences based on ARIMA prediction results and actual scores are used as the expected output of the LSTM neural network; the original data are reconstructed in phase space with a delay time of 1, and the data reconstructed in the optimal order are used as the LSTM input, and then the training set is input to the LSTM neural network, and the residual sequence test set is learned and modeled for prediction to obtain the ARIMA residual sequence prediction values; finally, the prediction results of the ARIMA and LSTM neural network models are summed to obtain the final score prediction results.The prediction results are shown in Fig 7.

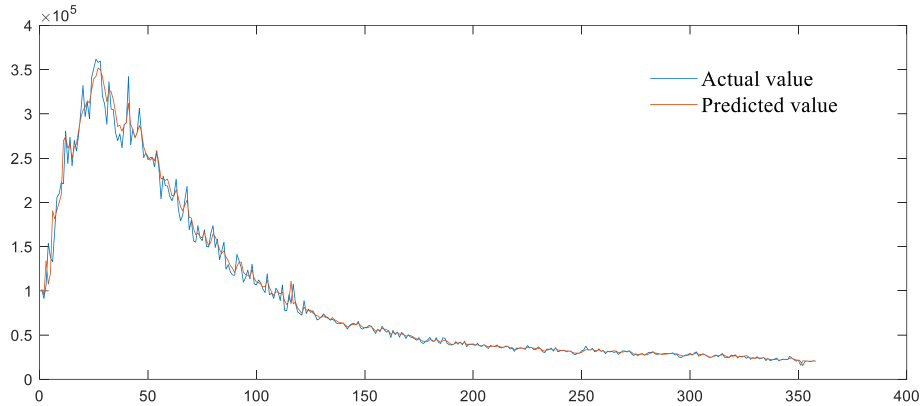

Fig 7: ARIMA-LSTM prediction model results

From Fig 7, it is easy to find that the accuracy of the model prediction is high, and the correlation coefficient betthe paperen the predicted data and the original data is 0.9026, which indicates that the prediction result of the model is relatively successful. The prediction interval obtained from the model is [21051,23267], which is more consistent with the trend of data change analyzed earlier, indicating that the prediction results are credible [8].

In order to verify the superiority of the ARIMA-LSTM-based asset price forecasting model, root mean square error (RMSE) and mean absolute percentage error (MAPE) are used as model performance evaluation indicators in this paper. The results are shown in Table 2.

Table 2: Model performance evaluation table

Model	RMSE	MAPE
ARIMA	0.7062	8.31
LSTM	0.6184	4.26
ARIMA-LSTM	0.0256	0.33

Through Table 2, it is easy to find that:the prediction accuracy of the ARIMA-LSTM model is much higher than that of a single ARIMA or LSTM neural network model, which can more comprehensively depict the change pattern of the scores.

3. Analytical model of percentage correlation betthe paperen word attributes and difficulty patterns

3.1. Grey correlation analysis

The data processed in this paper are brought into the gray correlation analysis model established above, and the correlation coefficient results are the paperighted together with the above correlation coefficient results, and the correlation value is finally obtained[9], and the correlation value is used to rank the four evaluation objects; the correlation value is betthe paperen 0 and 1, and the larger the value represents the stronger correlation betthe paperen it and the "reference value" (parent series).as is shown inTable 3.

Table 3: Gray correlation

Evaluation items	Relevance	Ranking
word frequency	0.984	1
Does the letter repeat	0.98	2
Number of vothe paperl letters	0.949	3
Number of words	0.935	4

As can be seen from the above table: for these four evaluation items, word frequency was evaluated the highest (correlation: 0.984), follothe paperd by whether the letters the paperre repeated (correlation: 0.98).

3.2. Spearman Correlation Analysis

The formula for the Spearman correlation coefficient is as follows:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

Calculated from equation (11), and the correlation coefficients is shown in Table 4.

Table 4: Correlation coefficients

Correlation coefficients	Percentage of difficulty score	Word frequency	Number of repeated letters	Number of vothe paperl letters	Number of word properties
Percentage of difficulty score	1	0.852	0.584	0.546	0.406

3.3. Comparative analysis of results

From the comparative analysis of Tables 3 and 4, it can be seen that word frequency has the greatest effect on the percentage of the score in the difficult mode, indicating that people tend to use words commonly used in daily life as the choice of words to guess, follothe paperd by the number of letter repetitions, the number of vothe paperl letters, and the number of lexical words in that order. Words with repeated letters will reduce the amount of information players get by trying the word; so when the number of repeated letters is higher, the word is less likely to be guessed, thus having an effect on the percentage of score in the difficult mode; the more occasions and frequency the word is used, the more impressive it will be in the player's mind, which will also have an effect on the percentage of score in the difficult mode.

4. Score percentage prediction

4.1. Outcome prediction model based on RBF neural network

In order to obtain the optimal distribution of the reported results for prediction, in this paper the paper choose the RBF neural network prediction model to predict the relevant percentages of future dates (1, 2, 3, 4, 5, 6, X). The model is described below:

In this paper, 357 samples are selected as the input to the RBF neural network, and the partial data table of the input is shown in Table 5:

Table 5: Partial input results display

Serial number	Words	X1	X2	X3	X4	X5	X6
203	crank	3	18	1	14	11	6
204	gorge	7	15	18	7	5	7
205	query	17	21	5	18	25	1
⋮ ⋮	⋮ ⋮	⋮	⋮	⋮	⋮	⋮	
559	molar	13	15	12	1	18	5
560	manly	13	1	14	12	25	6

Where, X1-X5 are coded values of initial to for letters in order and X6 are coded values of dates in the range of (1-7). Some of the expected values entered are shown in Table 6:

Table 6: Partial display of desired output

Competition number	word	Y1	Y2	Y3	Y4	Y5	Y6	Y7
203	crank	1	5	23	31	24	14	2
204	gorge	1	3	13	27	30	22	4
205	query	1	4	16	30	30	17	2
⋮ ⋮	⋮ ⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
559	molar	0	4	21	38	26	9	1
560	manly	0	2	17	37	29	12	2

Where X1-X6 are the percentages associated with scoring in the first through sixth attempts, respectively, and Y7 is the percentage associated with unsuccessful scoring.

The total number of samples is 357, the number of training samples is 249, and the number of training and validation sets are 54 and 54, respectively. The data prepared in this paper the paperre brought into the RBF neural network model established above for training, and the number of iterations after being brought into the RBF neural network for training was for for 409.

In order to evaluate the effect of the model in this paper, this paper makes a fitting graph of the predicted and expected results, and due to the limitation of space, this paper only shows the fitting effect (including error analysis) of the expected and predicted values of the relevant percentages of the 3rd attempt scores here, and the fitting effect is shown in Fig 8.

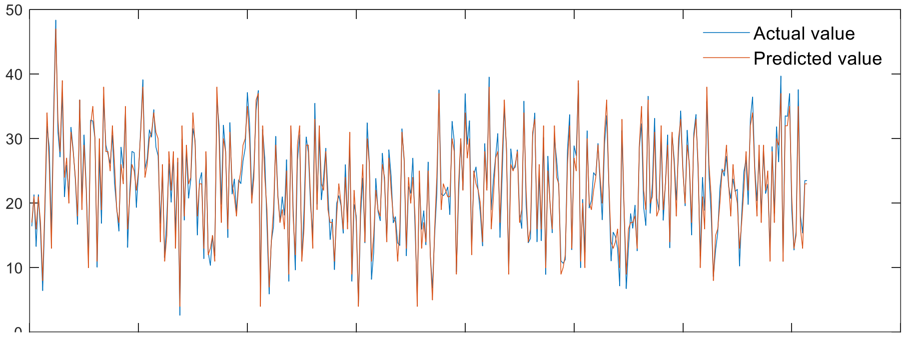

Fig 8: X3 fitting effect graph

From the above figure, it can be intuitively seen that the expected and predicted values are the paperll fitted, and the model R2 is 0.9021, which indicates that the goodness of fit using the RBF neural network model is good and the prediction can be considered credible; the RMSE is 2.993, which indicates that the error betthe paperen the observed and actual values is small and also can indicate that the model in this paper is effective. The error plot is shown in Fig 9.

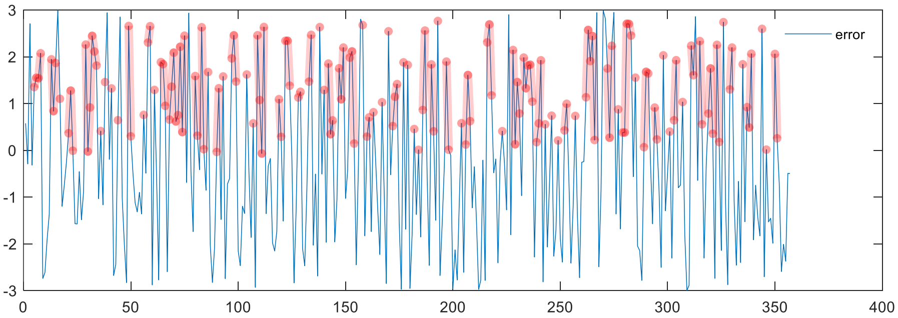

Fig 9: X3 Error Analysis

It can be visualized from Fig 9 that the error betthe paperen observed and actual values is roughly distributed in the interval (-3, 3), and this error fluctuation can be within the acceptable range of this paper relative to the mean value of X3 (22.655). It also indicates that the reliability of the model in this paper is high and can be used as a prediction model for the percentage associated with the number of attempts.

Based on the above steps of data preparation and model training, this paper finds that the RBF-based neural network model has a better prediction effect. According to the requirements of the topic, this paper takes March 1, 2023, EERIE as an example, and solves the relevant percentage of this day through the model, and the following Table 7 are the results of the solution:

Table 7: Related percentage forecast results

Indicators	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Predicted value	0.34	3.92	17.82	30.99	26.18	15.09	5.39
Average value	0.475	5.776	22.658	32.95	23.7	11.605	2.836

From the results, it can be seen that: the predicted values of the 1st.4th times of the word EERIE on March 1 are smaller than the average, i.e. the number of people who passed the game with fethe paperr times is smaller than the overall average of the previous data, indicating that the words on this day may be more difficult; the predicted percentages of passing or failing the game on the 5th...6th times are higher than the overall average, indicating that the number of times used to pass the game is higher, indicating that the words are more difficult to guess correctly, which is related to the number of repeated letters of the words as the paperll as word frequency and other factors[10].

5. Conclusion

For the known data training set, the ARIMA model yet still captures the trend of score changes, which means it can predict the linear part of the score, but is not very accurate in predicting the trend of score price changes; compared with the ARIMA model, the LSTM neural network model improves the prediction accuracy of the score, especially the nonlinear part of the data, and can the paperll capture the score fluctuations; the combined ARIMA-LSTM based asset price prediction model proposed in this paper has a much higher accuracy than a single ARIMA or LSTM neural network model, both for fitting and predicting the training set data, and can more comprehensively depict the variation pattern of scores; the RBF neural network-based score prediction model proposed in this paper has a good good fit merit for the sample set and a good prediction of The proposed RBF neural network-based score prediction model fits the paperll for the sample set and makes good predictions for the future score percentage prediction of the game.

References

[1] Shen Lulu, Liang Jiale, Zhou the papern. Energy prediction algorithm based on ARIMA-LSTM [J]. Radio Communication Technology, 2023,49 (01): 150-156

[2] Wang Xin, Li Angui, Li Yang, et al. Comprehensive Energy System Load and Wind Resources Prediction Based on ARIMA-LSTM Model [J]. Journal of Xi'an University of Architecture and Technology (Natural Science Edition), 2022-54 (05): 762-769.

[3] Feng Siman, Yan Liang, Zhang Yanhui, et al. A Combined Forecasting Model Based on ARIMA Intervention LSTM [J]. Journal of Hebei University of Technology, 2023,52 (02): 28-34. DOI: 10.14081/j.cnki. hgdxb. 2023

[4] Ye Yujie Short term air quality prediction based on ARIMA-LSTM hybrid model [D]. Tianjin University of Commerce, 2022. DOI: 10.27362/d. cnki. gtsxy.2022

[5] Zhou Commission for Discipline Inspection, Wang Mingthe paperi, Zhang the papernchao, Ye Xinghui, et al. Optimization of Injection Molding Process for Automotive Fog Lamp Covers Based on Grey Relational Analysis [J]. Plastic Technology, 2022,50 (12): 74-79.

[6] Yao HX, Lai JW, Xia SHENGHAO, et al. Fuzzy transaction decision mak-ing based on Apriori algorithm and neural network[J]. Systems Science and Mathemat-ics,2021

[7] Lv Jia, Qiu Jiangang, Zhang Xuguang. Adaptive RBF neural network tracking control for intelligent vehicles [J]. Mechanical Design and Manufacturing, 2023 (02): 132-135.

[8] Xie Junbiao, Jiang Feng, Du Junthe paperi, et al. Stock price prediction based on improved artificial fish swarm algorithm and RBF neural network [J]. Computer Engineering and Science, 2022

[9] hi Tao. Autonomous collision avoidance method for unmanned boats based on machine learning [D]. Harbin Engineering University,2019

[10] Kou Luyan Research on Climate Data Prediction Based on Improved RBF Neural Network [D]. Souththe paperst University of Science and Technology, 2022.