ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

Research on Wordle Based on Time Series Model and Grey Prediction Model

Jianguo Deng[1,*], Jiahang Liang[1], Yi Wu[2], Dan Wang[1]

[1]College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, 300222, China
[2]College of Information and Intelligence, Hunan Agricultural University, Hunan, Changsha, 410125, China
[*]Corresponding author: dengjianguo426@163.com

Abstract: *In this paper, the time series model and grey prediction model are established, and the stability of ARIMA model is tested by difference test, and the prediction is carried out. The prediction result is taken as the upper interval of the prediction interval, and the grey prediction result is taken as the lower interval of the prediction interval. The time series and grey prediction models are established respectively. Secondly, this paper establishes Pearson correlation coefficient model to analyze the relationship between lexical attributes and the percentage of difficulty pattern attempt scores. Three indicators are selected for the attributes of words: word complexity, letter frequency and the number of repeated letters, which are analyzed by taking the average number of attempts. It is found that these three indicators are significantly correlated with the number of attempts, the number of letter repetition and word complexity are positively correlated with the average number of attempts, and the letter frequency is negatively correlated with the average number of attempts.*

Keywords: *Time series, Neural network module, Elbow rule, K-Means clustering, Euclidean distance*

1. Background

Wordle is a popular jigsaw puzzle provided by the New York Times. Players need to guess a five letter English word. After each guess, they will get feedback. The yellow box indicates that the letter is in the word but the position is wrong, the green box indicates that the letter is in the word and the position is correct, and the gray box indicates that the letter is not in the word. The game supports regular and difficult modes, which require players to use the correct letters in subsequent guesses after finding them. Players can try six or fewer guesses in the game every day to guess words, and each guess must be an actual English word. Many players report their scores on Twitter, and MCM has generated daily result files from January 7, 2022 to December 31, 2022, including words on that day, the number of people reporting scores on that day, the number of players in difficult mode, the percentage of words guessed correctly and the proportion of problems that cannot be solved.

2. Research method

This study aims to solve the following problems:

How to explain the changes in the results of daily reports and establish a model to establish a prediction range for the number of results reported on March 1, 2023. Whether there is a word attribute affects the score percentage of the difficulty mode of the report.

In order to study the trend of daily reports, time series model and grey prediction model can be established respectively, and the prediction results can be used as the upper and lower bounds of the prediction interval. Analyze the relationship between the attributes of words and the percentage of attempted scores in difficulty mode. Through correlation analysis, determine which attributes have an impact on the percentage of scores.

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

3. Model establishment and solution

3.1 Trend of daily report

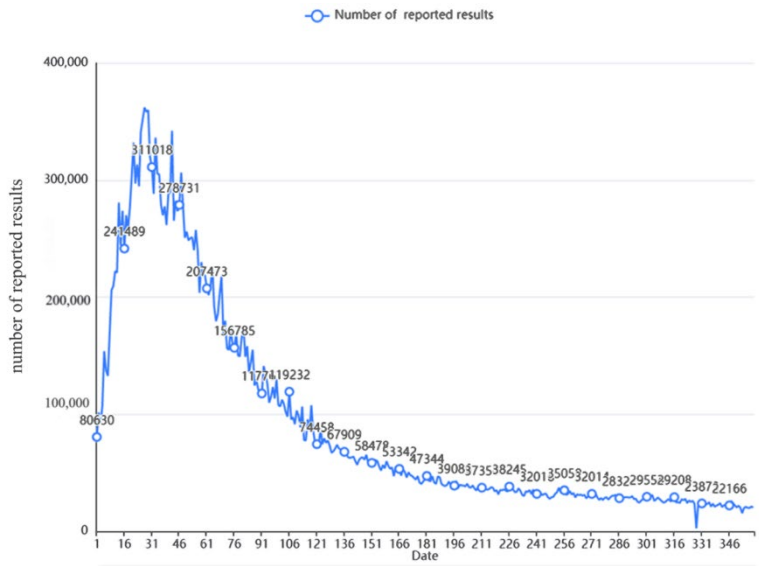

Figure 1: Number of reported results

As can be seen from the figure 1, the number of reported results rises sharply in a short period of time and then drops slowly after a period of time. As for this change trend, the explanation is that when the Wordle software was launched, it quickly received widespread attention, but as time went by, the popularity of the software gradually declined, and finally became stable.

3.2 Establish time series (ARIMA) and gray prediction model

3.2.1 Establishment of time series model

ARIMA (autoregressive moving average model) is a time series analysis technology, which is used to model and predict regular time series data[1-4]. It is based on autoregressive (AR), moving average (MA) and mixed (ARMA) definitions. It only considers the input data and does not consider external independent variables[5,6]. It is usually used to predict seasonal and trend effects. It builds models on existing data to predict future values[7]. In ARIMA model, the future value of the sequence is expressed as a linear function of the current period and lag period of the lag term and random interference term, that is, the general form of ARIMA (p, d, q) model is as follows[8-10]:

$$y_t' = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}' + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \quad (1)$$

(1) Stationary test of time series

Unit root test or difference test can be used to test the stability of time series. In this paper, we use difference test to test the stationarity of time series. The ADF test function is used to test the stationarity of the difference sequence, and then the stationarity result of the sequence is estimated through the p-value test.

The following table 1 shows the results of ADF test, including variables, difference order, T test results, AIC values, etc.

Table 1: ADF Test Results

ADF Inspection Form							
variable	Difference order	t	P	AIC	critical value		
					1%	5%	10%
Number of reported results	0	-3.867	0.002***	7203.313	-3.45	-2.87	-2.571
	1	-4.242	0.001***	7195.208	-3.45	-2.87	-2.571
	2	-10.663	0.000***	7176.608	-3.45	-2.87	-2.571
Note: * * *, * * and * represent the significance levels of 1%, 5% and 10% respectively							

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

It can be seen from the table that the results of the sequence test show that, based on the variable Number of reported results, when the difference is of order 0, the significance P value is 0.002 * * *, showing significance horizontally, and rejecting the original hypothesis. When the difference is of order 1, the significance P value is 0.001 * * *, showing significance horizontally and rejecting the original hypothesis. When the difference is of order 2, the significance P value is 0.000 * * *, showing significance horizontally and rejecting the original hypothesis. To sum up, this series is a stable time series(Figure 2 and Figure 3 and Figure 4).

Figure 2: Sequence diagram after 0 order difference

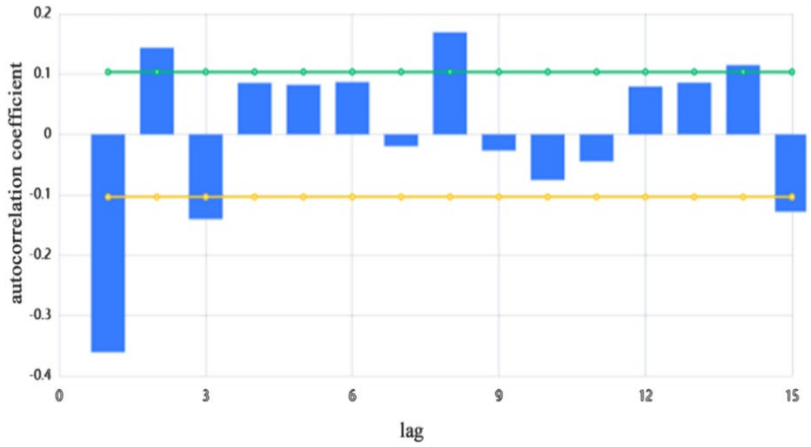

Figure 3: Autocorrelation Diagram of Final Differential Data (ACF)

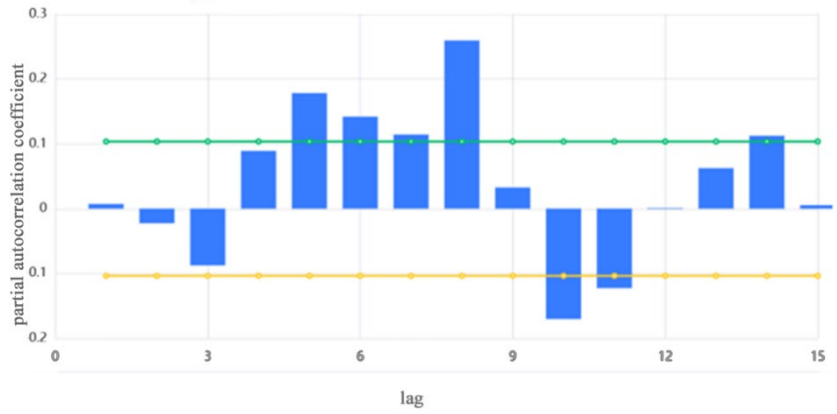

Figure 4: Partial Autocorrelation Diagram of Final Differential Data (PACF)

(2) Determine the order of the model

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

Table 2: ARIMA Test Results

Checklist of ARIMA model (1,1,0)		
term	Symbol	value
	Df Residuals	356
Number of samples	N	359
q statistic	Q6 (P value)	0.019(0.891)
	Q12 (P value)	24.347(0.000***)
	Q18 (P value)	54.224(0.000***)
	Q24 (P value)	82.246(0.000***)
	Q30 (P value)	98.285(0.000***)
Information guidelines	AIC	7749.064
	BIC	7760.705
Goodness of fit	R^2	0.982
Note: * * *, * * and * represent the significance levels of 1%, 5% and 10% respectively		

Find the optimal parameters based on AIC information criteria(Table 2). The model result is the ARIMA model (1,1,0) test table. Based on the variable: Number of reported results, from the analysis of Q statistics results, it can be concluded that Q6 does not show significance horizontally(0.019), and the assumption that the residual error of the model is a white noise sequence cannot be rejected. At the same time, the fit degree R^2 of the model is 0.982, the performance of the model is good, and basically meet the requirements.

(3) Parameter estimation and diagnostic test

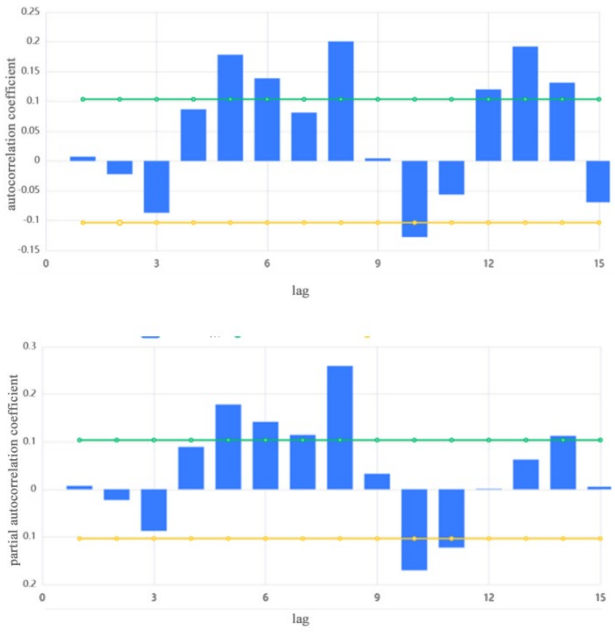

Figure 5: Model residual autocorrelation diagram (ACF) and Model residual partial autocorrelation diagram (PACF)

Table 3: Model Inspection Table

Model Parameter Table						
	coefficient	standard deviation	t	P>|t|	0.025	0.975
constant	-168.296	467.032	-0.36	0.719	-1083.661	747.069
ar.L1.D.Number of reported results	-0.362	0.049	-7.344	0	-0.459	-0.266
Note: * * *, * * and * represent the significance levels of 1%, 5% and 10% respectively						

It can be seen from the appeal chart that the model setting is basically correct (Figure 5 and Table 3).

(4) Predict with the established ARIMA model

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

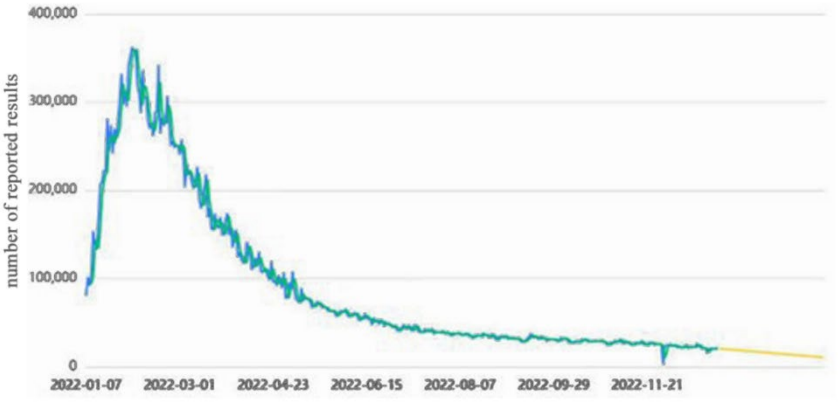

Figure 6: Time series prediction chart

The above figure 6 shows the original data graph, model fitting value and model prediction value of the time series model. The model is used to predict 60 days backward, and the predicted value of number of reported results on March 1 is about 21298. We take it as the upper interval of the prediction interval.

(5) Establishment of gray prediction model

The brief principle of GM (1,1) prediction model is: first, use the accumulation technology to make the data have exponential law, then establish a first-order differential equation and solve it, and then reduce the results to the gray prediction value, so as to predict the future. Since GM (1,1) model is only applicable to a small amount of data, the data from September to December are selected for prediction.

1) Before establishing the grey prediction model, it is necessary to ensure the feasibility of the modeling method, that is, it is necessary to carry out a level comparison test on the known original data. Set the initial non negative data sequence as:

$$X^{(0)} = \left\{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\right\} \tag{2}$$

Only when allσ(k) The model can be established only when it falls into the calculation range. The calculation and judgment formula of stage ratio are as follows:

$$\sigma(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \sigma(k) \in \left(e^{-\frac{2}{n+1}}, e^{-\frac{2}{n+1}}\right) \tag{3}$$

Obtained after accumulation $x^{(0)}$ The first order cumulative sequence of can be weakened $x^{(0)}$ Disturbance of:

$$x_k^{(1)} = \sum_{i=1}^{k} x_i^{(0)}, k = 1,2, \dots, n \tag{4}$$

$Z^{(1)}$ yes $X^{(1)}$ The sequence generated by the nearest mean of

$$Z^{(1)} = \left\{z^{(1)}(2), z^{(1)}(3), \dots, z^{(1)}(n)\right\} \tag{5}$$

$$z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right) \tag{6}$$

Therefore, the corresponding differential equation of GM (1,1) model can be obtained as follows:

$$x^{(0)}(k) + az^{(1)}(k) = b \tag{7}$$

among $Z^{(1)}$ Is the background value of GM (1,1) model.

According to the calculation results, all the stage ratios of the series after translation and transformation are within the interval (0.98, 1.02), which indicates that the series after translation and transformation is suitable for building a gray prediction model.

2) Build data matrix B and data vector Y, respectively as follows:

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

$$B = \begin{bmatrix} -z(2) & 1 \\ -z(3) & 1 \\ \vdots & \vdots \\ -z(n) & 1 \end{bmatrix} \qquad Y = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{pmatrix} \tag{8}$$

Then the least squares estimation parameter column of the grey differential equation meets:

$$u = [a \ \ b]^T = (B^T B)^{-1} B^T Y \tag{9}$$

Among them, the development trend of a main control system is called development coefficient. The size of b reflects the relationship between data changes and is called grey action.

3) Establish the model and solve the generated value and the restored value. According to the formula, the prediction model can be obtained:

$$\hat{x}^{(1)}(k) = \left[x^{(0)}(1) - \frac{b}{a}\right] e^{-a(k-1)} + \frac{b}{a} \tag{10}$$

$$k = 1,2,\cdots,n$$

After accumulation and subtraction, the predicted value of restoration is obtained.

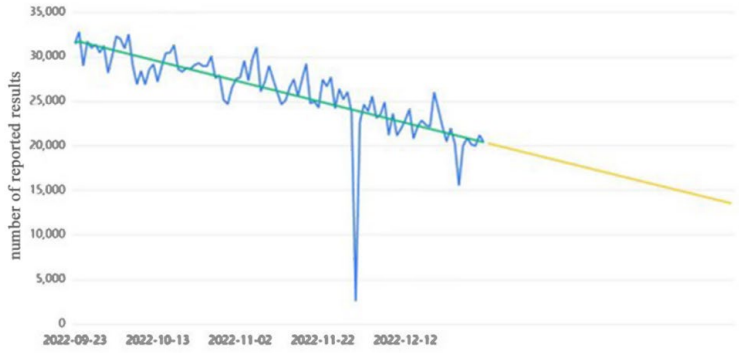

Figure 7: Gray prediction diagram

The above figure 7 is the model fitting prediction chart, and the average relative error of the model is 13.005%, which means that the model fitting effect is good. The predicted number of reported results on March 1 is about 17654. We use it as the lower range of the prediction range.

Combined with time series and gray prediction model, the prediction interval of number of reported results on March 1 is [1765421298].

3.3 Establish correlation model

3.3.1 Data preprocessing

Before determining the word attribute indicators, since most words in the data are 5 in length and only a few words are 4 or 6 (4) in length, the impact of this attribute can be ignored and the data of these words can be discarded.

3.3.2 Determine word attribute indicators

When a player guesses a word in Wordle, the attributes of the word itself will affect the number of times the player guesses the word. The attributes that may affect the number of guesses are the complexity of the word, the number of repeated letters in the word, and the frequency of the letters in the word in English.

We define the uncommon degree of words as the complexity degree of words. The complexity degree of words can be divided into five grades A, B, C, D, E according to the difficulty degree of words in the dictionary. In order to facilitate analysis, these five grades are quantified as 0.2, 0.4, 0.6, 0.8, and 1.0, that is, the larger the index, the more complex the words are. Therefore, the word complexity can be taken as an indicator.

Because the occurrence frequency of each letter in English is different, we can average the occurrence

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

frequency of the letters of a word, that is, sum up the occurrence frequency of each letter in the dictionary and then weighted the average as the attribute indicator of the word letter frequency.

The number of repeated letters in a word can also be used as an indicator. The maximum number of repeated letters in a word is taken.

In order to more intuitively analyze the relationship between text attributes and the percentage of attempts, average the percentage of attempts.

3.3.3 Establish Pearson correlation model

(1) Test the significance relationship

Before Pearson correlation coefficient, it is necessary to test whether there is a statistically significant relationship between the indicators to determine whether the P value is significant (P<0.05). As shown in Table 4.

Table 4: Correlation coefficient and p-value table

	Letter frequency	**Number of repeated letters**	**Average attempts**	**Word complexity**
Letter frequency	1(0.000***)	0.024(0.646)	-0.359(0.000***)	-0.039(0.466)
Number of repeated letters	0.024(0.646)	1(0.000***)	0.386(0.000***)	0.028(0.594)
Average attempts	-0.359(0.000***)	0.386(0.000***)	1(0.000***)	0.233(0.000***)
Word complexity	-0.039(0.466)	0.028(0.594)	0.233(0.000***)	1(0.000***)
Note: * * *, * * and * represent the significance levels of 1%, 5% and 10% respectively.				

(2) Calculate Pearson correlation coefficient

Let the data of word complexity X: {X1, X2,., X359}, the average number of attempts Y: {Y1, Y2,., Y359}, the letter frequency N: {N1, N2,., N359}, and the number of repeated letters M: {M1, M2,., M359}.Choose the data of word complexity and the average number of attempts for examples.

The mathematical expression of the population mean is

$$E(x) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{11}$$

Where x is the independent variable and n is the number of samples. Substitute the data of word complexity X and the average number of attempts Y into formula (11) to find the respective overall mean values.

The mathematical expression of the standard deviation is

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - E(X))^2}{n}} \tag{12}$$

Substitute the data of word complexity X and the average number of attempts Y into formula (12) to find the standard deviation.

The mathematical expression of the overall covariance is:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - E(X))(Y_i - E(Y))}{n} \tag{13}$$

Substitute the data of word complexity X and the average number of attempts Y into formula (13) to find the respective overall covariance.

The mathematical expression of the Pearson correlation coefficient of the population is

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \frac{(X_i - E(X))(Y_i - E(Y))}{\sigma_X \quad \sigma_Y}}{n} \tag{14}$$

Substitute the respective population mean and the respective population covariance into Formula (14) to obtain the respective population Pearson correlation coefficients. Similarly, the overall Pearson correlation coefficient between other indicators can be calculated.

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

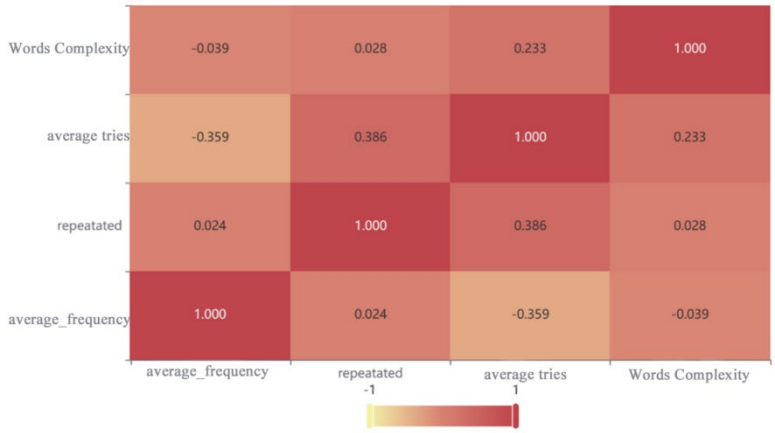

Figure 8: Correlation coefficient diagram

It can be seen from the figure 8 that letter frequency, number of repeated letters, and word complexity are all significantly related to the average number of attempts, that is, these three indicators will affect the reported percentage of attempts in difficult mode. And the letter frequency is negatively correlated with the average number of attempts, that is, the higher the frequency of the letters in the word in English, the less attempts. The number of repeated letters and word complexity are positively correlated with the average number of attempts, that is, the fewer the number of repeated letters in a word, the simpler and more common the word, and the fewer attempts.

4. Conclusions

In conclusion, this study has successfully developed and implemented time series and grey prediction models to forecast the upper and lower intervals of a prediction interval. The stability of the ARIMA model was confirmed through the difference test, ensuring the accuracy and reliability of the established models. The findings reveal a significant correlation between lexical attributes, such as word complexity, letter frequency, and the number of repeated letters, with the average number of attempts made by students in solving difficulty pattern problems. The results suggest that these three indicators can be used as valuable references for educators and learners to predict and improve the difficulty level of vocabulary learning. Furthermore, the Pearson correlation coefficient model established in this study has effectively analyzed the relationships between these lexical attributes and the percentage of difficulty pattern attempt scores. The positive correlation between word complexity and the number of attempts indicates that complex words tend to require more attempts for comprehension, whereas the negative correlation between letter frequency and the average number of attempts suggests that words with higher letter frequencies are more easily understood. These insights can contribute to the development of more effective vocabulary teaching strategies and assist learners in making more informed decisions when selecting vocabulary to learn. In summary, the combination of time series and grey prediction models, along with the analysis of lexical attributes and their correlation with difficulty pattern attempt scores, offers valuable insights into the complexity of vocabulary learning. The findings of this study can serve as a valuable reference for educators, learners, and policymakers in designing and implementing effective vocabulary teaching and learning strategies. Moreover, future research can further explore the applicability of these models in different educational contexts and expand the scope of vocabulary attributes investigated to develop more comprehensive understanding of vocabulary difficulty.

References

[1] Xiaolong Z,Congjun R,Xinping X, et al. Prediction of demand for staple food and feed grain by a novel hybrid fractional discrete multivariate grey model[J]. Applied Mathematical Modelling, 2024, 125(PB).

[2] Junting Z,Haifei L,Wei B, et al. A hybrid approach of wavelet transform, ARIMA and LSTM model for the share price index futures forecasting[J]. North American Journal of Economics and Finance, 2024,69(PB).

[3] Rui S , Shu-Hua W , Dong-Lian G ,et al.Research of combined forecasting model based on time

ISSN 2616-5775 Vol. 6, Issue 13: 145-153, DOI: 10.25236/AJCIS.2023.061321

series model and gray model[J].Journal of Yanshan University, 2012.
[4] Babu M S V K,Pratyush C,Mayukha P . Planning of fast charging infrastructure for electric vehicles in a distribution system and prediction of dynamic price[J]. International Journal of Electrical Power and Energy Systems,2024,155(PA).
[5] Sareh G H,Abbasali V,Reza M S . Desertification simulation using wavelet and box-jenkins time series analysis based on TGSI and albedo remote sensing indices[J]. Journal of Arid Environments, 2023, 219.
[6] Bin M,Xing G,Penghui L . Adaptive energy management strategy based on a model predictive control with real-time tuning weight for hybrid energy storage system[J]. Energy,2023,283.
[7] Yangyang S,Yuhang X,Yingjie W, et al. Reconstruction of incomplete flow fields based on unsupervised learning[J]. Ocean Engineering,2023,288(P1).
[8] Şenol Halil,Çakır İlkay Türk,Bianco Francesco. Improved methane production from ultrasonically -pretreated secondary sedimentation tank sludge and new model proposal: Time series (ARIMA). [J]. Bioresource technology,2023.
[9] Pantelis L,Vasilis P,Theodor P, et al. CO2 concentration forecasting in smart cities using a hybrid ARIMA–TFT model on multivariate time series IoT data[J]. Scientific Reports,2023,13(1).
[10] Ke W,Changxi M,Xiaoting H . Research on traffic speed prediction based on wavelet transform and ARIMA-GRU hybrid model[J]. International Journal of Modern Physics C,2023,34(10).