ARIMA-Based Prediction for the Number of People Reporting to Wordle

Yun Shen [*, #], Jiayi Cao [#], Yifan Wu [#]

School of Mathematical Sciences, Soochow University, Suzhou, China, 215006

* Corresponding Author Email: yvonne _s_h@163.con

[#]These authors contributed equally

Abstract. As soon as Wordle is launched, people all over the world are posting their scores on Twitter with great enthusiasm. The projection of the number of people helps the development of the software. To test the applicability of the ARIMA model in this area. In this paper, the data is fed into the model for prediction. The analyses show that the value of R^2 is 0.982, so the model fits well. The predicted values also show that the number of people reporting scores has a downward trend. This paper divides the attributes of the word into six dimensions. We use the Least Squares Estimation, Spearman Correlation Analysis and ADF Test to analyze the relationship between six attributes and the percentage of scores reported that were played in Hard Mode. The test shows that there was neither linear nor non-linear relationship between them.

Keywords: Wordle, ARIMA Model, Predictive Analysis.

1. Introduction

In English-speaking countries, Crossword Games have a long history and a wide range of audience, and changes are also varied [1]. Crossword is a representative example. By contrast, Wordle seems greatly simplified in both difficulty and form [2,3]. The game is updated daily to guess a five-letter word. As soon as the game is launched, it arouses players' interest and desire to challenge, and other countries have followed and developed versions in different languages [4-6]. Qi Yang et al. developed a series of evaluation models to predict Wordle words, and the results showed that the word prediction models based on convolutional neural network models performed moderately well, and the evaluation models were generally effective [7]. The ARIMA method has been widely used in forecasting fields [7], such as runoff forecasting, disease forecasting, and carbon emission forecasting [8-10]. However, this model is used because it is not widely used in this research field. Wei et al. used ARIMA model to develop prediction of monkeypox cases, and the results showed that ARIMA (2, 2, 1) model performed best on global monkeypox dataset with MAPE value of 0.040, while ARIMA (2, 2, 3) performed best on US and French datasets with MAPE values of 0.164 and 0.043, respectively [11]. Zhao Jiwei et al. used ARIMA model to predict monthly historical precipitation data for 1973 - 2021 in Luoyang City. The indicators show that the model predictive performance outperforms the EMD-LSTM (Empirical Mode Decomposition), EEMD-LSTM, EEMD-ARIMA combined models and the single models, and the model has high confidence in the prediction results of future precipitation [12].

In conclusion, we will use ARIMA model to predict the number of Wordle users, and the research results will bring huge profits to Wordle formula.

2. ARIMA model

In this paper, we determine the following ARIMA parameters. As shown in Table 1.

Table 1. ARIMA parameters

Symbols	***Definition***
p	The number of lags in the timing data itself used in the prediction model
d	Order of differential differentiation required for time series data
q	The number of lags in the prediction error used in the prediction model

And make the following derivation. Suppose y represents the difference of Y at time t

$$d = 0, y_t = Y_t \tag{1}$$

$$d = 1, y_t = Y_t - Y_{t-1} \tag{2}$$

$$d = 2, y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \tag{3}$$

Hypothesis p, q, d is known. The mathematical formula of ARIMA can be expressed as:

$$\hat{y}_t = \mu + \varphi_1 * y_{t-1} + \ldots + \varphi_p * y_{t-p} + \theta_1 * e_{t-1} + \ldots + \theta_q * e_{t-q} \tag{4}$$

We use ADF test on the data to determine whether the time series is stationary. When the difference is 0, the p value is 0.002 (<0.05). Therefore, the null hypothesis is rejected and the sequence is stable.

The data are delayed for 6, 12 and 18 steps successively, and we can observe that all ***p*** values of the model haven't surpass 0.05. Therefore, we reject the null hypothesis and conclude that the correlation exists between the sequence itself.

The sequence meets the condition of stationarity and pure randomness, so ARIMA model can be adopted. According to the autocorrelation analysis and the partial autocorrelation analysis, we can obtain that p=1 and q=1. Further, we test the model——ARIMA (1,0,1). The result shows that R2 was 0.982, so the model performs well and basically meets the requirements.

The formula of the model obtained is as follows: y_t=-170.68-0.369×y_(t-1), from which the number of reported results on March 1 can be predicted to be 10,315. The 95% prediction interval is (5803, 33,075). As in Figure 1.

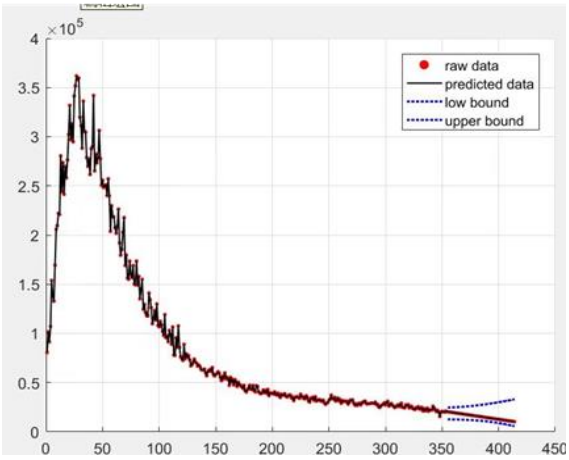

Figure 1. ARIMA Model Prediction Interval

Considering the impact of early rising data on model prediction, we intercept first-order difference stabilisation data to predict, As in Figure 2.

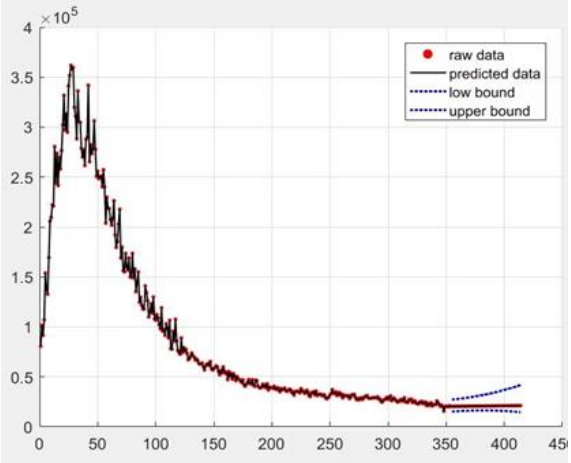

Figure 2. Improvement ARIMA Model Prediction Interval

March 1 can be predicted to be 21,184. The 95% prediction interval is (14,521, 41,703)

Since the rising data will interfere with the subsequent data at the beginning of the game, we chose the improved model.

3. Impact Analysis

In this section, an in-depth analysis of the influencing factors is presented. The influencing factors are divided into the following 6 categories:

High-frequency Letter: 'E', 'T', 'A,' 'O', 'I', are selected as the first five high-frequency letters in all words, and the data is classified by the number of high-frequency letters that contain them. Type of Letter: Classify data by numbers containing different letters. Number of Consonants: Data are classified by the number of consonants they contain. High-frequency Initial: 'T', 'A', 'O', 'W', 'S', as the first five high-frequency initials in all words and classify the data by the number of high-frequency initials containing them.

Familiarity: Find the frequency use of 330,000 words, sort it, and the sorted value corresponds to the data, and then standardize them.

Special Data: If both letters in a word appear twice, or if one letter appears three times, it is defined as special data. As in Figure 3.



Figure 3. Word Attributes Affect the Percentage

Firstly, we calculate that the VIF value equal to 1 represents no multicollinearity, so we can use the ordinary least squares method to judge the linear relationship of the model. Checking the model according to the results of the F-test, where the p value > 0.05, we can easily get that the model is invalid. In other words, there is no linear relationship.

When we use the Spearman hierarchy correlation coefficient to determine whether there is a strong nonlinear relationship between the models, the relationship obtained is shown in the Figure 4. The correlation is extremely low, so there is no nonlinear relationship.

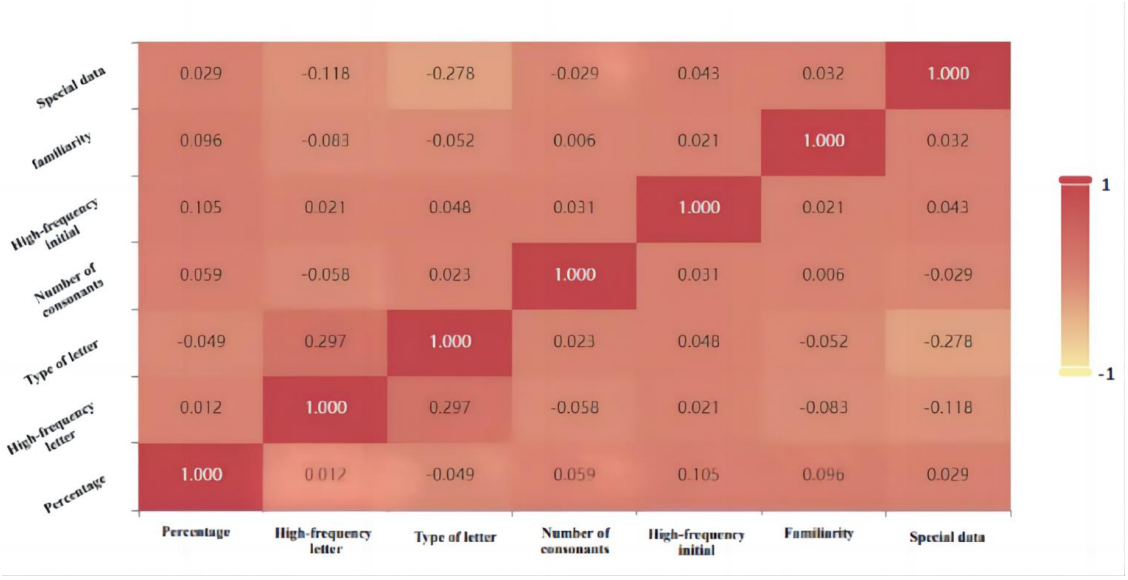

Figure 4. Heatmap of Spearman Correlation Analysis

In the case of first-order difference, this is shown in Figure 5, it is stationary and hence exhibits reversion to the mean, so we can say the time series is stationary and check it with the ADF test.

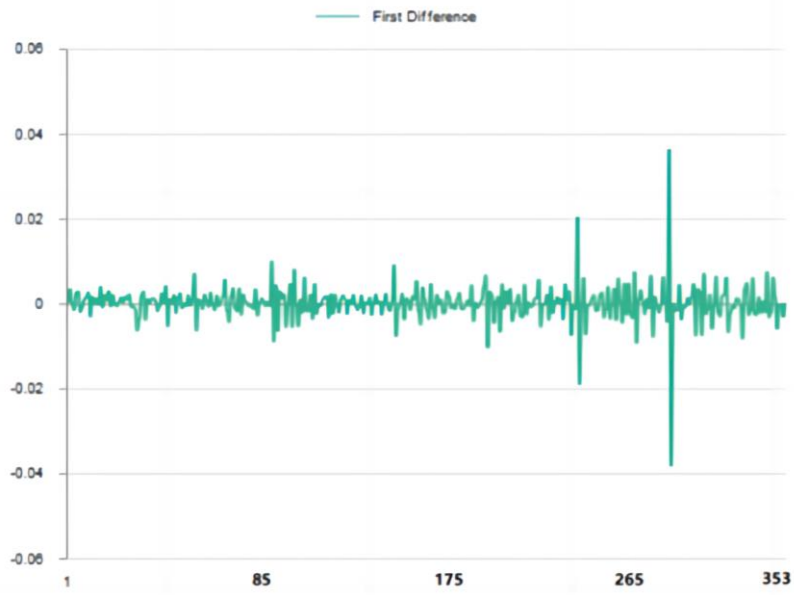

Figure 5. First-order Difference Plot

In general, for Problem 1, we first use two ARIMA models to predict the number of participants in March this year, and our view is that the current rules of the game and other conditions remain unchanged, the number of participants will level off in a short period of time. After that, using Regression analysis methods, it was discovered there is no attribute of the word affect the percentage of scores reported that were played in Hard Mode.

4. Conclusions

In this study, we aimed to predict the user count for the popular word-guessing game, Wordle. By analyzing various factors, such as social media trends, user engagement, and game mechanics. In conclusion, this study demonstrates that predicting the user count for Wordle can be achieved by considering social media trends, user engagement, and game mechanics. By monitoring these factors and adapting strategies accordingly, game developers and marketers can gain valuable insights into Wordle's growth trajectory and make informed decisions to foster its continued success.

References

[1] Bao Chen, Wang Yunhai. Review of word cloud Visualization [J]. Journal of Computer Aided Design & Computer Graphics,2021,33(04):532-544.

[2] Du Hua. English Reading Teaching Design and Practice of Text Cloud Map: A Case study of text cloud map tool Wordle [J]. Modern Educational Technology,2012,22(09):65-69.

[3] Wang Yifan, Wu Xiangyang, Wang Jikang, et al. Fast Generation of Shape constraint Word Cloud with Free Focus [J]. Journal of Computer Aided Design & Computer Graphics,202,34(06):869-877. (In Chinese)

[4] Short Martin B. Winning Wordle Wisely—or How to Ruin a Fun Little Internet Game with Math [J]. The Mathematical Intelligencer,2022,44(3): 227-237.

[5] Feng Siman, Yan Liang, Zhang Yanhui, et al. Combinatorial prediction model based on ArMI-Interventive-LSTM [J]. Journal of Hebei University of Technology,2023,52(02):28-34.

[6] Zheng Minggui, Yu Ming, Fan Qiurong, et al. Lithium carbonate demand Forecast in China from 2025 to 2035: a combined model based on grey correlation analysis and ArMI-GM-BP neural network [J]. Advances in Earth Sciences, 2019,38(04):377-387.

[7] Qi Yang, Yifan Fang, Yu Zheng. Word Data Research and Prediction Based on Wordle Game [J]. Academic Journal of Computing & Information Science, 2023,6(4),18-19.

[8] Chang Jianbo, Fu Xingtao, Zheng Meijun, et al. Spatio-temporal evolution and prediction of carbon emissions from energy consumption: Based on night light data in Shanxi Province [J]. Environmental Science and Technology,2023,46(03):81-89.

[9] Liang Jianguo, Feng Yongliang, Dong Li, et al. Application of ARIMA product season model in the prediction of tuberculosis in Shanxi Province [J]. Disease Surveillance, 2019,38(03):332-338. (In Chinese)

[10] ZHANG Wanlin. Prediction and Analysis of Ningbo Port Logistics Demand based on GM (1,1)-ARIMA combined model [J]. Logistics Engineering and Management,2023,45(01):32-35.

[11] Wei Wudi, Wang Gang, Tao Xing, et al. Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, GM (1, 1) and LSTM deep learning methods. [J]. The Journal of general virology,2023,104(4),33-36.

[12] Zhao Jiwei, Nie Guangzheng, Wen Yihao. Monthly precipitation prediction in Luoyang city based on EEMD-LSTM-ARIMA model[J]. Water Science and Technology,2023,87(1).

[13] Chang Jianbo, Zhao Zhilin, Shi Zhanxiang, et al. Optimization and Control of synchronous orbit entry time for Mars Probe [J]. Mathematical Modeling and Applications,202,11(02):80-87.