Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 3, 2023

Machine Learning based Data Analysis for Wordle Game

Zidong Wu *

Department of Hangzhou Dianzi University, Zhejiang, China
*** Corresponding author Email:** wzd1052390788@163.com

Abstract: As wordle game is becoming more and more popular all over the world, the study of the intrinsic patterns and problem-solving techniques of the game has become a popular topic. In this paper, we obtained daily result files of wordle over a period of time, which included data such as words of the day, the number of people who reported their scores on the day, and the number of players who entered the difficult mode, and further analyzed the data. In order to predict the results reported on a future day, this paper first eliminates the abnormal data in the dataset, and then builds the LSTM model and ARIMA model. On the test set, the MAPE (mean absolute error) of the LSTM model is 5.643, and the LSTM is significantly better than the ARIMA model. Secondly, in order to predict the distribution of word results for a given future date, the 12 features of the network training data were first subjected to PCA dimensionality reduction, and the results showed that the percentages were consistent with time by performing the Shapiro-Wilk normalization test on the correlation percentages. Based on this observation, we built a BP-LSTM parallel model to extract word attributes and extract percentage features over time. The model has a MAPE of 6.09, which outperforms a BP neural network that can only extract word attributes.

Keywords: Wordle; ARIMA; LSTM; BP-LSTM-Parallel.

1. Introduction

Wordle is a popular word guessing game on the Internet [1]. Players have up to six chances to guess a five-letter word. For each guess, the system gives a hint based on how well the guessed word matches the correct answer. The game is designed so that you can easily share your word-guessing patterns after guessing the answer. The game can only be played once a day and people all over the world are guessing the same word. As a result, the process of winning players sharing their word-guessing results is very social. Thanks to its fun rules and healthy social attributes, Wordle has been played by tens of thousands of people for more than 300 days. We tried to analyze the results of future game reports based on past data and attempted to analyze the difficulty of each word in the questions.

In this paper, we have information on the date and number of Wordle matches, the daily match questions, the number of people reporting scores each day, the number of players participating in the difficult mode each day, and the percentage of guesses (1, 2, 3, 4, 5, 6, X) attempted by each player. A time series model is created based on the number of results reported each day to predict the range of results on a future date. Also in this paper, an attempt is made to create a predictive model to further predict the distribution of reported results for a single player attempt to guess on a particular day in the future.

2. Data Preprocessing

At the beginning of data preprocessing, we need to make sure that the data is available.

Outlier handling. In general prediction problems, the model is usually a representation of the overall structure of the sample data. The model focuses on the general properties of the sample as a whole, and points where the properties are completely inconsistent with the sample as a whole are called outliers.

First, we found that the number of outcome reports for November 29, 2022 was only 2,569, which is much less than the number of reports for December 1, 2022 and November 29, 2022, so we deleted the data for this day so that it would not participate in the model construction. Second, we also calculated the sum of the correlation percentages for specific dates (1, 2, 3, 4, 5, 6, X).

Smoothing Filter. The time series are predicted and trained one by one in terms of time. In this example, we are predicting March 1, almost 60 days before the last day of the dataset. If the prediction is done on a day-by-day basis, the prediction is difficult and prone to large errors. In order to reduce the prediction error, we use the average of 10 days as the time unit for training and prediction, which reduces the span of the time unit for prediction. The following formula is the relevant formula for smoothing filtering.

$$y(d) = \frac{1}{N}\sum_{k=0}^{N-1} x_{c}(d-k) \tag{1}$$

Standardized Processing. In machine learning, we often need to process different types of data. Here, we normalize the number of results reported each day. The relevant expression is as follows: where μ and σ are the mean and standard deviation of the sample data, respectively.

$$\frac{X_d - \mu}{\sigma} \tag{2}$$

3. Forecasting the Number of Reports

3.1. ARIMA Model

ARIMA model, fully known as Auto Regressive Integrated Moving Average Model, is a time series detection method proposed by Box and Jenkins in the early 1970s [2]. The ARIMA model is determined by three important parameters

(p, d, q), where p is the autoregressive term, d is the number of difference terms when the time series is stationary, and q is the number of moving average terms.

This model is a combination of autoregressive (AR) and moving average (MA), which can transform non-stationary time series into stationary time series, and then regresses the lagged values of the dependent variable, the present value of the random error term, and lagged values into the established model.

The ARIMA prediction model can be written as the following formula:

$$\hat{p}^{\{t\}} = p_0 + \sum_{j=1}^{p} \gamma_j p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \varepsilon^{\{t-j\}} \quad (3)$$

Where p is the order of the autoregressive model (AR), q is the order of the moving average model (AM), $\varepsilon^{\{t\}}$ is the error term between time t and $t-1$, γ_j and θ_j are the fitting coefficients, and p_0 is a constant term [2].

After calculation, it can be concluded that p=1, q=1, and d=3 in our established ARIMA model. we utilize the established ARIMA model for prediction and the prediction results are shown in Figure 1.

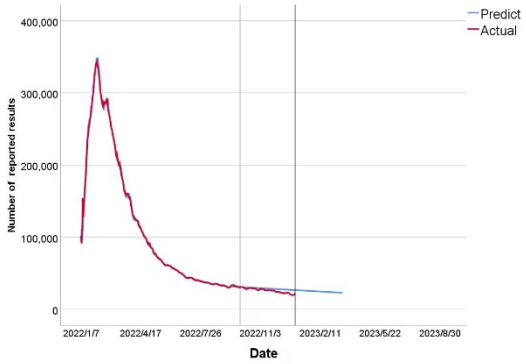

Figure 1. ARIMA Model Prediction Results

3.2. LSTM Model

In addition to using traditional machine learning methods to predict the number of reported results on March 1, 2023 we also tried to use deep learning methods.

In recent years, deep learning algorithms have been widely used in the field of time series forecasting models. Compared with machine learning, the most important feature of deep learning models is that they do not require the use of a priori knowledge to construct feature engineering, and can automatically learn the features of continuous text from data. There are various types of deep learning models, among which RNN (Recurrent Neural Network) is more suitable for processing time series. LSTM (Long Short-Term Memory) neural network is a variant of RNN, which improves the performance of RNN by memorizing the long prior information to solve the problem of gradient vanishing and explosion. In this example, we use the LSTM model to predict the number of future reported outcomes.

LSTM, or Long Short-Term Memory, is a special kind of recurrent neural network. It introduces the concepts of input threshold, forgetting gate and output gate, and is widely used in speech recognition, text processing, etc. The structure of LSTM neural network is shown in Figure 2.

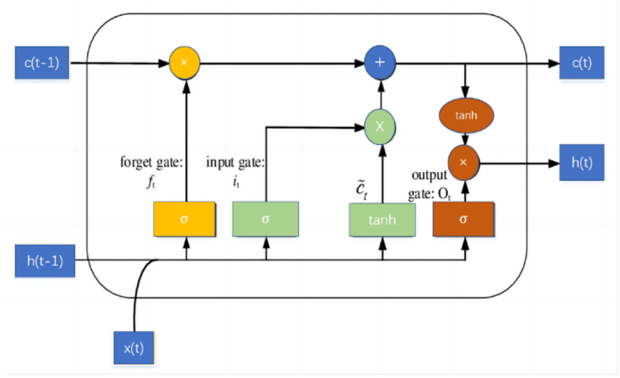

Figure 2. Decision tree basic structure

The LSTM controls C_t through an input gate and a forgetting gate. the forgetting gate ft controls the ratio of the previous state C_{t-1} to the state of the C_t cell at the current moment. Thus, the LSTM can retain a large amount of past information through the forgetting gate. The input gate i_t controls the ratio of the current moment's input to the current moment's cell state x_t. This helps to ensure that the content of valid information enters the memory. It also prevents some useless information from being saved again [3].

The activation function of the LSTM uses a sigmoid function. the sigmoid function is a smooth step function that is derivable and therefore mitigates the problem of vanishing gradients and can nonlinearly compress values between 0 and 1 to indicate how much information is passing through the sigmoid layer.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_t) \quad (5)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (6)$$

$$o(t) = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$\mathrm{h}_t = o_t \cdot \tanh(c_t) \quad (8)$$

We use the established LSTM model for prediction and the prediction results are shown in Figure 3.

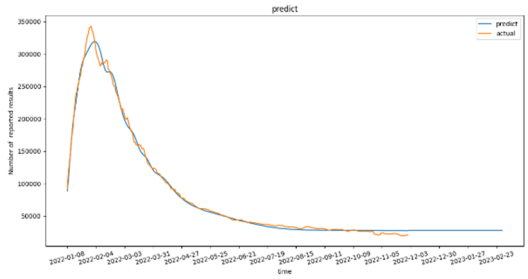

Figure 3. LSTM Model Prediction Results

3.3. Model Comparison Analysis and Questions and Answers

By comparing the training and validation accuracies of ARIMA model and LSTM model, we understand that LSTM model gives better prediction results. The model performance evaluation table is shown in Table 1.

Table 1. Model performance evaluation table

Model	MSE	MAPE
ARIMA	0.163	10.332
LSTM	0.051	5.643

4. Wordle Outcome Distribution Prediction Model

4.1. Predictive Theory

According to the requirements of the topic, we need to build a model to predict the distribution of reported results corresponding to a word on a certain date in the future. In other words, a multiple-input multiple-output prediction model is built with the relevant attributes of the word and the date of occurrence as input variables, and the relevant percentages of answers (1, 2, 3, 4, 5, 6, X) as output variables. (1, 2, 3, 4, 5, 6, X) as output variables. Therefore, we try to build relevant deep learning models.

Let the independent variable be time d, the sequence of independent variable word attributes $(P_1, P_2, P_3, \ldots, P_{12})$, and the sequence of percentage distribution of output report results $(Y_1, Y_2, Y_3, \ldots, Y_7)$. Its related expression is shown in the following equation.

$$Y_1^d, Y_2^d, Y_3^d, \cdots, Y_7^d = f(P_1^d, P_2^d, P_3^d, \cdots, P_{12}^d, Y_1^{d-1}, Y_2^{d-1}, Y_3^{d-1}, \cdots, Y_7^{d-1}) \quad (9)$$

4.2. Forecasting Methods

First, we need to determine the relevant attributes of the words. Still based on the word attribute labels used in the first question, each word attribute is vectorized in the same way.

As traditionally understood, a player's problem-solving performance should only be related to the word difficulty of the day. Each day's word difficulty should be independent of time. Therefore, we try to use only word attributes as dependent variables and build the model using BP neural network.

After testing, we found that the accuracy of the model was low. Then we considered that there are some loyal players in Wordle game. They play the game every day. In the day-to-day training, their proficiency will get higher and higher, so it will gradually improve the overall completion of the game over time.

In order to prove that the answer situation is related to time, we conducted a normalization test on the distribution data of the seven answer results. After the test, it was found that the data of the results of each trial did not satisfy the normal distribution, which proved that the original data was related to time.

In order to introduce the time variable, we introduced the LSTM neural network into the original BP neural network and constructed a new network structure. The specific structure of the BP-LSTM-Parallel model is shown in Figure 4.

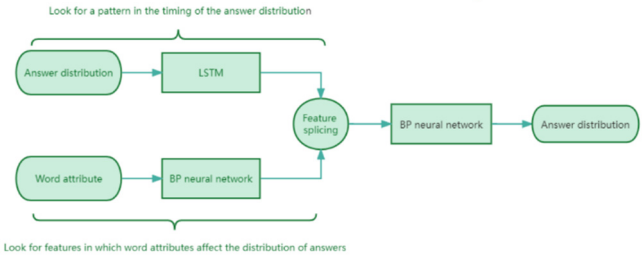

Figure 4. The Specific Structure of BP-LSTM-Parallel Model

4.2.1. BP Neural Network

First of all, as can be seen from the name, BP neural network can be divided into two parts, BP and neural network. BP is the abbreviation of Back Propagation, which means back propagation [4].

BP networks can learn and store a large number of input-output pattern mapping relationships without revealing the mathematical formulas describing the mapping relationships beforehand. Its learning rule is to minimize the sum of squared errors of the network by continuously adjusting the weights and thresholds of the network through backpropagation using the steepest descent method. The BP neural network flowchart is shown in Figure 5.

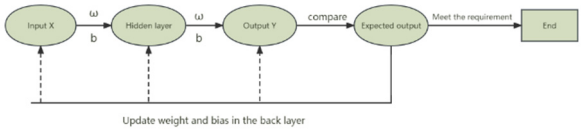

Figure 5. BP Neural Network Flow Chart

4.2.2. Test for Normality of Answer Distribution

In order to determine whether the percentage of guessing 1, 2, 3, 4, 5, or 6 words correctly or not solving the puzzle was time-dependent, we attempted to test for normality of the percentages for each day. Due to the small total amount of data for this question, we chose the Shapiro-Wilk method for the test. For the results of the normality test, if the p-value is greater than 0.05, we consider that it conforms to a normal distribution; conversely, if the p-value is less than 0.05, we consider that it does not conform to a normal distribution. The results of the normality test are shown in Table 2.

Table 2. Shapiro-Wilk test results

Attempts	1try	2try	3try	4try	5try	6try	7ormoretries(X)
P value	0.000	0.001	0.034	0.003	0.151	0.000	0.000

From the results of the normality test shown in Table 2, it can be seen that most of the variables do not conform to a normal distribution. Therefore, we believe that the relative percentage of time spent answering questions is related to time. Therefore, we introduced the time variable in our model.

4.2.3. BP-LSTM Parallel Modeling

In building this model, we first input the answer distribution features with temporal features into the LSTM branch to extract the temporal features and get the feature vector reflecting the temporal dimension features. We input the word attribute features without temporal features in another branch and output the reflection. Word attribute features, then the features obtained from the two branches are spliced, and finally the distribution of answers is obtained through the output of the fully connected layer [5]. The specific structure of the model is shown in Figure 6.

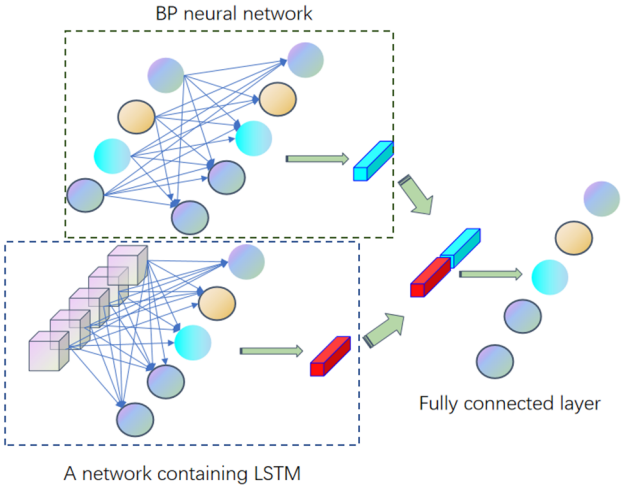

Figure 6. Detailed Structure of BP-LSTM-Parallel Model

4.3. Comparative Analysis of Model Prediction Performance

We select 80% of the data as the training set to train the model and select the remaining 20% of the data as the test set to validate the model. The results of the model on the training and test sets are shown below.

4.3.1. Performance During Model Training

We train the model. In general, the self-built network model is trained better than the traditional BP neural network. However, our training method introduces some error into the self-built network. When we use the model for prediction, we use day-by-day prediction, so when predicting, we need to set the word attributes to random numbers for prediction and add EERIE word attributes for prediction until March 1st. Randomly generated word attributes may lead to biased results, and small amount of data and high jitter may lead to overfitting of the model.

Intuitively analyzing the PCA dimensionality reduction test results for the test set of the above two models, it can be seen in Figure 7 that the BP neural network gradually deviates from the true value, while the BP-LSTM-Parallel model has a smaller error from the true value and the trend of the true value is the same, which can be seen that the BP-LSTM-Parallel model is better than the BP neural network.

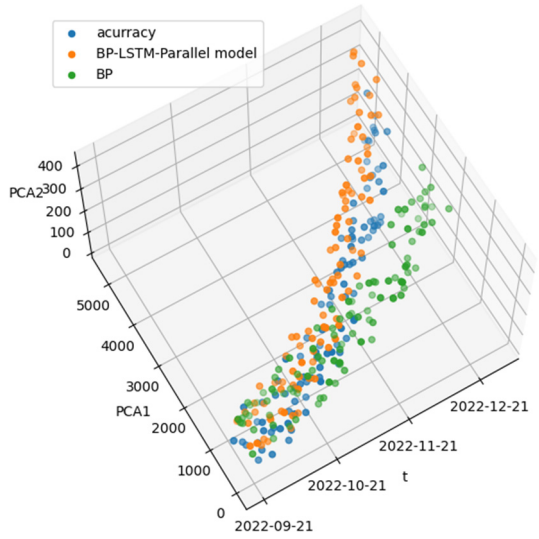

Figure 7. Comparison of Two Neural Networks

4.3.2. Performance During Model Testing

The BP-LSTM-Parallel model performs well when testing the model. The MAPE of the model on the test set is 6.09, so the deviation of the model prediction results will not exceed 6.09%. The correlation results are shown in Table 3.

Table 3. Model Performance Evaluation Table

Model	MSE	MAPE
BPNN	0.49	31.10
BP-LSTM-Parallel Model	0.3	6.09

5. Model Evaluation

As the original dataset fluctuates, LSTM can fit and predict the data better than other models when the data fluctuates. Therefore, LSTM model can achieve better results. The BP-LSTM-Parallel model not only takes into account the effect of word attributes, but also finds the temporal regularity of the percentage, which leads to better prediction.

References

[1] Du Hua. Design and Practice of English Reading Teaching with Word Cloud Mapping--Taking Wordle, a Word Cloud Mapping Tool, as an Example[J]. Modern Education Technology, 2012,22(09):65-69.

[2] Weng Zixia. Stock Price Analysis and Forecasting Based on ARIMA Model--Taking Construction Bank as an Example[J]. Modern Information Technology,2023,7(14):137-141. DOI: 10.19850/ j.cnki.2096-4706.2023.14.029.

[3] QIN Jiabing, LU Lihua, JI Chengfeng. LSTM stock price prediction model for sector ETF funds[J]. Fujian Computer, 2023, 39(08):15-19.DOI: 10.16707/j. cnki.fjpc. 2023. 08. 004.

[4] YANG Qiuge, WU Peng, YUAN Jing et al. Research on early warning of students' learning status based on BP neural network [J]. Science and Technology Innovation, 2023 (20): 133- 137.

[5] WANG Wei-Qiang, YU Jin-Quan, ZHOU Yi-He et al. Vehicle speed prediction based on BP-LSTM combined neural network model [J]. Intelligent Computer and Application,2022, 12(06): 54-59.