Journal of Electronics and Information Science (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2023.080601
ISSN 2371-9524 Vol. 8 Num. 6

Strategies for analyzing the guessing game "Wordle"

Peixin Guo[1,*], Baoqi Wang[2], Zuyou Fan[2]

[1]*College of Transportation Engineering, Chang'an University, Xi'an, Shaanxi, 710061, China*
[2]*School of Energy and Electrical Engineering, Chang'an University, Xi'an, Shaanxi, 710061, China*
[*]*Corresponding author: g983846671@163.com*

Keywords: Wordle; Curve Fitting; PSO-LSTM

Abstract: Nowadays, games have become indispensable for people's entertainment. Among them, the five-letter decryption game "world" launched by the New York Times has swept the world. Many players also reported their scores on Twitter. Through these published data, we found some interesting information. According to the information and requirements given by the topic, the table data attached to the topic is preprocessed. Find an exception in the data and delete it. The attributes of a given word are extracted by data encoding. We draw a line graph of the data, observe its trend change and perform curve fitting. The study found that the first half of the curve rose rapidly and the second half fell slowly. The fitting curve function was obtained, and the goodness of fit were 0.9521 and 0.9629, respectively. Using the curve, the quantitative range of results reported on March 1, 2023 is [5805.390,6075.43]. The sensitivity analysis was carried out by changing the four parameters of the fitting curve. The results show that the predicted value is within the reasonable range. Based on the parts of speech classification and the number of letter repeats, we investigate whether they affect the percentage distribution of difficult sentence patterns. Secondly, we optimize the LSTM model based on particle swarm optimization algorithm, and carry out hyperparameter optimization processing to build the PSO-LSTM model. Compared with LSTM, it is found that its model expression is better than that of single LSTM model. The MAPE value of the test set is 1.248, which means the uncertainty is 1.248%, so we have 98.752 percent confidence in the accuracy of the model. The EERIE data were encoded and put into the established PSO-LSTM model, and the correlation percentages were 0.432, 3.631, 19.326, 30.291, 26.954, 14.234 and 5.132, respectively.

1. Introduction

During the New Crown pandemic, many people began working from home [1]. To make work more fun and hands-on, creative engineer Josh Wardle developed a charades game called world [2]. It provides the player with one word per day as a puzzle [3]. The player needs to guess a 5-letter word no more than 6 times [4]. During the process of guessing the word, the player gets some hints by changing the color of the sticker. The game contains two modes. In regular mode, when submitting a word, if the patch turns green, it means that the letters in the post are in the word and are correct [5]. If it turns yellow, the letters in the post are in the word but in the wrong place. If it turns gray, the letter is not in the word. The difficulty model requires that once the correct letters are found, they must be used in subsequent guesses. In the title release, the author only updates the title once a day,

which is unsatisfying for the players who go through the game [6]. In the communication part of the game, by default, only the position of the colder squares can be disseminated and the answers cannot be announced directly, which makes the whole process more mysterious and enhances people's desire to explore [7]. It not only adds a lot of fun to people's boring family life, but also enriches people's vocabulary and improves people's thinking ability, so it was warmly welcomed upon its introduction. It not only adds a lot of fun to people's boring family life, but also enriches people's vocabulary and improves their thinking ability, so it was warmly welcomed once it was introduced. We assume that the vocabulary of an individual has no effect on the final number of guesses in the game.

2. The basic fundamental of analysis models

Before we begin modeling, we first need to ensure the accuracy and availability of the data. The data sample chosen for this paper provides basic data on game statistics: date, number of tournaments, vocabulary per day, number of people reporting scores, number of hard-modeled players, and number of attempts. The data is small but noisy. In this example, the words guessed in the wordle game all consisted of 5 letters, so the data for words that did not consist of 5 letters was removed, i.e., the data for words such as "rprobe, clen, and tash" was removed to prevent erroneous data from adversely affecting subsequent modeling.

In order to facilitate clustering analysis of words with different difficulty levels and correlation analysis of the number of people with different registration difficulties, we extracted word attributes from the words. However, there are various methods for extracting word attributes, including the commonness of the word, the length of the word, the semantic range of the word, the structure and spelling of the word, and so on. As a result, the extracted information is characterized by confusion and complex logical relationships, which are not suitable for analysis. Data coding can solve this problem well. Data coding is a commonly used technique to convert class variables to decimal, an example of which is shown in Appendix [1].

The specific steps for applying data encoding in the word vector domain are as follows:

Step 1: Count all the words contained in the data, and then number each word one by one from a to z. If the word is an adjective, it will be counted as an adjective, and if it is an adjective, it will be counted as a word. Meanwhile, if the word is an adjective, it is labeled as [1,0,0], an adverb as [0,1,0], and a noun as [0,0,1].

Step 2: Create an 8-dimensional vector for each word. Each dimension of the vector represents a letter and the word composition of its part of speech. Define A as 1, b as 2 ..... . z is defined as 26, and arrange them in order to form an ordered array.

For example, the word [manly] in the text is [8 1 13 7 25]. The advantages of data encoding are undoubtedly enormous. It solves the problem of discrete data in classification that is difficult to deal with, takes into account the order of words and words, and greatly expands the operation space of data. However, its disadvantages are also obvious. First, it is a bag-of-words model. Second, it assumes that words are independent of each other (whereas in most cases, words interact with each other). Finally, the features obtained are relatively discrete and sparse, which is not conducive to further statistical analysis.

Overall, data coding is a simple, effective and easy-to-handle technique for converting categorical variables into decimal numbers.

Overall, data coding is a simple, effective and easy-to-operate technique for converting categorical variables to decimal numbers, which helps to improve the performance and accuracy of the model. Based on the above method, all the words are coded and extracted for subsequent use.

Firstly, we draw a scatter line graph of the data distribution, observe the trend of the number of people reported by wordle game over time, and make a general trend analysis based on the trend and

the general law. The resulting discount graph is shown in Figure 1.

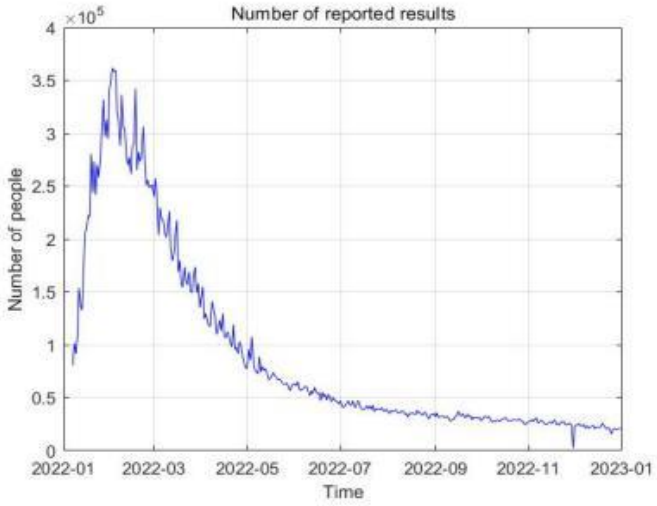

Figure 1: Line chart of the number of people changing over time

Figure 1 above shows the evolution of the number of reports over time. It can be seen that wordle game showed a clear trend of steep increase in January, reached a high level in February and March, and then continued to decline, with the downward trend gradually slowing down. From August onwards, the data tends to be consistent, maintaining a level of around 20,000 or 30,000 players. The distribution of data can also be seen in the box plot below. Mainly within 40,000, the data shows a relatively discrete distribution around 4-13w.

Observations from the first phase show a general upward trend in numbers, with numbers growing over time. As time continues to grow, the number of people generally shows an upward trend, and the upward trend is uncertain. Combined with the rising magnitude of the curve, we hypothesize that the time range may be linear, polynomial, exponential, etc. In the second stage, we can conclude from the observation that the overall number of people shows a decreasing trend with the change of time. Combined with the downward trend of the curve being faster and then slower, we speculate that the time range may be a polynomial curve, an exponential curve, a power curve, and so on.

The short-term memory network unit is another module in the RNN. From an abstract point of view, LSTM stores the long-term dependent information of the text. In traditional RNN networks, the hidden layer is very simple. This simple structure cannot effectively link historical information together, but LSTM can record and learn historical information very well [8].

It is called a storage block and contains a storage unit and three main gates, namely the storage gate, the input gate and the output gate. The horizontal line at the top is the unit state, which controls the transfer of information to the next moment. The two tanh levels in the figure correspond to the inputs and outputs of the cell. From the figure we can see that the flowchart of LSTM consists of three main steps [9]:

Step 1: The "oblivion gate" layer is controlled by the sigmoid function, which determines which information can pass through the cell state. The "oblivion gate" layer will pass or partially pass the output of the previous moment.

Step 2: Generate the required new information. This step consists of two parts: the first part is the "input gate" layer, which decides which values need to be updated by means of the sigmoid function, and the second part is the tanh layer, which generates new candidate values and adds them to the previous ones to get the final candidate value for that part. By combining these two steps, new information is added and unwanted information is discarded.

Step 3: Obtain the model output. First get the initial output from the sigmoid layer, then use tanh

to scale that value to between - 1 and 1, then multiply the output with the sigmoid pair to get the final output of the model.

By constructing a well-developed deep learning, machine learning models have good results in problems such as regression, but usually there are many hyperparameters in the model. The setting of variable parameters can directly or indirectly affect the performance of the model [10].

Model hyperparameter optimization generally has six steps:

1) Construct the corresponding model and determine the search space.

2) Select hyperparameters from the search space.

3) Apply the hyperparameter combination to the model, train the data, and evaluate its performance on the verification data.

4) Select the next set of hyperparameter combinations according to the evaluation results.

5) Repeat steps 2-4 until the number of iterations or the specified time is reached.

6) Finally evaluate the performance of the model on the test data.

PSO algorithm is used to optimize LSTM related parameters, and the parameters to be optimized by LSTM are input into LSTM as particle swarm. At this time, LSTM is the optimal model optimized by PSO, and its prediction accuracy is improved. This process is not affected by human parameter adjustment factors, and the randomness is low, so the optimization is relatively stable

The main idea of improving the PSO-LSTM model is to use the good parameter ability of PSO algorithm to optimize the relevant parameters of LSTM and improve the prediction effect of LSTM.

3. Results

3.1 Data sources

A sample of data analyzed by our team was taken from the website : https://www.mathmodels.org/Problems/2023/MCM-C/index.html.

3.2 Analysis of experimental results

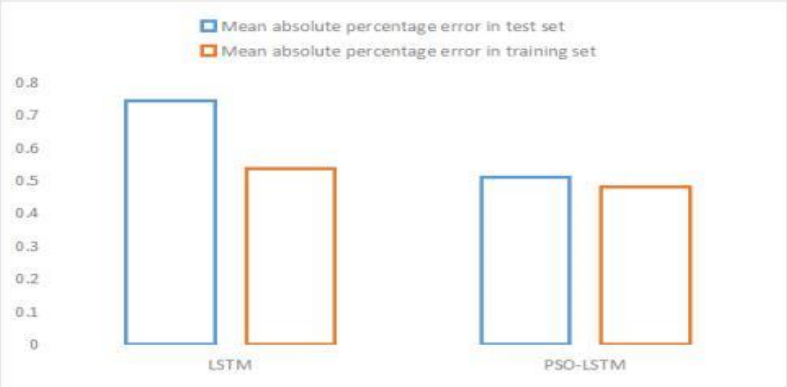

Figure 2: Contrastive analysis chart

Our team visually compared the analysis of the results before and after applying PSO optimization using bar charts, as shown in Figure 2. It can be clearly seen that after applying the PSO algorithm for optimization, the results of both the training set and the test set have been significantly improved, and the results of the model have also been significantly improved. Here, the EERIE data should be coded first.

Substituting the correlation percentage of EERIE words (1,2,3,4,5,6, Y) into the model, the following prediction results can be obtained, as shown in Table 1.

Table 1: Predicted value

distribution	1try	2tries	3tries	4tries	5tries	6tries	7tries or more
Forecast result	0.432	3.631	19.326	30.291	26.954	14.234	5.132

With the composite score level of the data processing part as the dependent variable, the data were divided into training set and test set, in which the test set accounted for 30%. After the model training is completed, the confusion matrix of the model classification results is shown in Figure 3 above, and the model performance parameters are calculated based on the prediction results of the test set. It can be seen that the model results are excellent.

Based on the decision tree model of automatic optimization genetic algorithm, it can be obtained that the prediction probability of EERIE is 0.07648993, the medium prediction probability is 0.74494204, and the difficult prediction probability is 0.1895767. Therefore, the difficulty of EERIE is in the medium level.

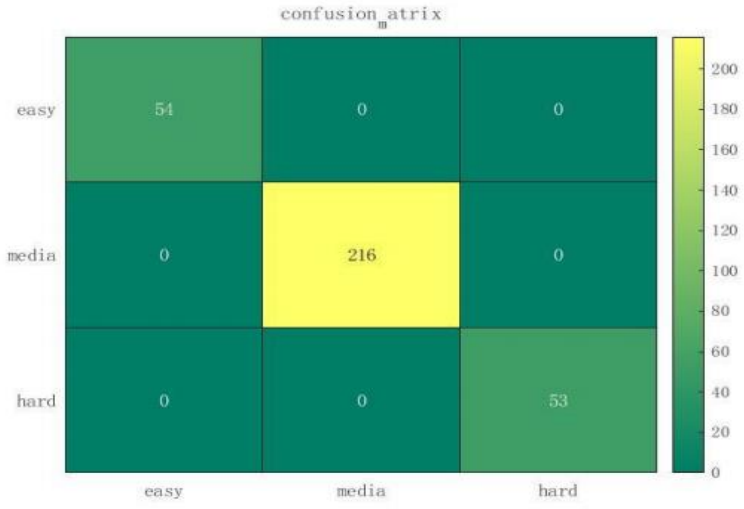

Figure 3: Thermodynamic analysis

4. Conclusions

Once the prediction is complete, we want to measure the difficulty of the word. We use Spearman's correlation coefficient to analyze the correlation between the number of attempts by different players and use entropy weighting to establish a quantitative index of difficulty based on the previous data. Word difficulty is categorized into three levels: easy, medium and difficult. Through modeling, we can get the probability that EERIE belongs to easy is 0.0764, the probability that it belongs to difficult is 0.1896, and the probability that it belongs to medium is 0.7449, which means that the word belongs to medium difficulty level.

We also found an interesting result. We found that the percentage of people who passed the game 1, 2, and 3 times is negatively correlated with the percentage of people who passed the game 5, 6, and x times. This suggests that the ability to pass the game the fourth time has a significant impact on the difficulty of the game. We also analyzed the relationship between good senior players and the number of times they completed the game, and concluded that the more informative the word, the lower the probability that a player will attempt to pass the game.

References

[1] Alexander B. Instructional Strategies for Analyzing Media Literacy in the Classroom[J] Texas Speech

Communication Journal. 2016.
[2] TANG YF, CAI Y, ZHOU Shuai, WANG Member, YANG Zelin, CHEN Xinghong. Research on fault diagnosis method of rotating machinery based on PSO-BP-Adaboost [J]. Journal of Sichuan University of Light and Chemical Engineering (Natural Science Edition), 2023, 36(04):26-33.
[3] TIAN Jie, LI Yang, ZHANG Lei, LIU Zhen. Adaptive control of temporary support force based on PSO-BP neural network [J]. Industrial and Mining Automation, 2023, 49(07):67-74.
[4] Deng Yuxuan. Research on cost prediction of urban railroad based on adaptive PSO-BP neural network[J]. Construction Economy, 2023, 44(S1):92-96.
[5] SHI Peilong, CHANG Hong, WANG Cairui, MA Qiang, ZHOU Meng. Research on path tracking control of driverless car based on PSO-BP optimized MPC[J]. Automotive Technology, 2023, (07):38-46.
[6] Cai Hairong. Research on cold chain logistics service provider selection based on rough PSO-BP neural network[J]. China Business Journal, 2023, (13):71-74.
[7] Nick Smurthwaite. OBITUARIES: Irving Wardle[J]. The Stage, 2023, (10).
[8] Ross Norbert. the intersection of violence and early COVID-19 policies in El Salvador. [J]. American anthropologist, 2022, 124(3).
[9] Khosravifar B , Alishahi M , Bentahar J ,et al.A Game Theoretic Approach for Analyzing the Efficiency of Web Services in Collaborative Networks[C]//IEEE International Conference on Services Computing.IEEE, 2011.DOI: 10. 1109/ SCC.2011.76.
[10] Alfaro Karla, Soler Montserrat, Maza Mauricio, Flores Mauricio, López Leticia, Rauda Juan C. , Chacón Andrea, Erazo Patricia, Villatoro Nora. Mumenthaler Eveline, Masch Rachel, Conzuelo Gabriel, Felix Juan C., Cremer Miriam. Cervical Cancer Prevention in El Salvador: Gains to Date and Challenges for the Future[J]. Cancers, 2022, 14(11).