The Application of Data Analysis in A Web-Based Word Game: Worldle

Zien Zhu

School of Finance, Jiangxi University of Finance and Economics, Nanchang 330000, China

zhuzien210926@163.com

Abstract. We will analyze the game Worldle based on its data in 2022. First, we use the EEMD Model to analyze the trend of the number of players over time in this year.We need to forecast the number of players in the next 60 days and reduce the prediction error, we use the Moving Average and build Rolling Window to reduce the forecast error. We use both LSTM model and EEMD-LSTM for training and comparing the effects of the two and find EEMD-based LSTM model is better. Then, we come up with several properties of the solution words that we think would affect the difficulty of the puzzle, and after Correlation Analysis, we find that this is indeed the case and give some explanations. **Second**, we mainly use Deep Learning. To predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a given future solution word on a future date, we use a Genetic Algorithm and Back Propagation Neural Network (GA-BP Neural Network) to analyze the percentage distribution of given words. After that, the distribution of the number of answers for the example can be obtained by prediction as (1.1831,6.2839,21.3307,28.6500,24.7627,14.7354,3.8857). **Third**, how difficult is it to guess a word correctly? For this problem, we use a method belonging to Machine Learning - PSO Decision Trees. Before this, our group first used RSR and the Pearson Correlation Coefficient to divide the difficulty level of words.

Keywords: EEMD-LSTM Model, Moving Average, GA-BP Neural Network, RSR, Decision Tree.

1. Introduction

We can obtain the total daily number of people who play the game, who play the hard mode, and the score percentage distribution of the daily puzzles from game results on Twitter every day. We need to analyze and predict the number of players, score distribution (Develop a model to predict score percentages of (1, 2, 3, 4, 5, 6, X)), attributes, difficulty, and so on. we will predict the condition for the word EERIE on March 1, 2023 and analyze it.

2. Analyzing Variation in the Number of Reported Results

Based on the original data, we get the following picture by EEMD.

● Trend component: Trend of development

The correlation coefficient between trend item and original sequence is 0.89, and variance ratio to original sequence is 51.6%, higher than other IMFs, indicating that the trend item is the most important part in the process of the number of reports changing. The trend reflects the change trajectory. When Wordle is in heyday, its trajectory is above the trend item; when it's in infancy, its running track is below the trend item. This reflects decline in enthusiasm of Wordle.

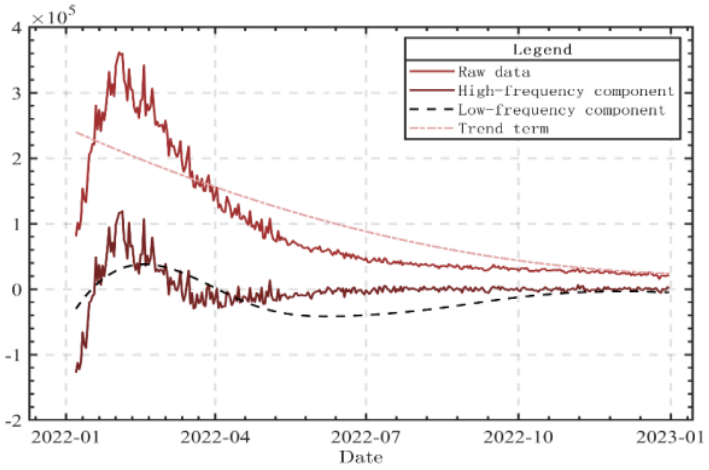

Fig. 1 Time series of number of players

● Low-frequency component: The influence of main factors

The variation fluctuation is mainly caused by low-frequency components. The correlation coefficient between the low-frequency component and the original sequence is 0.73, the variance ratio to the original sequence is 6.8%, and the average period of the low-frequency IMF is 179.5 weeks, showing that the impact of low-frequency items on people's changes is difficult to eliminate in a short time. We think it's variation of trend in time due to these reasons: People's enthusiasm was high in the previous quarter, but dropped significantly in the second quarter, and the subsequent recovery may be due to "I haven't played it for a long time and want to play it again". Besides, we suspect the difficulty of solution words also plays a role in the trend.

● High-frequency component: Random fluctuation

The coefficient between the high component and the original sequence is 0.56, and the variance ratio to the original sequence is 9.5%. It can be seen that the average period of high-frequency components is relatively short, which reflects the randomness of population changes and is affected by factors such as psychological fluctuations and nature. Seen from Fig.1, it's a curve with high volatility. Intuitively, this can explain that it has a large variance, which reflects its large volatility.

3. Forecasting through EEMD-LSTM and Discussing Attributes That Affect the Score

We both use LSTM and EEMD-LSTM to compare results and find the latter is better. So we choose EEMD-LSTM model to predict the number of March 1st in the future. We get a predicted value of 27327.36 for March 1, 2023. According to the prediction error during training, we get the prediction interval of this day as (25339, 29316), found that the low-frequency items have a larger increase in subsequent predictions. The reason may be that some users who are tired of playing have started to play again, similar to infectious diseases. People in incubation period began to rise largely.

Tab. 1 Conditions of different methods

	MAPE	RMSE	MAE	R2
LSTM	0.1069	25305.562	12651.664	0.9111
EEMD-LSTM	0.0811	24037.0895	10838.385	0.9272
Improvement rate	0.2413	0.0501	0.1433	0.0161

We determined four features: the frequency of letters *Lf*, the frequency of daily use of five-letter words *Wf*, common letter pairs *LPf*, common ligature letter pairs *LLPf*, arguing that these features have impacts on the distribution of score percentages in the game.

There are some interesting points. First, *Lf* is obtained through the following process. If we want to use *Lf* as an attribute of the solution word for analysis, we need to consider that: Wordle uses 5-

letter words. Therefore, we first use / usr / share / dict / words as the etymology, and get the 24-letter frequency table corresponding to the 5-letter word, as shown in the left of Tab.2.

But, is Wordle using the source we mentioned earlier as the source of its questions? Using / usr / share / dict / words as an resource is convenient, but we doubt these words may not the source of the vocabulary in Wordle. Won't word game players use a dictionary, like Collins? For example, plurals of four-letter words are also accepted by Wordle. So we found two sites *wordfind.com* and *www.allscrabblewords.com*, and obtained relevant information. After several searches and replacements, corrections were made to the original frequency. The following list of letter frequencies for five-letter words was obtained, as shown in Tab.2. We take the right of Tab.2 as the final letter frequency of 5-letter words.

Tab. 2 Left: / usr / share / dict / words

Right: from 2 webs: wordfind.com and www.allscrabblewords.com

letter	Frequency(%)			letter	Frequency(%)		
a	**10.5**	h	3.1	**s**	**10.4**	p	3.1
e	**10**	m	3.1	**e**	**10.3**	m	3
r	**7.2**	p	3	**a**	**8.9**	h	2.7
o	6.6	b	2.7	o	6.7	g	2.5
i	6.1	g	2.6	r	6.5	b	2.5
s	5.6	k	2.1	i	5.9	k	2.1
t	5.6	w	1.6	l	5.5	f	1.8
l	5.6	f	1.6	t	5.2	w	1.5
n	5.2	v	1.6	n	4.5	v	1.1
u	4.4	z	1.1	d	3.9	z	0.6
y	3.8	**x**	**0.6**	u	3.8	**x**	**0.5**
c	3.6	**j**	**0.4**	c	3.3	**j**	**0.4**
d	3.3	**q**	**0.2**	y	3.1	**q**	**0.2**

Second, the frequencies of the five words use the frequency ratings from the Corinthian dictionary website. We divide the frequency of word usage into five grades (1~5), the higher the grade is, the more frequently the word is used.

Third, the *LPf* and *LLPf* are derived from Herbert S. Kim's work *Codes and Secret Writing*. Common letter pairs are TH HE AN RE ER IN ON AT ND ST ES EN OF TE ED OR TI HI AS TO, and common ligature letter pairs are LL EE SS OO TT FF RR NN PP CC.

Based on the four attributes proposed above, we analyze the correlation of these four attributes with the scores of words. The results are shown in Tab.3. Since these four attributes have no correlation with 7 or more attempts, we no longer put them in Tab.3.

Tab. 3 the frequency of letters Lf, the frequency of daily use of five-letter words Wf, common letter pairs LPf, common ligature letter pairs LLPf

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries
Wf	0.068(0.201)	0.117(0.027**)	0.083(0.116)	-0.058(0.277)	-0.113(0.033**)	-0.061(0.251)
Lf	0.211(0.000***)	0.363(0.000***)	0.255(0.000***)	-0.144(0.006***)	-0.243(0.000***)	-0.189(0.000***)
LPf	0.166(0.002***)	0.313(0.000***)	0.254(0.000***)	-0.15(0.004***)	-0.314(0.000***)	-0.152(0.004***)
LLPf	-0.116(0.028**)	-0.228(0.000***)	-0.211(0.000***)	0.09(0.090*)	0.254(0.000***)	0.108(0.042**)

Excluding the first attribute, the number of times required to correctly answer the latter three attributes shows a strong correlation. Of course, the first attribute is also highly correlated with 'twice' and 'five', and it is positively correlated with the former and negatively correlated with the latter. Judgments made on the frequency of words, meaning that words used more often tend to come to mind. The reason why there is a negative correlation with "try five times to get it right" is that we guess that it may be because once the user has not thought of using this word for the first few times, the probability of subsequent thought will be very small, which leads to the emergence of negative correlation. Then, observing the following three characteristics, it’s easier for us to conclude these three attributes are very correlated with '1 try' to '6 tries'.

Secondly, it can be seen that both 'letter frequency' and 'common letter pairs' are positively correlated with '1 try', '2 try' and '3 try', and negatively correlated with the latter three attempts. However, common ligature pairs, by contrast, were negatively correlated with the first three and positively correlated with the last three. This is an interesting phenomenon to think about, and it may be because people are more familiar with the frequency of letters in five words and the common letter pairs (many letter pairs are some common combinations of consonants and vowels, and two consonants Letters can form certain combinations of consonants). According to this, people mostly guess words through these attributes. On the contrary, although there are commonly used ligature letter pairs in daily life, for Wordle, one less letter often makes players uncomfortable, which may make it difficult to think of using a ligature letter pair instead.

Beyond that, some specific letters have striking correlations. Letter 's', which has extremely strong significance for the second and third guesses. Players tend to like to guess the plural forms of some words first try, and the prototype of these words is often composed of 4 letters. Wordle recognizes plurals of words.

4. Multi-input multi-output neural network prediction

Seen from Tab.4, the mean value of MAPE of the improved BP neural network model with GA is significantly better than that before the improvement.

Tab. 4 BP vs GA-BP

Bp Neural Network			GA-BP		
	train	test		train	test
1 try	0.964	6.406	1 try	0.637	2.581
2 tries	0.374	0.384	2 tries	0.023	0.596
3 tries	0.201	0.202	3 tries	0.002	0.276
4 tries	0.11	0.132	4 tries	0.001	0.154
5 tries	0.149	0.196	5 tries	0.001	0.226
6 tries	0.379	0.551	6 tries	0.002	0.6
7 or more tries (X)	1.104	2.129	7 or more tries (X)	0.08	1.795
mean	0.468714	1.428571	mean	0.106571	0.889714

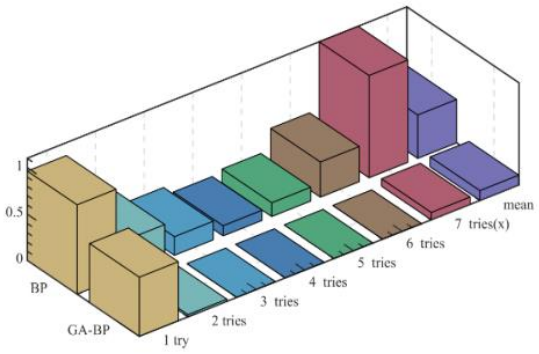

Fig. 2 MAPE of models of train

It's also clear from this 3D histogram that the improved model using the genetic algorithm predicts better results. As GA-BP results are better, we finally chose a GA-BP neural network model for prediction of 'eerie', which corresponds to a word vector of. The results are **(1.1831,6.2839,21.3307,28.6500,24.7627,14.7354,3.8857)**

Based on the errors in the training and test sets, the goodness of fit, i.e., the certainty that the model predicts accurately, is 99.737%, and the
uncertainty is 0.263%.

5. Comprehensive score and classification of difficulties based on Spearman and RSR

- Spearman correlation analysis

By analyzing the correlation coefficient of Spearman correlation percentage (2,3,4,5,6, X), the difficulty of words can be measured. As 1 try is caused by chance, this variable is removed. It can be seen that the forward slash in the lower left corner of the matrix indicates that the two variables in the cell are positively correlated, whereas the forward slash in the lower left corner indicates that the two variables are negatively correlated, and the darker the color indicates that the correlation is stronger. Green indicates positive correlation, while yellow indicates that there is a negative correlation.2 tries and 3 tries are positively correlated, while they are negatively correlated with percentage.

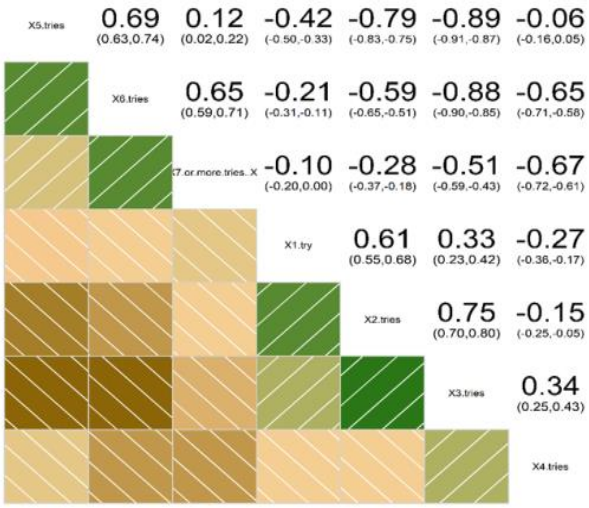

Fig. 3 Spearman correlation of tries'

of 4-X, so we can divide 2 tries and 3 tries as negative indicators, that is, the higher the percentage, the lower the difficulty of the word. 4 tries-X is a positive index, that is, the higher the percentage, the higher the difficulty of the word. The entropy weight method is used to calculate the weight of each index, and the results are shown in Tab.5.

Tab. 5 The weight of each index

items	Information entropy value e	Information utility value d	weight (%)
5 tries	0.985	0.015	9.027
4 tries	0.994	0.006	3.472
7 or more tries (X)	0.903	0.097	59.089
6 tries	0.966	0.034	20.508
2 tries	0.996	0.004	2.388
3 tries	0.991	0.009	5.517

- rank-sum ratio comprehensive evaluation method (RSR)

We obtain dimensionless statistic RSR through rank conversion, apply the concept and method of parametric statistical analysis, study the distribution of RSR, and rank the difficulty of words according to the RSR value.

The formula for calculating RSP is:

$$RSR_i = \frac{1}{mn}\sum_{j=1}^{m} R_{ij}$$

Where n is the number of words 358, m is the word attribute 30, i= 1,2... ,n; J = 1, 2,... m; R is the rank of the element in row i and column j.

Tab. 6 file sort threshold table

File sort threshold table			
grade	Percentile threshold	Probit	RSR critical value (fit value)
1	<15.866	<4	<0.1103
2	15.866 ~	4 ~	0.1103 ~
3	84.134 ~	6 ~	0.2902 ~

Probit value (the probability unit corresponding to the cumulative frequency) was taken as the independent variable and the RSR value as the dependent variable to calculate the linear regression equation and fit the corresponding RSR estimate. The fitting equation is as follows:

$$Y = 0.249 + 0.09 * \mathrm{Probit}$$

The difficulty is divided according to Tab.6: Level 1 is easy, level 2 is medium, and level 3 is difficult:

6. Decision tree model training classification based on PSO tuning

Tab. 7 Parameters of PSO

Parameters	optimal value
Max_depth	5
Min_samplies_split	2
Min_samplies_leaf	1
Min_weight_fraction_leaf	1
Min_impurity_decrease	0.1
Ccp_alpha	0.2
Learning_fate	0.15
Max_leaf_nodes	0

The data is split into a training set and a test set, where the test set accounts for 20%, and the optimal parameters obtained after training with the PSO-decision tree model are shown in Tab.7.

The two-node partition path of the DTM is used to obtain the threshold of word difficulty partition.

Based on the DTM obtained from training and inputting the attributes of EERIE, we can get the probability that EERIE is predicted to be easy is 0.1538, the probability that it is predicted to be medium is 0.7338, and the probability that it is predicted to be hard is 0.1124. Therefore, the difficulty of EERIE is considered to be medium.

References

[1] Lei Y, He Z, Zi Y. Application of the EEMD method to rotor fault diagnosis of rotating machinery [J]. Mechanical Systems and Signal Processing, 2009, 23(4): 1327-1338.

[2] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling [J]. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6): 275-285.

[3] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures [J]. Neural computation, 2019, 31(7): 1235-1270.

[4] Cohen I, Huang Y, Chen J, et al. Pearson correlation coefficient [J]. Noise reduction in speech processing, 2009: 1-4.

[5] Cao W,Fan R. Performance evaluation analysis of city business firms based on rank-sum ratio comprehensive evaluation method [J]. Academic Journal of Business & Management,2023,5(11).

[6] Hui C,Yulin W,Chuanwang S, et al. Prediction of Surface Subsidence Based on PSO-BP Neural Network [J]. Journal of Physics: Conference Series,2022,2400(1).