Result Prediction of Wordle Word Guessing Game Based on SVM And BP Neural Network Multi-Input-Multi-Output Regression Prediction Model

Ruichen Liu [1, *, #], Yinxi Li [1, #], Mengdan Li [2, #]

[1] Jinan University - University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511436

[2] College of Cyber Security, Jinan University, Guangzhou, China, 511436

* Corresponding Author Email: Yinxi.Li0122@outlook.com

[#]These authors contributed equally

Abstract. Wordle is a popular daily puzzle game in which players can guess a five-letter word in six or fewer attempts. In order to predict the distribution column of user participation and guessing results, this paper established a support vector machine (SVM) regression model based on the historical data set of game results to predict the number of reports on March 1, 2023. Then, by extracting attribute features such as word frequency from words, the relationship between word attribute and report quantity under difficult mode is studied through multiple regression analysis and correlation coefficient analysis. Finally, a BP neural network multi-input-multi-output regression prediction model was established. The goodness of fit (R) of the model was 0.93671, and the distribution of correct guesses of the word "EERIE" was predicted. Through this regulation, the popularity of the game can be maintained to some extent, so that the number of people playing the game can maintain stability or growth.

Keywords: Supported Vector Machine, BP Neural Network, Wordle.

1. Introduction

1.1. Problem Background

Wordle is a free web crossword puzzle written by British engineer Josh Wardle to pass the time during the COVID-19 pandemic. Players have to guess a five-letter word in six chances. The correct word is shown with a green color, a yellow color means the word contains the letter but is in the wrong place, and a gray color means the letter is not included in the word. After guessing correctly, players can share the completion time and number of guesses on social media platforms and compete with their friends on the speed of guessing words.

Wordle became an instant hit in early 2022, with millions of players a day. On February 1, 2022, The New York Times announced its acquisition of Wordle, and in October 2022, it partnered with toymaker Hasbro to release a board Game called Wordle: The Party Game. Google announced its list of the world's most-searched games in 2022 on December 7, 2022. The results showed Wordle at the top of the list as the most searched word of 2022, beating popular search terms such as "World Cup," "the late Queen Elizabeth," and "iPhone 14"."Everyone is obsessed with the game, and you can see it in the numbers." Simon Rogers, head of Google's Trends data team, told the press.

Based on the popularity of Wordle, this paper predicts the number of Wordle users and studies the relationship between word attributes and the number of reports in difficult mode.

1.2. Literature Review

In order to predict the number of reports and understand the relationship between word and the number of reported results, it is necessary to evaluate the difficulty of word quantitatively. Domestic and foreign scholars have carried out a great deal of research on the difficulty of English words. In previous studies [1], it shows how to distinguish students' performances on vocabulary assessments as a means of understanding what contributes to the ease or difficulty of vocabulary knowledge. In

addition, the model in research [2] evaluated word familiarity rather than word length as a stand-in for word difficulty. The way they calculate the specific difficulty factor parameters can reflect the difference of the degree of disharmony between different words, and thus directly affects the rationality of the final difficulty value of words. The defects in the definition of vocabulary difficulty will inevitably lead to the cognition of vocabulary attributes and the classification of vocabulary.

Meanwhile, a study [3] shows a novel method for using agent experiences gathered through an embodied simulation to ground contextualized word vectors to object representations. Based on this, on the basis of literature review, this study proposes a specific method to quantify the difficulty of words to make it more scientific and practical and shows the quantification process by taking the words in the game given in the 2023 MCM/ICM as an example.

In addition, SVM is mainly used for the learning rules of small sample data, which can obtain good generalization ability under the condition of limited samples [4], the combination of SVM model, BP neural network can effectively complete the model establishment of the number of reported results, word difficulty and the correlation between them in the future.

1.3. Our Work

First, we build support vector machines (SVM)[5] to explain changes in reported results and create a forecast range for the number of results reported on March 1, 2023. And the relationship between any attribute of the word and the reported percentage in difficult mode is verified.

Next, we build BP-neural network model to predict the distribution of attempt times (1,2,3,4,5,6, X) in the future day by word attributes.

2. The Data Pre-processing

The data is from the 2023 MCM/ICM.

First of all, with a box graph for the number of reported results to characterize the original data distribution[6], which showed the data of median, standard deviation, maximum, minimum, upper and lower bounds of the box graph, as shown in Figure1 and Table 1.

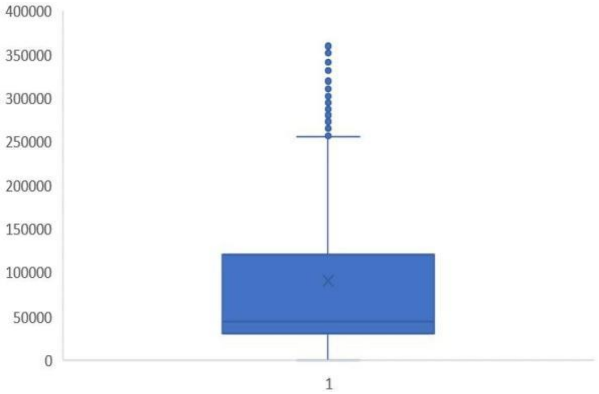

Figure 1 Box type diagram 1

Table 1 Statistics 1

Index	Value
Mean	90918.877437
Std	89274.374730
Min	2569.000000
25%	30308.500000
50%	44578.000000
75%	120294.000000
Max	361908.000000
Upper bound	104669.75
Lower bound	165286.75

Thus, there are 36 outliers in the number of reported results. For the deleted data, in order to maintain the integrity of the data and avoid the impact of datareduction on the accuracy of the model, we used Lagrange interpolation to complete the deleted data, and the results are shown in Table 2.

Table 2 Original and supplementary values of outliers of Number of reported results

Date	Original value	Lagrange interpolation supplementary values
2022/11/3	2569	21388
2022/3/28	173696	143266
2022/3/27	165468	147173
2022/3/24	169066	151732
2022/3/21	173636	161526

By the same token, next we draw a box diagram for the Number in hard mode and deleted the abnormal data.

3. Supported Vector Machine

3.1. Selection of indicators

The most obvious factor when considering the number of attempts is the difficulty of the word. There are many factors that affect the difficulty of a word, such as word length, word frequency, the ratio of syllables to word length (which is called word harmony in this article), the presence or absence of repeated letters, and so on. Since the game gives feedback for getting the right letter, the number of repeated letters in the word is chosen as an indicator and called it repetition. Since a word with fewer syllables is likely to be easier to pronounce, which also means the word is easier to think of and easier, the number of syllables are chosen as the word attribute and called it harmony. And part of the syllable count results are shown in Table3.

Table 3 Number of repeated letters and syllables in words

Date	Word	Repeat letter count	Syllable count
2022/12/14	usual	1	2
2022/12/13	spoke	0	1
2022/12/12	apply	1	2
2022/12/11	naive	0	1
2022/12/10	knock	0	1

3.2. Building of the model

The basic idea of Support Vector Machines(SVM for short) is to find the separation hyperplane with the largest interval in the feature space of a data dimension for efficient binary classification of target data. The maximum interval makes it different from perceptron, which not only effectively solves the dimension disaster problem, but also has good scalability.

The number of syllables in the word and the number of repeated letters in the word into the model, so that different predicted values can be obtained by changing the number of syllables in the word and the number of repeated letters in the word to form the interval.

Finally, the classification decision function (1) is obtained:

$$f(x) = \text{sign}((-1.9398, -0.3244, -0.1383, 0.1063)^{T} \cdot x + (0.0.9011, 0, 0, 0)^{T}) \quad (1)$$

70% of the original data were randomly selected as the training set and 30% as the test set to train the model. After completion, the date was re-inserted into the model to obtain the predicted value of the Number of reported results from January 7, 2022 to December 31.

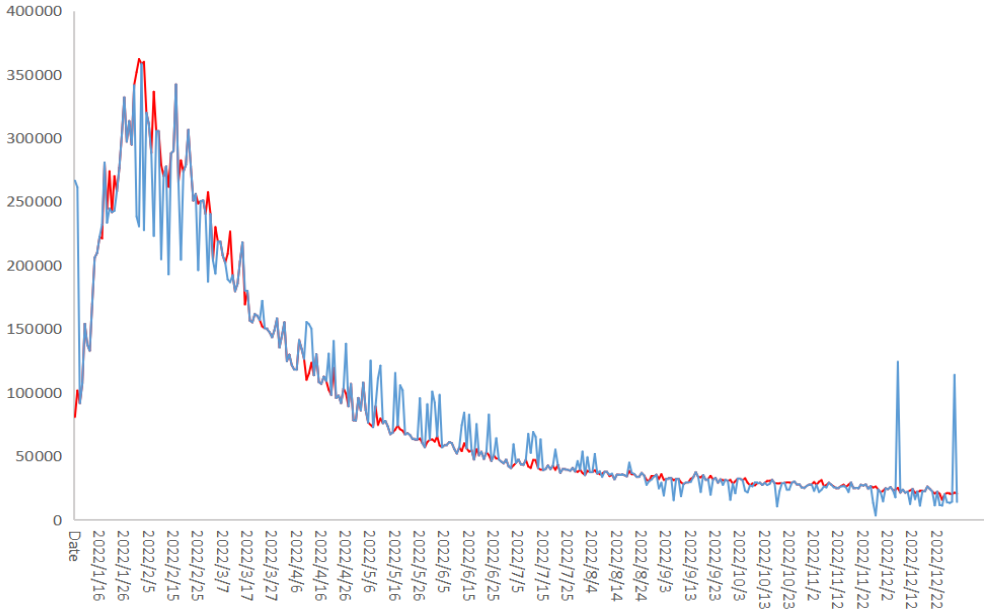

Figure 2 Raw versus SVM predicted values for the Number of reported results.

It can be seen from Figure 2 that the predicted value of the model has a good fitting effect on the original data, but there is a large deviation in the predicted value in January-February and December, which may be due to the fact that these two periods are at the beginning and end of the data, and there is no more data before and after to assist the model to judge it, leading to the first error larger than the data in the middle part.

Table 4 Error statistics

	max	average	count
Error	185805.3166	7901.29054	109

Comparing Table 4, the maximum error of the forecast occurs on January 7, 2022, the first day of data, which may also be due to the absence of earlier data to assist the model prediction.

3.3. Predicted results

Through the establishment of the SVM above, the Number of reported results on March 1, 2023 are shown in table5.

Table 5 Predicted intervals for the Number of reported results on 1 March 2023

Date	Minimum	Maximum
2023/3/1	13158.43325	25908.2883

3.4. Relevance determination

In addition, since people tend to input common words they know, the word frequency are chosen as word attribute and call it word frequency, which comes from the word frequency queried in the dictionary (per million)[7]. Due to space limitation, only 5 pieces of data are shown in Table 6. Here, the proportion of the first letter appearing times of the answer words in all days are used as the word attribute. And the multiple regression method was used to analyze the word.

Table 6 Lexical frequency of words

	waltz	libel	sneak	carry	flout
Word frequency	174	800	236	9198	59

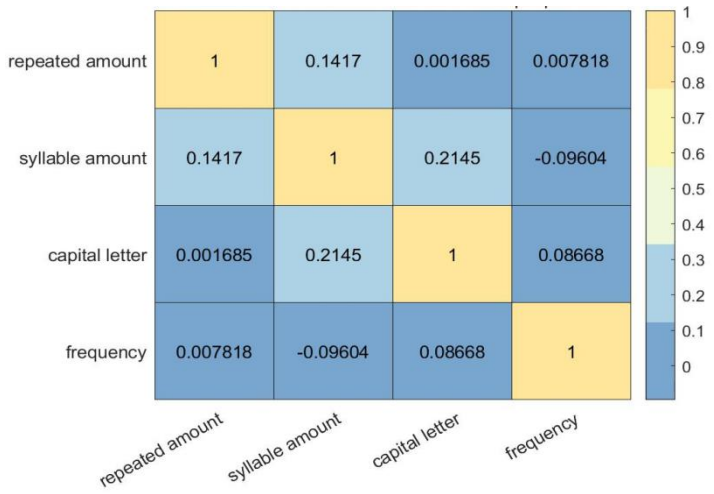

Figure 3 Correlation coefficient matrix

From Figure3, it can be found that among the four factors, their Pearson correlation coefficients are all lower than 0.3, indicating that their linear correlation is very weak.

Table 7 Variance inflation factor

	Degree of repetition	Commonness	Initial ratio	Harmony
VIF	1.022	1.022	1.062	1.086
Tolerance	0.978	0.978	0.941	0.921

The VIF of the regression equation is <1.1, and the tolerance values are all greater than 0.9, are shown in Table7, which indicating weak collinearity of the regression equation. It follows that word attributes do not affect the percentage of difficulties reported[8].

4. Attempts distribution prediction by BP-Neural network

4.1. Selection of indicators

Given a certain term, predict the distribution of attempts (1,2,3,4,5,6,X) on the future day. The attempt distribution is mainly related to the attributes of the word, the date, as well as the number of reported results and number in hard mode.

The properties of words can determine the difficulty of the game, that is to say, they have an impact on the distribution of attempts. If the word is more difficult, the proportion of more attempts will increase, and vice versa. For word attributes, word repetition, word harmony, word commonness, and first letter association are selected.

Initial association is the proportion (%) of the word with the same letter in all English words. Due to space constraints, only 5 pieces of data are shown in Table8.

Table 8 Proportion of initial letters in all English words (%)

	A	B	C	D	E
Capital letter association	7.8	1.75	2.56	4.5	12.5

For the case of a given date, it is also introduced that the contest number, number of reported results and number in hard mode on that day as the basis for predicting the distribution of the number of attempts. And the data of 5 words is shown in table 9.

Table 9 Overview of input data of neural network model

Word	Contest number	Number of reported results	Number in hard mode	Word repetition	Word harmony	Capital letter association	Word frequency
study	529	21388	2405	0	2	6.4	21314
undue	528	23739	2316	1	1	3	486
tepid	527	26051	2484	0	2	8.7	79
happy	526	25206	2356	1	2	6.1	10266
clean	525	26381	2424	0	1	2.56	6123

4.2. The Establishment of BP-Neural network

BP-neural network [9] is an artificial neural network with back propagation function, which is composed of input layer, output layer and hidden layer. The hidden layer often has several layers [10].

The dimension of input data of BP neural network model is 7, which are respectively used as input data set and target data set, forming a multi-input-multi-output model. After parameter adjustment and repeated training, the number of hidden layers is set to 9, and the model is obtained.

At the same time, in order to prevent over fitting of the model, sample data will be randomly divided, and the ratio of training, testing and verification data is 70:15:15. The mean square error(MSE) is used to measure the network performance. When the number of training reaches 1000, the iteration stops.

The model fitting accuracy is shown in Figure4 and Figure5. The parameters of training set, test set and verification set are all greater than 0.92, and the predicted curve is similar to the actual curve, indicating that the model has a good fitting effect.

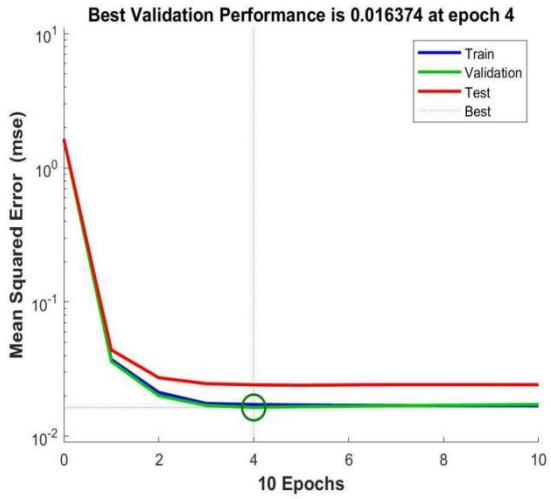

Figure 4 Regression goodness of model

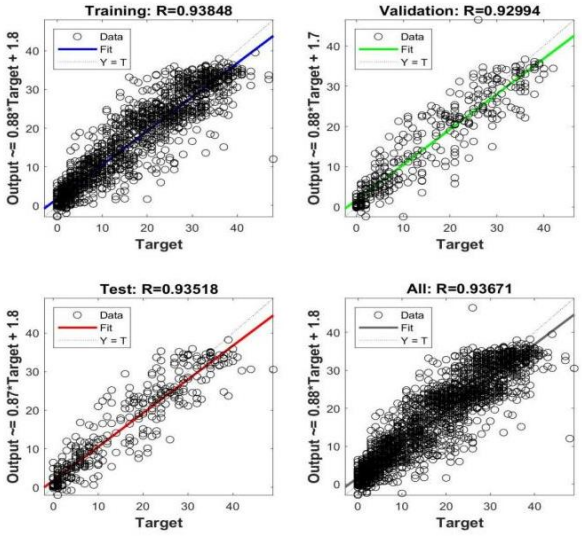

Figure 5 Training performance of model

4.3. Results

The parameter condition of the word EERIE: word repetition, word harmony, capital letter association can be obtained directly, and the frequency of EERIE words obtained by looking up in the dictionary is 1 per million.

For the word EERIE on March 1st, 2023, the information is shown in Table 10. Above, we obtained the Number of reported results predicted by SVM as 20476.91156 on March 1st, and x=620 was replaced into the regression equation through curve fitting function. The proportion of predicted difficult mode is: 0.0946.

Table 10 Overview of EERIE data on 1 March, 23

	Contest number	Number of reported results	Number in hard mode	Word repetition	Word harmony	Capital letter association	Word frequency
Word frequency	620	20476.91	1931.83	1	1	12.5	1

After the relevant parameters are substituted into the data, the distribution result of the number of attempts predicted by the model is shown in Table 11.

Table 11 Predicted EERIE attempt number distribution on March 1, 2023

Word	1	2	3	4	5	6	X
EERIE	0.0575	4.8679	21.3727	33.5346	25.4730	11.6570	1.94714

4.4. Model evaluation

According to the model training effect drawing, the BP neural network model has good fitting effect. For the analysis of the results, the sum of EERIE's attempt times distribution on March 1, 2023 is 98.9084, which is within the actual control range. As the data volume is still small and relies on the SVM model, it is not possible to estimate the data distribution far beyond December 31, 2022.

5. Conclusion

This paper uses relevant data to predict the number of results reported on March 1, 2023. The prediction model shows that it is mainly related to contest number, which indicates that the number of reported results is closely related to game popularity. Based on this prediction, the distribution of people's attempts is predicted if the word of the day is "EERIE". This shows that the distribution of the number of attempts on this day can be predicted by adjusting the words through the above model. Through this regulation, the popularity of the game can be maintained to some extent, so that the number of people playing the game can maintain stability or growth.

However, the attributes of the words did not affect people's choice of difficult patterns. Because the hard mode is selected in the Settings page of the game, the player cannot predict the word of the day.

The model of predicted the level of word "EERIE" can be used for analyzing the difficulty of article, which helps to judge the most proper grade of child to read. It can be used more than words, such as we can collect the information of users reading habit and offer more satisfactory contents. The same difficulty article can replace the basic topic, the same quantity of conjunction can avoiding fatigue. Even more we can use this model for analyzing pictures, games or music with some new features .

References

[1] Hiebert E H, Scott J A, Castaneda R, et al. An analysis of the features of words that influence vocabulary difficulty [J]. Education Sciences, 2019, 9(1): 8.

[2] Yılmaz K, Temizkan V. The effects of educational service quality and socio-cultural adaptation difficulties on international students' higher education satisfaction [J]. SAGE Open, 2022, 12(1): 21582440221078316.

[3] Sadaf Ghaffari, Nikhil Krishnaswamy (2023) Grounding and Distinguishing Conceptual Vocabulary Through Similarity Learning in Embodied Simulations

[4] Liu Z. (2021). Red tide disaster prediction algorithm based on neural network and SVM.

[5] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018 Jan-Feb; 15(1): 41-51.

[6] Sahu R T, Verma M K, Ahmad I. Interpreting Different Timeslot Precipitation Characteristics in the Seonath River Basin, Chhattisgarh During 1901–2017[M]//Recent Advances in Sustainable Environment: Select Proceedings of RAiSE 2022. Singapore: Springer Nature Singapore, 2022: 21-37.

[7] Chen Z, Huang L, Yang W, et al. More than word frequencies: Authorship attribution via natural frequency zoned word distribution analysis [J]. arXiv preprint arXiv:1208.3001, 2012.

[8] Salmerón Gómez, Román, Ainara Rodríguez Sánchez, Catalina García García, and José García Pérez. 2020. "The VIF and MSE in Raise Regression" Mathematics 8, no. 4: 605.

[9] Long Qian, Jianbin Zhao, Yue Ma, Option Pricing Based on GA-BP neural network, Procedia Computer Science, Volume 199, 2022, Pages 1340-1354, ISSN 1877-0509.

[10] Wang J, Wang X, Wen H. The Use of BP Neural Network Algorithm and Natural Language Processing in the Impact of Social Audit on Enterprise Innovation Ability. Comput Intell Neurosci. 2022 May 18; 2022:7297769.