Analysis Of Wordle Games Based on Multiple Logistic Regression and Clustering

Guangsheng Liu [*, #], Binjian Cai [#], Junlin Liao [#]

South China Normal University Department of School of Mathematics,510631, Guangzhou, China

* Corresponding Author Email: lgs2461574798@163.com

Abstract. The development of social media has led to an increasing number of people participating in the online guessing game Wordle, which has become a popular phenomenon on Twitter. Understanding the difficulty of the Wordle game is important for both players and game designers. This article aims to gain an in-depth understanding of the game by analyzing its gameplay reports and word difficulty. Firstly, this article predicts the distribution of Wordle report results. Using a multivariate logistic regression model, the probability distribution prediction is converted into a classification problem, with word attributes and the average number of guesses per player as independent variables for prediction. At the same time, this article introduces random noise into the model prediction to address low probability issues. The model is evaluated using MAE and RMSE, and sensitivity analysis is conducted using k-fold cross-validation. In order to further evaluate the difficulty of words in the game, this article uses the entropy weight-TOPSIS method combined with the Bi-Kmeans classification model to classify word difficulty. The model results show that word difficulty can be divided into five categories. Finally, this article evaluates the performance of the model using Davies-Bouldin Index and the Silhouette Coefficient, and the results show that the model has strong reliability and persuasiveness.

Keywords: Wordle, multivariate logistic regression model, Kmeans, word attributes, topsis.

1. Introduction

Similar to the historically popular crossword puzzle, wordle is an online word guessing game that requires players to enter letters into five squares in sequence in up to six attempts to guess the word given by the system. Thanks to the game's simplicity, convenience, and ease of sharing results, wordle is spreading worldwide, with hundreds of players posting their wordle results on Twitter every day.

In the past studies, scholars mainly focused on the research of wordle game strategy [1] and the improvement of wordle game interface [2], but there are few articles on data mining and analysis of wordle game results. Although Twitter does not cover all the game results data, data analysis of existing wordle data on Twitter can further understand the development status of wordle games and the future development trend. Therefore, this paper uses multiple logistic regression model, K-means clustering algorithm and entropy-weighted-topsis evaluation model to mine and analyze the found wordle game results data.

2. Empirical analysis

2.1. Data crawling and pre-processing

In this paper, we use python software to text-mine the wordle comments shared by Twitter between January 7, 2022 and December 31, 2022 to derive a dataset about the wordle results, including the date, the game number, the word of the day, the number of people who reported their scores on that day, the number of players who were in the difficult mode, and the number of players who made one, two, three, four, five, six attempts, the percentage of words guessed or puzzles that could not be solved (denoted by X). Considering the possibility of a small amount of anomalous data in the dataset, we performed data preprocessing on the dataset. It was detected that there were two types of outliers in the dataset: the number of letters in the given word was not 5, and the sum of the number of times required to solve each word was not 100%.

Therefore, the samples in which outliers were present were excluded.

Based on the words provided in the dataset each day, the values of their word attributes were counted. The word attributes selected are shown in Table 1:

Table 1: Word attribute selection

Word attribute name	Explanation
l_f	word occurrence frequency in Corpus of Contemporary American English from 2005 to 2019
l_{max}	maximum number of repeated letters in a word
l_l	the number of repeated letters
l_n	the number of orthographic neighbors to the word

2.2. Model Theory

2.2.1 Multiple Logistic Regression Model

Logistics regression analysis, a generalized linear regression analysis model, is commonly used in data mining, automatic disease diagnosis and other fields. Since the logistic function takes values from 0 to 1, it is often used to build risk level or probability prediction models. However, it is mainly used to solve the problem of dichotomous variables and is not applicable to the multicategorical variables. Therefore, it is necessary to make logistic regression model extended to the multiclassification case, i.e., the multiple logistic regression model. The goal of the multiple logistic regression model is to predict the probability distribution of each category. Specifically, the model outputs probabilities for each category that sum to 1. The output of the model shows the probability that a new input sample belongs to each category [3-4].

The multiple logistic regression model first computes a weighted sum of the characteristics of the input samples and the model parameters. This weighted sum is called a linear predictor. Then the linear predictors are fed into a **nonlinear** function called **Softmax**[5], which converts each linear predictor into a nonnegative probability that sums to 1. Specifically, given an input variable$\mathrm{X} = \{\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{p}\}$, The linear predictor of the model can be expressed as:

$$z_1 = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p \quad (1)$$

$$z_2 = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p \quad (2)$$

$$z_{K-1} = \beta_{(K-1)0} + \beta_{(K-1)1}x_1 + \cdots + \beta_{(K-1)p}x_p \quad (3)$$

K is the number of categories, $\mathrm{z_k}$denotes the linear predictor belonging to the $\mathrm{k^{th}}$ class, $\beta_{\mathrm{k0}}$is the intercept, $\beta_{\mathrm{kp}}$is the coefficient of the $\mathrm{p^{th}}$ feature.

Then, input the linear predictor into the softmax function:

$$\hat{p}_k = \frac{\exp(z_k)}{\sum_{j=1}^{K-1}\exp(z_j)} \quad (4)$$

$\hat{p}_k$denotes the probability that the input vector belongs to the k^{th} class, The probability of the last category can be obtained by normalizing.

In practical applications, the model is usually trained using a cross-entropy loss function. By minimizing the loss function on the training set, the optimal model parameters are gotten. After the training is completed, the paper can use the model to make predictions on new input samples.

2.2.2 Bi-kmeans clustering model

Bi-Kmeans is a clustering algorithm that extends the traditional K-means algorithm to better handle datasets with unevenly sized clusters [6-8]. It works by first applying K-means clustering to the input dataset, and then applying K-means clustering again on the resulting clusters to further refine

the clustering. The algorithm repeats this process until the clusters are sufficiently small or the desired number of clusters is reached.

The key innovation of Bi-Kmeans is that it applies K-means clustering to the clusters themselves, rather than to the entire dataset. This allows the algorithm to handle datasets where the clusters are of very different sizes, by recursively splitting larger clusters into smaller ones until a desired granularity is achieved.

Overall, Bi-Kmeans is a useful clustering algorithm for datasets with unevenly sized clusters and can be applied to a wide range of applications such as image segmentation, text classification, and market segmentation.

2.2.3 The entropy method-TOPSIS evaluation model

TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution),[9] an evaluation method to make relative merit comparisons from all existing evaluation objects, is widely used in the construction of comprehensive evaluation systems, but the TOPSIS method does not respond to the variables However, the topsis method does not reflect the degree of correlation and importance between variables. Therefore, it is essential to apply the weight evaluation method to attach the corresponding weight coefficients to each index sample before using TOPSIS to compare the advantages and disadvantages [10-11].

Compared with the correlation coefficient method and the variance coefficient method, the entropy method has better robustness and is more reflective of the importance of indicators by means of information entropy. Therefore, this paper mainly uses the entropy method to assign weight to the indicators, which constitutes the entropy method-topsis combination model [12].

2.3. Evaluation of the model

To verify the rationality of the model prediction, the paper uses the MAE and RMSE point prediction evaluation indicators to evaluate the prediction accuracy of the Multiple Logistic Regression Model. The evaluation indicators are shown in Table 2.

Table 2: Description of evaluation indicators

	Point estimation	
Symbols	MAE	RMSE
Meaning	Average absolute error	Root mean square error
Calculation formula	MAE=$\frac{1}{N}\sum\|\bar{y}_i - y_i\|$	RMSE=$\sqrt{\frac{1}{N}\sum(\bar{y}_i - y_i)^2}$

Where, $\bar{y}_i$ is the predicted value, y_i is the true value.

To verify the accuracy of the Bi-kmeans clustering model, the paper introduces two evaluation metrics, Davies-Bouldin Index [13] and Silhouette Coefficient [14], with the following formulas:

$$SC = \frac{b(x_i)-a(x_i)}{\max\{b(x_i),a(x_i)\}} \tag{5}$$

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\frac{avg(c_i)+avg(c_j)}{den(c_i,c_j)} \tag{6}$$

Where, $b(x_i)$is the minimum distance from other classes of samples to sample x_i, $a(x_i)$is the average distance of the sample from other samples in the same category,$avg(c_i)$is the average distance of the sample of class i,$den(c_i, c_j)$ is the distance between the center of class i and class j.

3. Model Establishment and Resolution

3.1. Prediction of the distribution of reported results

3.1.1 Results of the prediction

The multiple logistic regression model can naturally handle the case where the dependent variable sums to 1, so the probability prediction for solving this question performs well. We take the four-word attributes selected in the previous section as independent variables and input the distribution of each word reported outcome as the dependent variable into the model. The model is trained to learn the data that have been preprocessed to fit the optimal parameters. Then, we input the attributes of the specified test words into the trained model to obtain the distribution of reported results for the specified words. Specifically, taking EERIE as an example, the predicted distribution of reported results for this word is shown in the following Table 3.

Table 3: Prediction results of EERIE under multiple logistic regression

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 tries
0	0	0.28023231	0.28036615	0.18686272	0.24795541	0.00458341

From Table 3, it can be seen that the results obtained after the prediction of multiple logistic regression have only five categories of probability distributions. We can find that multiple logistic regression needs to transform the probability distribution labels of training data into classification labels in the process of training and learning. This process requires Softmax smoothing of the probability distribution and its classification by argmax function. Since the proportion of '1 try' and '2 tries' categories is always low, even close to 0 at one time, machine learning considers them as almost impossible events in this transformation process, and thus the final output does not contain the probability of these two categories. Figure 1 shows its schematic representation from probability distribution to classification.

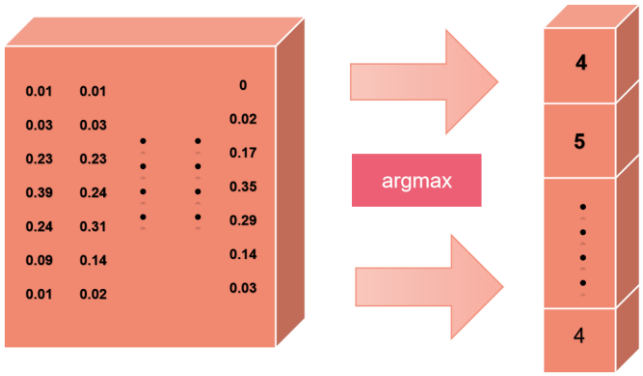

Figure 1: Probability distribution to category diagram

However, we know that '1 try' and '2 tries' to complete the game still exist in Wordle games. This indicates that these two categorical categories are small probability events in real life, rather than unlikely to happen. Therefore, we need to make further corrections to the results obtained from the multiple logistic regression model to make its results more relevant to the real-life situation.

In order to make the predicted probability distribution results more realistic, we will additive noise to the prediction results of multiple logistic regression. Additive noise is a type of random noise, which adds a varying constant value to the original signal. We use the additive noise method in this subsection to add a Gaussian distributed random number obeying a mean of 0.0005 and a standard deviation of 0.001 to each element of the probability vector. The additive noise perturbation has a random nature. To reduce the effect of this randomness, the original probability vector is perturbed 10,000 times, and the final result is averaged over these 10,000 perturbations. The final result is shown in Figure 2.

before	disturbance
1 try	0
2 tries	0
3 tries	0.2802323
4 tries	0.2803662
5 tries	0.1868627
6 tries	0.2479554
7 tries	0.0045834

Perturbed entry		
Column 1	…	Column 10000
x_{11}	…	$x_{10000,1}$
x_{12}	…	$x_{10000,2}$
x_{13}	…	$x_{10000,3}$
x_{14}	…	$x_{10000,4}$
x_{15}	…	$x_{10000,5}$
x_{16}	…	$x_{10000,6}$
x_{17}	…	$x_{10000,7}$

after	disturbance
1 try	0.0005
2 tries	0.0005
3 tries	0.2798
4 tries	0.2799
5 tries	0.1867
6 tries	0.2476
7 tries	0.0050

Figure 2: Diagram of the final result of the probability distribution

3.1.2 Accuracy and Sensitivity Analysis of the Model

Although this paper has obtained the results of the distribution of future solution words by the model, it cannot verify whether the model results are reliable by it because the real distribution of future solution words is unknown. To evaluate this, the paper needs to divide the known data into a training set and a test set. The model learns the training set by training, and then makes predictions on the test set. Eventually the paper then compares the predicted results together with the true distribution to verify the reliability of the model. It is important to note that during the model training process, we still have to strictly follow the prediction steps, adding perturbation impact to the model prediction results.The paper divided the known data into a training set as well as a test set in the ratio of 7:3, and evaluated the accuracy of the prediction model using three evaluation metrics, MAE and RMSE, in order to demonstrate its effectiveness. The values of MAE and RMSE are 0.0696 and 0.0936 respectively.With these data, the paper can get that the results of the prediction model are plausible.

Because the data in reality are often subject to various anomalies and uncertainties, it is difficult to cope with these situations if the model lacks robustness. The prediction results of such a model will be unreliable or cannot be applied in real-world scenarios. Therefore, in order to verify whether the prediction model is really applicable in real situations, it is needed to perform sensitivity analysis on it.

Common methods of sensitivity analysis are adjusting the coefficients of each variable, changing the values of the input variables, or changing the number of input variables. Most of the above methods can be implemented by cross-validation. In this subsection, the paper evaluates the sensitivity of the model for different datasets by cross-validation. This paper observes the performance of the model on different datasets by dividing the dataset into different training and testing sets. Here the paper will use the same two evaluation metrics, MAE and RMSE, to represent the evaluation results.

This paper collecst the evaluation metric values under different parameters by continuously changing the grouping parameters of the K-fold cross-validation. Finally, by visualizing the collected assessment index values, the paper observes its range of variation. The results are shown in Figure 3. Here, the parameter are changed from 5 to 15.

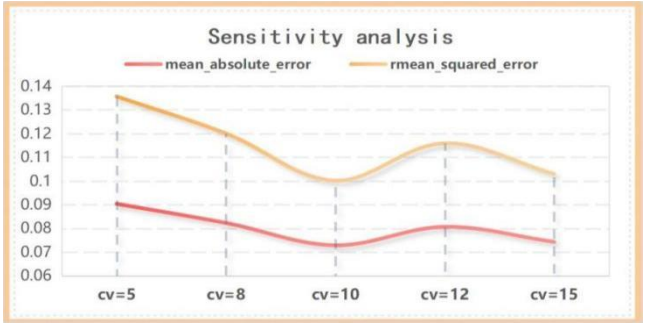

Figure 3: Model sensitivity analysis

From Figure 3, it can be seen that the range of variation of MAE assessment index under different parameters is within about 0.04. And the range of variation of RMSE under different parameters is about within 0.02. All these ranges of variation are small. This shows that the prediction model has good robustness and low sensitivity. It is applicable to practical situations.

3.2. Difficulty classification of Wordle words

3.2.1 Establishment of the evaluation model

The guessing difficulty of a word is determined by its own attributes. In order to achieve a better word difficulty division effect, this article needs to select representative indicators from the word attributes. Based on the principles of systematicity and scientificity of the evaluation system, and in reference to literature, this article selects the following indicators as our evaluation indicators. After selecting the evaluation indicators, this article constructs a comprehensive evaluation model using the entropy weight-TOPSIS evaluation model. By inputting the attribute indicators corresponding to different words into the constructed comprehensive evaluation model, the comprehensive scores of the guessing difficulty of each sample word can be obtained. Figure 4 describes in detail the selection of each evaluation index.

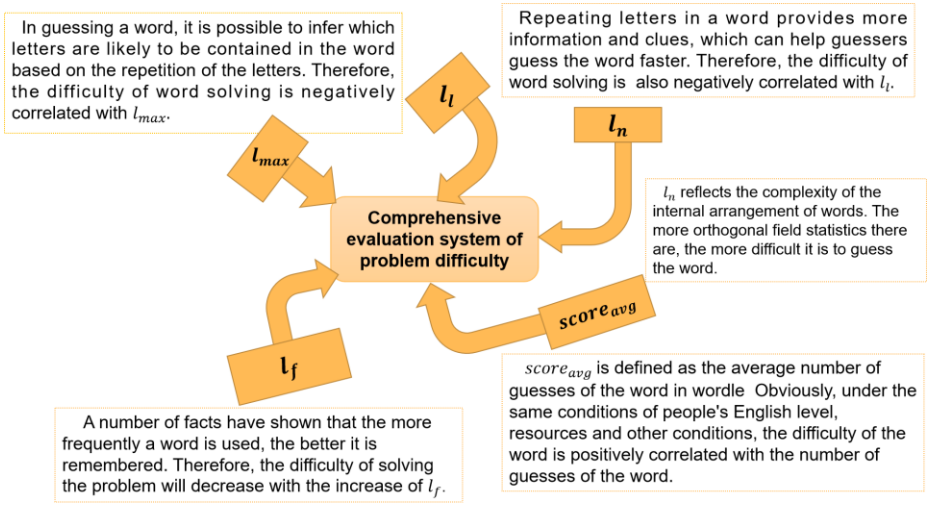

Figure 4: Selection of evaluation indicators in the evaluation system

3.2.2 Establishment of the Classification Model and Analysis of Results

After obtaining the comprehensive scores of the difficulty of solving each sample word using the comprehensive evaluation system constructed in Section 3.2.1, this article will use the bi-means clustering algorithm to cluster each sample based on the difficulty of solving the problem. Bi-kmeans is an improved model designed to address the flaw that the k-means model may fall into local optima. Based on the principle of minimizing SSE, this model first regards all data points as a cluster, then divides the cluster into two, and then selects one of the clusters to continue the division. The selection of the cluster to be divided depends on whether its division can minimize the value of SSE to the greatest extent. Based on this, this article selects the clustering number k=5, and the clustering result of the bi-means clustering model is shown in Figure 5.

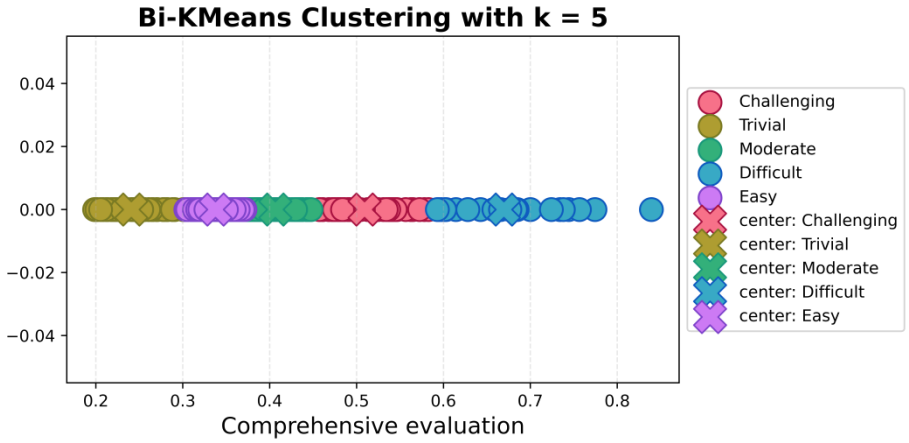

Figure 5: Clustering results of bi-means clustering model with k=5

According to the magnitude of the overall rating of the difficulty of the solution in the comprehensive evaluation system, the categories are named from left to right: **Tricvial, Easy, Moderate, Challenging, Difficult**. In the end, we will substitute the term "ERRIE" on March 1st into the comprehensive evaluation model to obtain its difficulty score. According to the difficulty classification criteria obtained by bi-kmeans clustering, the difficulty level of the word "ERRIE" is "Trivial".

It is worth noting that since we did not obtain the real percentage distribution data for March 1st in advance, we used the percentage distribution data predicted in section 5.1 to calculate the "ERRIE" word.

3.2.3 Accuracy and Sensitivity Analysis of the Model

This article evaluates the established clustering model using two indicators, the Davies-Bouldin Index (DB) and the Silhouette Coefficient (SC). First, to verify the accuracy of the bi-kmeans model in Section 3.2.2, this article calculated its DB and SC values, which were 0.41 and 0.60, respectively. The SC of the model is greater than 0, and the DB is less than 1, indicating that it has a good and accurate classification effect.

This article also conducted sensitivity analysis on the classification model by changing the clustering number k value of the model evaluation indicator, and the specific results are shown in Figure 6. As shown in Figure 6, regardless of the SC value or the DB value, the magnitude of the indicator changes is within 0.05 after changing the k value, and the changes are not significant, indicating that the classification model has good stability.

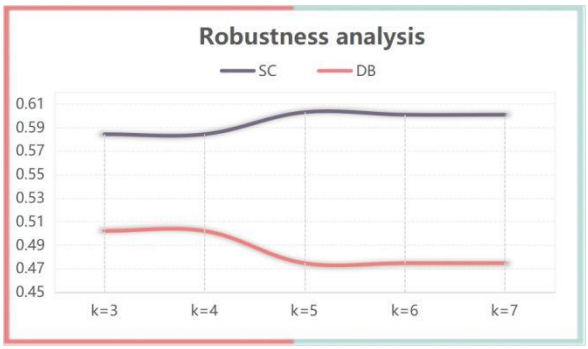

Figure 6: Sensitivity analysis of classification model

4. Conclusion

Based on the data about wordle game results from Twitter comment text mining, multiple logistic regression model, K-means clustering algorithm and entropy method-topsis evaluation model are used to mine and analyze the data, and the following conclusions are drawn:

(1) The multiple logistic regression model is reasonable, reliable and stable in the prediction of the distribution of the number of word guesses.

(2) In the estimation of word difficulty, the entropy-topsis model is able to measure the word difficulty coefficient based on the existing word attribute index values, while the bi-kmeans clustering model can classify the word difficulty more accurately based on the word difficulty coefficient values, and then estimate the word difficulty based on the word difficulty coefficient coefficient values of each word.

References

[1] Liu C L. Using wordle for learning to design and compare strategies[C]//2022 IEEE Conference on Games (CoG). IEEE, 2022: 465-472.

[2] Pamungkas N A R. The Effects of Wordle Media on Students' Vocabulary Mastery [J]. JETAL: Journal of English Teaching & Applied Linguistic, 2021, 2(2): 56-61.

[3] Liu X, Teng SH, Long FAN et al. Research on early warning model of road icing based on multiple logistic regression[J]. Hunan Transportation Science and Technology,2022,48(04):101-107.

[4] Ma, J., & Sun, Y. (2015). Multinomial logistic regression-based health risk assessment using biomonitoring data from the Canadian Health Measures Survey. Environmental Research, 138, 185-192.

[5] Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[6] Li, C., Li, Y., Li, D., Li, Y., & Wang, J. (2018). Bi-KMeans algorithm based on fuzzy clustering for intrusion detection. International Journal of Distributed Sensor Networks, 14(1), 1-12.

[7] Ma T, Yu J, Zhang J, et al. A collaborative filtering recommendation algorithm based on hybrid clustering and fused user interests [J]. Electronic Technology Applications,2022,48(04):29-33.DOI:10.16157/j.issn.0258-7998.212086.

[8] Ma T. Research on the recommendation method of students' practical training direction based on knowledge graph[D]. North Central University, 2022. DOI:10.27470/d.cnki.ghbgc.2022.000483.

[9] Cheng Yuteng,Yan Su,Zhao Qin. Research on the quality evaluation of civil aviation recruitment students based on entropy weight TOPSIS method[J]. Journal of Civil Aviation,2023,7(01):132-136.

[10] Luo Zhen. Research on dynamic and comprehensive evaluation method of listed companies' performance [D]. Hunan University, 2009.

[11] Wang Shunjiu, Zhou Jiye.A review of the application of comprehensive meteorological multi-factor evaluation methods in China[J]. Alpine Mountain Meteorology Research,2019,39(04):88-96.

[12] Wu Jizhao. Research on the competitiveness of inland river ports based on CRITIC-entropy power method and TOPSIS method [D]. Chongqing Jiaotong University, 2022

[13] Yang Houyi. Workpiece Positioning and Grasping Based on Vision [D]. Southwest University of Science and Technology,2018.

[14] Zhang Denrong, Du Yao, Xun Dandan, Liu Ting. Kernel-kmeans: A Spatial Clustering Algorithm Based on Kernel Density Estimation [J]. Journal of Hangzhou Normal University (Natural Science Edition),2017,16(03):324-329.