Analyzing and Predicting Player Behavior by SARIMAX and Neural Networks

Dengke Ruan *

School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, 611731

* Corresponding Author Email: ruan_dengke@outlook.com

Abstract. This article focuses on the trend analysis of Wordle puzzles and predicting player behavior through advanced models. This paper found based on the number of wordle users on social platforms in 2022 that this number grows in the early burst and then the region declines smoothly. This paper introduce an exogenous variable-word difficulty-to optimize the model, The model that has been changed is called SARIMAX. The predicted player count for March 1, 2023 is between 15,647 and 29,059. Our forecast distribution for the word 'EERIE' on March 1, 2023 is 2 to 4 occurrences, with a likelihood of 89%. This article choose to fuse the word difficulty evaluation, calculated from dataset, with the classical K-means clustering algorithm to classify the words into 5 classes according to their difficulty. The model introduces a neural network classification algorithm that eventually classify words based on their own properties with high accuracy. Substituting EERIE into this model, the word was found to be of medium difficulty.

Keywords: SARIMAX, BPNN, K-means, PSO.

1. Introduction

In recent years, advancements in data science and machine learning have transformed various industries, including gaming. By analyzing the vast amount of data, they can dig for obscure pieces of information that are useful to the. As such, in the game industry, by analyzing players 'game logs, companies can design models that predict players' behavior to help developers optimize game design and improve user experience. [1] But these prediction models tend to focus on the user's behavior and ignore the impact of the nature of the game itself. Wordle, one of popular crossword games, has created a new wave of crossword puzzles on social platforms since its launch. By using a neural network, different level of difficulty are taken into consideration, which help designers to understand the dynamics of user engagement in casual games through analyzing the impact of game difficulty on player behavior based on the existing data and find a better game design scheme.

2. Seasonal AutoRegressive Integrated Moving Average with exogenous regressors

2.1. SARAMAX-based model for predicting the number of people reporting

The paper introduce an exogenous variable for the SARIMA model - the difficulty of the word [2]. The new model with the addition of this variable is called the SARIMAX model. By including the exogenous variable in the SARIMAX model, the time series can be predicted more accurately and the factors that affect the time series can be better understood. The SARIMAX (p, d, and q) (P, D, and Q) m model can be written as formula (1):

$$y_t = c + \sum_{i=1}^{p} \varphi_i \, y_{t-i} + \sum_{i=1}^{q} \theta_i \, \varepsilon_{t-i} + \sum_{i=1}^{P} \phi_i \, y_{t-im} + \sum_{i=1}^{Q} \theta_i \, \varepsilon_{t-im} + \sum_{i=1}^{k} \beta_i \, x_{i,t} + \varepsilon_t \quad (1)$$

Where $x_{i,t}$ denotes the value of the ith external variable at time point t and β_i denotes the coefficient of the external variable$x_{i,t}$.

Word difficulty was rated for each word in the given dataset using the method described above, and the difficulty of each word was used as an exogenous variable affecting the number of daily

reports, and word difficulty for the prediction interval was randomly generated based on the average of word difficulty over the past year. Predictions were made based on this model and the given data set, and the resulting data are shown in Figure 1.

The final projection is 22,353 reported on 2023/3/1.this paper can see that the SARIMAX model's predictions for future data remain largely stable, and the overall decline in the number of reports over the past year has moderated and gradually stabilized.

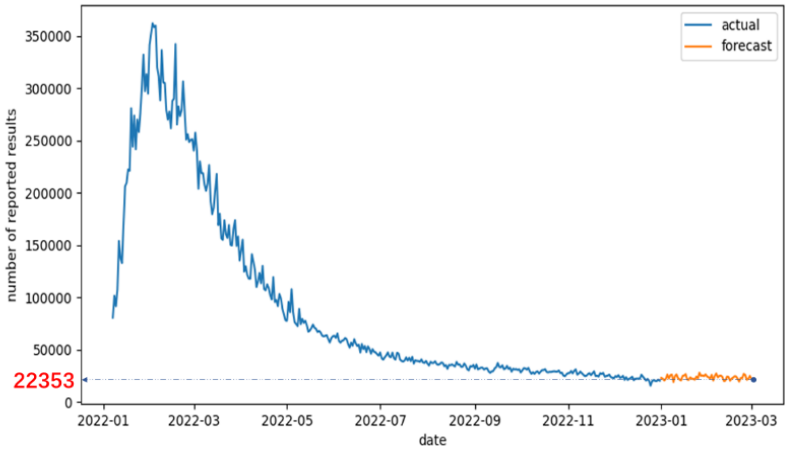

Figure 1. SARIMAX model.

2.2. Construction of prediction intervals

By constructing the SARAMAX model, this paper predicted the data situation after two months by given, but because the prediction has uncertainty, this paper cannot confirm the value as the final result, but this paper can give a prediction interval to ensure that the probability that the real data is within the interval is greater than 95%.

Specifically, this paper use the data from 2022-1-1 to 2022-11-1 as the training set, and the data from 2022-11-1 to 2022-12-31 for a total of 2 months as the validation set, and finally find that the prediction for the validation set can reach 95% correct rate in the case of 30% error tolerance interval, so this paper can say that our prediction interval is within the final prediction result of + within 30%, 15647 to 29059, can achieve 95% correctness. The obtained results are shown in Figure 2.

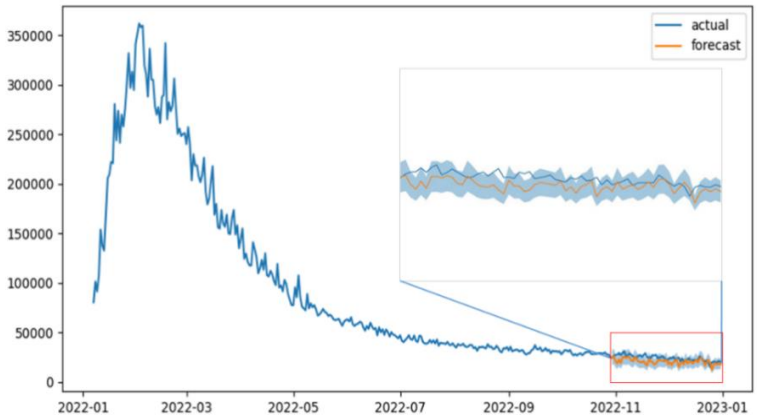

Figure 2. SARIMAX model.

2.3. The relationship between word attributes and the people's behavior.

Each word has its own attributes, and to determine whether there is a word attribute related to the percentage of people choosing the difficult mode that day, the first thing that should be done is to extract the possible word attributes. However, there are many attributes of words, and it is obvious that it is not possible to take all of them into account, so this paper can only consider some of the word attributes mainly.[3], [4].

Through the calculation of information entropy [5] and prediction expectation, it is clear that the number of people playing HARD mode per day as a percentage of the total number of reports and each of the above seven word attributes were subjected to correlation analysis, and the correlation test was performed using Pearson correlation coefficients, which are statistics measuring the linear correlation between two variables with values between -1 and 1 (1 represents a perfectly positive linear correlation between the two, 0 is no linear correlation, and -1 is a perfectly negative linear correlation), and the results obtained are shown in Table 1.

Table 1. Correlation test table.

variate	spearman correlation coefficient /P
Words Frequency	-0.074/0.374
Expected tries	0.032/0.550
Character frequency	-0.014/0.823
Fixed character Frequency	-0.024/0.656
Vothis paperl frequency	0.08/0.134
Repeat characters	-0.087/0.102
Words density	0.0129/0.715

According to the data analysis, this paper can find that the correlation coefficients are all around 0 and the statistically significant relationship p-values are large, (statistically $P<0.05$ when our variables are considered relevant), so this paper can conclude that the percentage of the number of people choosing difficulty each day is not related to the word attribute of the day. This combine the game mechanics with the fact that users proportion depends entirely on the user group preferences.

3. Predicting the n-try Distribution of Words

3.1. Regression Model Based on Word Attributes

The score distribution of words is directly related to the difficulty of the words. Therefore, this paper need to extract some features related to the difficulty of words. This paper extracted 32 features that may affect the difficulty of words, including the first six features and 26 features related to whether the word contains specific letters. For the first six features, this paper selected two words with significant differences in difficulty and compared their feature data. This paper also studied the relationship between all features and the difficulty of words with a sanky diagram. Based on all features, this paper performed separate regressions for each try and finally concatenated the complete distribution of words for that day. A part of results are shown in Figure 3 and Figure 4.

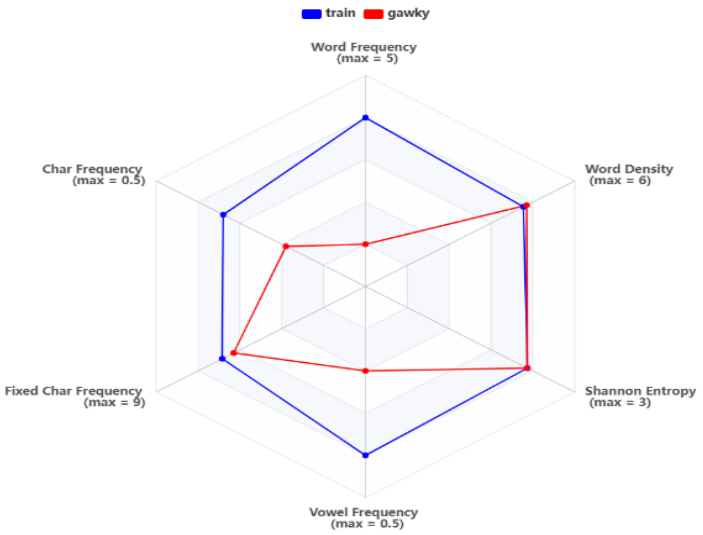

Figure 3. Features of word ‘train’ and ‘gawky’.

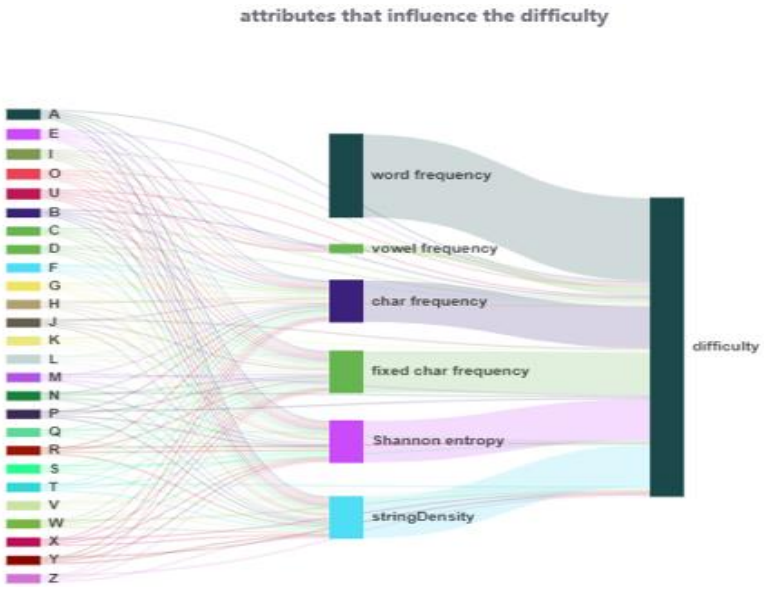

Figure 4. Relationships between features and difficulty.

3.2. BP neural network [6] with particle swarm optimization algorithm

Particle swarm optimization's (PSO) core idea [7] is to use the sharing of information among individuals in a group to produce an evolutionary process from disorder to order in the problem-solving space, in order to obtain feasible solutions to the problem. This paper kept the initial data consistent and added the particle swarm optimization algorithm. The neural network's prediction accuracy was used as the objective function, and different network parameters were used as the population, iteratively searching for the optimal parameters. The results are as Table 2:

Table 2. BPNN prediction results on 3 tries using PSO.

	MES	RMSE	MAE	MRPE	R^2
Training set	22.089	4.7	3.781	17.841	0.615
Test set	32.015	5.658	4.625	22.37	0.512

With reference to the optimal network parameters, this paper conducted training and prediction and found that, taking 3-tries as an example, this paper can obtain more reliable results in Figure 5.

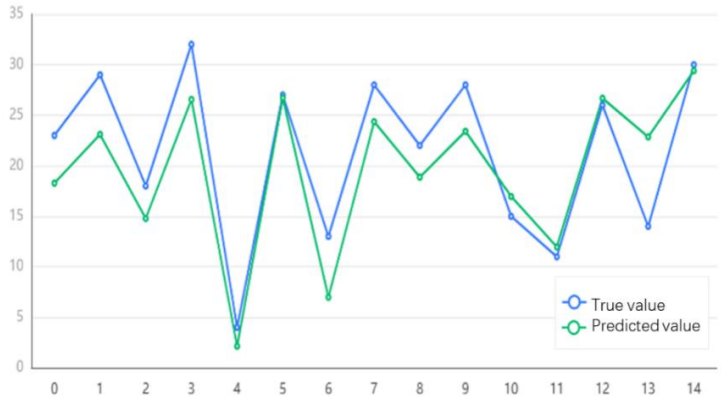

Figure 5. BPNN prediction results on 3 tries using PSO.

This paper divided the feature values of 3try in the given dataset, where 300 (80%) were used as the testing set and 60 as the validation set. Taking 3try as an example, the R-square value was 0.615 in the training set and 0.512 in the testing set, showing acceptable performance in both sets. Upon observing the graph, this paper found that the neural network can almost predict the actual values accurately. Considering that machine learning requires a large dataset for learning, this paper only have 300 data for learning, hence the obtained results are acceptable.

Therefore, using the optimized BP neural network, this paper predicted the distribution for March 1, 2023, as shown in the table 3.

Table 3. Score distribution on 2023/3/1.

Try	1	2	3	4	5	6	X
ERRIE	0	5.9975	27.9427	38.9339	22.3349	3.0018	1.5059

This paper used the predictive model to make separate predictions for 1try-xtries, and found that the score distribution on March 1st summed up to 96.7. Although the seven variables were predicted separately, the sum is close to 100 (the training set data was also processed to ensure that the sum of the seven variables is 100), so this paper believe that this model has good credibility. This paper controlled the variables and used the model to predict the distribution of the same word on January 1st. The results are shown in the graph, and the sum is 98.1. Therefore, this paper can infer that the accuracy of the prediction decreases gradually with time, but it still has high accuracy.

4. Wordle solution difficulty classification model

4.1. a clustering model based on the K-means algorithm.

K-means [8] is a commonly used clustering algorithm based on Euclidean distance, which considers that the closer the distance between two targets, the higher the similarity. For a given dataset, K-means algorithm can be used to cluster words into K classes by setting K centroids.

Firstly, K initial cluster centers $C_i(1 \le i \le k)$ are randomly selected from the dataset. Then, the Euclidean distance between each data object and each cluster center C_i is calculated. Each data object is assigned to the cluster center$C_i$corresponding to the closest distance. After that, the average value of data objects in each cluster is computed as the new cluster center, and the process is repeated until the cluster center no longer changes or the maximum number of iterations is reached [9].

The Euclidean distance [10] between data objects and cluster centers in space is as formula (2):

$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})^2} \tag{2}$$

Where x is a data object, C_i is the i-th cluster center, m is the dimensionality of the data object, and X_j, C_{ij} are the j-th attribute values of x and C_i, respectively. The formula (3) is concerning calculating the sum of squared errors (SSE) for the entire dataset:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |d(x, C_i)|^2 \tag{3}$$

Here, SSE represents the goodness of the clustering results, and k denotes the number of clusters. This paper set k=5 based on the frequency classification in the dictionary, and apply the K-means algorithm to cluster the data. Different difficulty levels of words will be automatically classified into different clusters, with each cluster representing a specific level of difficulty. The resulting clusters are shown in Figure 6.

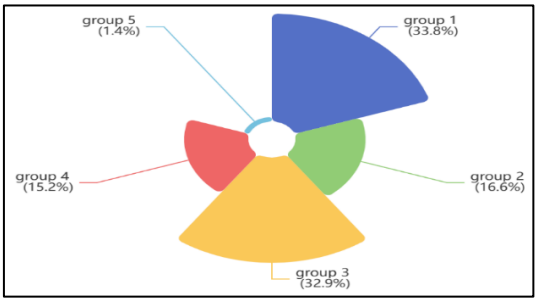

Figure 6. Cluster analysis results.

The table 4 shows the evaluation metrics of the clustering results obtained by the K-means algorithm. For a sample set, the silhouette coefficient is the average of all sample silhouette coefficients. The value range of the silhouette coefficient is [-1, 1], with a higher score indicating a better clustering result where samples within the same cluster are closer to each other and farther away from samples in different clusters [11].

Table 4. Cluster analysis results.

Clustering category	Frequency	Percentage (%)
Very easy	54	15.211
Difficult	120	33.803
Moderate	117	32.958
Very easy	59	16.62
Very difficult	5	1.408

4.2. Word Classification Model Based on Neural Network

This paper clustered the words in the dataset into 5 categories using the K-means algorithm. The classified words were then used as the training set for the classification neural network. The final trained model's performance is shown in the figure 7, and the figure are shown in Table 5.

Table 5. The classification results.

	Accuracy rate	Recall Rate	Precision rate	F1
Training set	0.623	0.623	0.635	0.61
Test set	0.62	0.62	0.695	0.603

The accuracy on the test set reached 70%, which is acceptable under the limited conditions of the dataset. At the same time, it can be determined that the neural network has successfully learned the features related to word classification. For instance, as shown in the figure 8, the word difficulty of "ERRIE" was classified as belonging to the third category in the neural network classification, but this paper found that the probability of it belonging to the first category is only slightly less than 1%. As this paper can see, the first category and the third category are neighboring categories. Therefore, the difficulty of this word may lie at the border between the first and the third categories, but it leans more towards the third category. Thus, according to our evaluation criteria, this word belongs to the category of moderate difficulty.

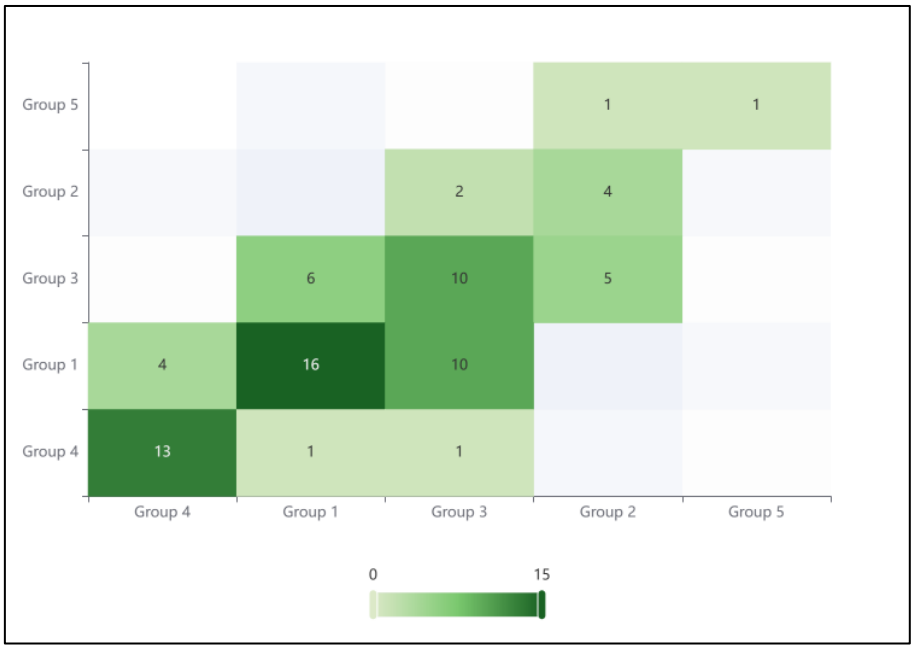

Figure 7. The confusion matrix of the classification results.

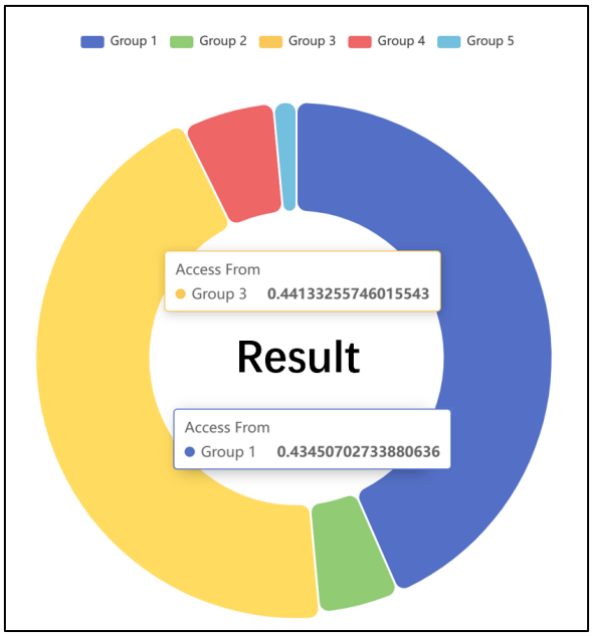

Figure 8. Possibility that 'ERRIE' belongs to each group.

5. Conclusions

This paper analyzed reported results using the SARIMAX model and found a rapid rise, peaking on February 2, 2022, followed by a rapid fall and slow decline, stabilizing over time. Changes showed a 7-day cycle, with volatile early data and flatter later data. The model predicted 15,647 to 29,059 results on March 1, 2023.BP neural network model was developed using a machine learning regression model and particle swarm algorithm. Trained with 32 manually extracted features, the model predicted EERIE for March 1, 2023. The correlation coefficient R was 0.78, significant but uncertain due to an inadequate dataset. This paper combined K-means and BP neural network classification for the given dataset, first using the clustering algorithm to classify the given word EERIE, followed by inputting the given word EERIE to get the word difficulty classification, and got the difficulty classification of EERIE as normal, but its probability is close to that of the partial simple class. In the future, for other games, more feature parameters can be extracted based on the analysis method, and more influencing factors can be added to achieve better prediction effect. The results can help software companies analyze their users' needs and preferences to develop better games and software.

References

[1] Liu Xufeng. Player data mining and behavior prediction in the virtual game scenario [D]. Xidian University, 2022.

[2] Alharbi F R, Csala D. A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach [J]. Inventions, 2022, 7(4): 94.

[3] Shannon C E .A mathematical theory of communication [J].The Bell System Technical Journal, 27[2024-05-06].

[4] Kherif F, Latypova A. Principal component analysis [M]//Machine learning. Academic Press, 2020: 209-225.

[5] Liu Z, Wang Y, Cheng Q, et al. Analysis of the information entropy on traffic flows [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18012-18023.

[6] Song S, **ong X, Wu X, et al. Modeling the SOFC by BP neural network algorithm [J]. International Journal of Hydrogen Energy, 2021, 46(38): 20065-20077.

[7] Shami T M, El-Saleh A A, Alswaitti M, et al. Particle swarm optimization: A comprehensive survey [J]. Ieee Access, 2022, 10: 10031-10061.

[8] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation [J]. Electronics, 2020, 9(8): 1295.

[9] Li S, Li S, Liu D, et al. Hardness prediction of high entropy alloys with machine learning and material descriptors selection by improved genetic algorithm [J]. Computational Materials Science, 2022, 205: 111185.

[10] Marks J, Andalman B, Beardsley P A, et al. Design galleries: A general approach to setting parameters for computer graphics and animation [M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 73-84.

[11] Řezanková H. Different approaches to the silhouette coefficient calculation in cluster evaluation[C]//21st international scientific conference AMSE applications of mathematics and statistics in economics. 2018: 1-10.