Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2420494

Decrypt Wordle: Explore the Secrets behind the Words

Summary

In 2022, a puzzle game called Wordle exploded and gradually swept the world due to its simple rules and interesting challenges. Many players reported their scores on social networking sites such as Twitter to share their gaming experience. The purpose of this report is to analyze the data provided by MCM and build mathematical models to explore the secrets behind the words.

Step 1: First, we pre-process the data and plot the relationship between The number of reported results and Date. Then, based on the plots and the corresponding data, we develop a model for analyzing and predicting the number of reported results. On the one hand, the model uses **Product Life Cycle Theory** to explain the reasons for the data changes and divides them into three phases, which are **Rapid Growth Period**, **Extreme Decline Period** and **Stabilized Decrease Period**. On the other hand, by analyzing the data provided, we used the first-order difference **ARIMA (2,1,6) model** to obtain the prediction interval of the number of reported results on March 1, 2023 as **[7128, 28106]**.

Step 2: First, we define the $Score$ of Wordle. Then, based on the data analysis, we select five word attributes that may affect the percentage of scores reported that were played in Hard Mode. Then, we quantify word attributes to investigate the relationship between each word attribute and $Score$. Ultimately, we reject the effect of Word frequency and identify **Sentiment Polarity**, **Repeated letters**, **Letter frequency**, and **Word similarity** as the four word attributes whose specific effects on $Score$ are investigated.

Step 3: To categorize the solution words based on their difficulty, first, we use **Analytic Hierarchy Process (AHP)** to assign different weights to different numbers of tries, indicating that they reflect different degrees of difficulty in solution words. Next, we calculate a composite score based on the weights and categorize all solution words into **Hard Category** and **Easy Category**. We use the four attributes of the quantified words as inputs and the different categories as outputs to predict the difficulty of solution words using the **BP Neural Networks**. The model predicts with more than **70%** accuracy and predicts with more than **85%** certainty that the EERIE is in the Hard Category.

Step 4: Using **Euclidean Distance Theory**, we establish a distance model. For each given future solution word, we compute its Euclidean distance from the rest of the words in the dataset, using the quantized four word attributes as indicators. We predict the percentage of tries for each future given word by obtaining the two words with the closest distances. For the word EERIE, we find that the two words closest to it are **MUMMY** and **FLUFF**, which results in the prediction of **(0,1,4,19,35,30,11)%** for each percentage of EERIE and consider the prediction to be plausible.

Step 5: We list and describe other interesting features in the dataset. We summarize all our findings in a letter to the Puzzle Editor of the New York Times.

Eventually, we perform a sensitivity analysis of the model to investigate the effect of changes in the parameters of the model variables on the results.

Keywords: Product Life Cycle Theory, ARIMA, Word Attributes, AHP, BP Neural Network, Euclidean Distance

Contents

1 Introduction

1.1 Problem Background

1 word, 5 letters, 6 chances. From late 2021 to early 2022, this puzzle game conquers the world. Within a few short months, English social media was swamped with daily shares from players. It is now growing in popularity and has appeared in more than 60 languages. The game, called Wordle, comes courtesy of the New York Times.

The rules of Wordle are simple: players need to guess a word that consists of five letters with no hints, and each person has six chances per day. If both the letter and the location are correct, the color of the tile is shown as green; if the letter is correct but the location is not, the color of the tile is shown as yellow; if the letter does not exist in the answer, the color of the tile is shown as gray.An example of how wordle is played is shown below:

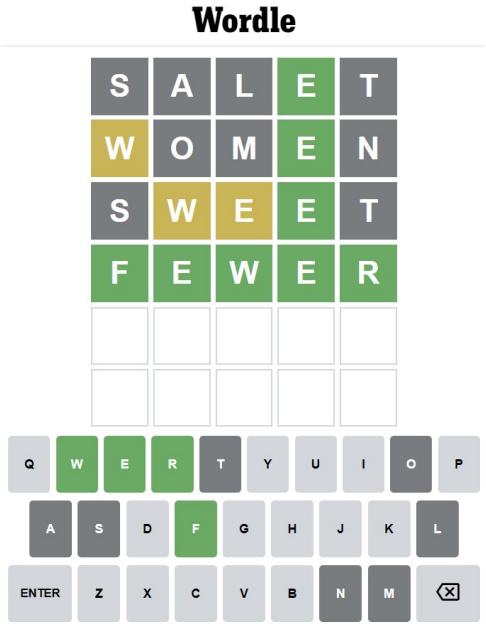

Figure 1: Example of how to play[1]

Additionally, players can choose between normal and hard modes. In hard mode, once the player finds a correct letter, which can also be said to appear as a yellow or green tile, each subsequent guess must result in the inclusion of those words, thus making the game more difficult. The results provide a rough idea of the player's word stock and familiarity with English word spelling rules.

1.2 Restatement of the Problem

Through in-depth analysis and research of the background of the problem, in conjunction with the data file, and taking into account the specific constraints of the question, the restatement of the problem can be expressed as follows:

- **Task 1** Develop a model to analyze the change in the number of reported results and use the model to predict the interval for the number of reported results on March 1, 2023. Determine whether the attributes of words affect the choice of game mode and explain why.

- **Task 2** Create a model for predicting the relevant percentage of different guesses for a given word on a given day. Use the model to predict the reported outcome scores for the word EERIE on March 1, 2023, to validate the accuracy of the model.
- **Task 3** A categorization model was developed to classify solutions based on difficulty, identifying the attributes of the words associated with each classification. The accuracy of the model is discussed by analyzing the difficulty of the word EERIE.
- **Task 4** Analyze and characterize other features of the data file.

1.3 Data Pre-processing

By observing and analyzing the data, the following anomalies are identified:

- **Word spelling errors** Words in data set 314, 525, and 545 consist of four or six letters that do not meet the requirement of "five letters and only five letters". In addition, words in data set 545 contain non-letter symbols. These words are considered to be misspellings, and the necessary corrections have been made to make them both relevant and meaningful.
- **Percentage error** In data set 281, the percentage sum of the reported results is significantly different from 100%. Therefore the data in this group is not meaningful. The original correct data is not available.
- **Quantity error** In data set 529, the number of reported results differ too much from the other sets and is considered to be an input error. This is dealt with by Lagrange interpolation.

The specific data processing results are shown below:

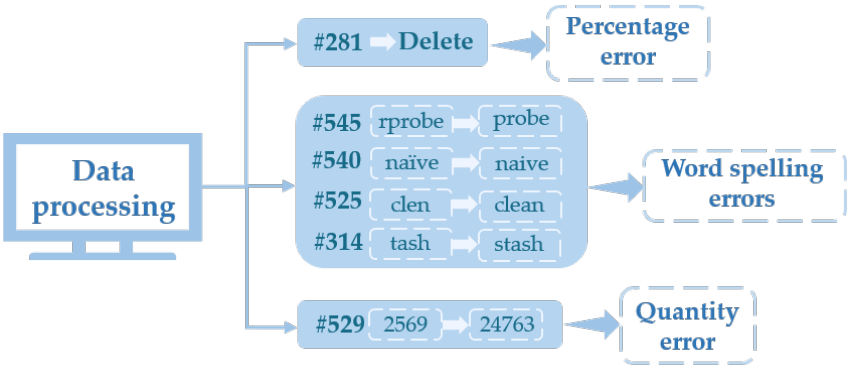

Figure 2: Data processing results

1.4 Our Work

Our work mainly includes the following:

Step 1 Analyze trends in the number of reported results using Life Cycle Theory and build an ARIMA model to create prediction intervals for the number of results reported on March 1, 2023.

Step 2 Determine the words' attributes and analyze whether they affect the percentage of scores in the hard mode.

Step 3 Classify the solution words according to their difficulty through the AHP model. Identify the word attributes of EERIE and determine its difficulty through the BP neural network model.

Step 4 Predict the percentage of correlation between different numbers of tries at a future date by calculating the Euclidean distance between word attributes. Predict the distribution of EERIE reported results using the model developed.

Step 5 Find other interesting features of changes in the number of resulted reports and describe them.

In order to avoid complicated description, intuitively reflect our work process, the flow chart is shown in Figure 3

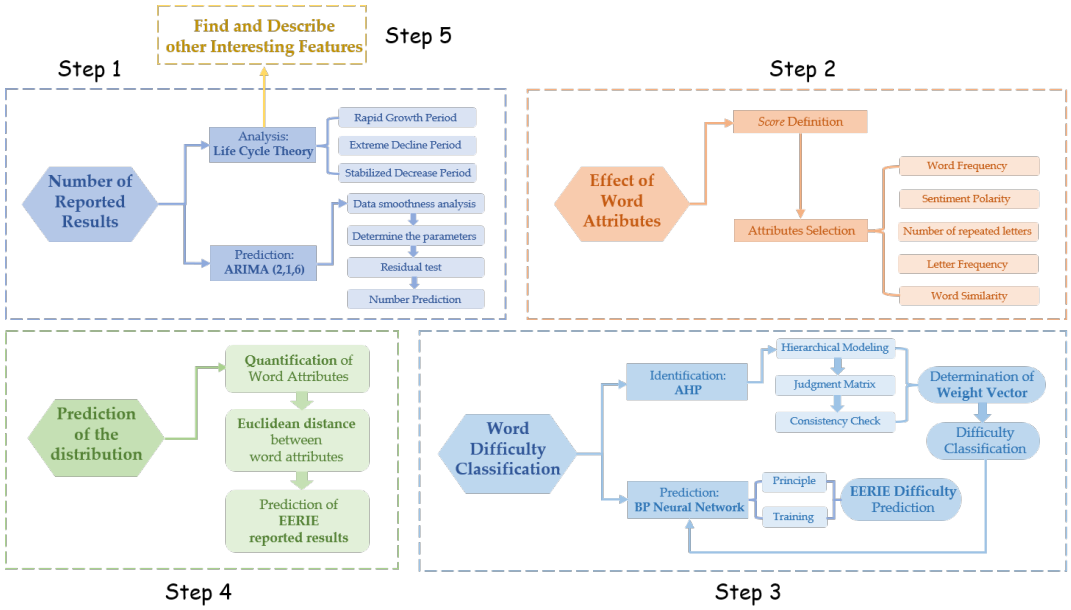

Figure 3: Flow Chart of Our Work

2 Assumptions and Notations

2.1 Assumptions and Explanations

Considering that practical problems always contain many complex factors, first of all, we need to make reasonable assumptions to simplify the model, and each hypothesis is closely followed by its corresponding explanation:

Assumption **1: The number of players is the same as the number of reported results, and the number of different tries is the same as a percentage of the total number of players as a percentage of the total number of reported results.**
Explanation: The assumption suggests that not all users will report results on Twitter. This makes the number of reported results lower than the number of players. This makes it possible to analyze the trend in the number of players to get a trend in the number of results reported.

Assumption **2: All players have normal spelling ability and the uploaded report scores are real and reliable without cheating.**

Explanation: To ensure that all the data used in the model analysis are real and reliable and that the calculations are relevant.

Assumption **3: Wordle's development process and the Product Life Cycle Theory are basically a good fit.**
Explanation: Exclude the effect of extraneous changes in the external environment on the number of wordle players. Only the development of this game in its natural state is considered.

Assumption **4: The results of the data pre-processing are all reasonable.**
Explanation: As the methods of pre-processing the data are not unique, different processing methods may have different impacts on the analyzed results. Since the percentage of anomalous data is very small, it can be assumed that the impact on the results is negligible.

Assumption **5: The total tries percentage is the same as the tries percentage in hard mode.**
Explanation: To simplify the analysis of Wordle's percentage scores in reported results in the hard mode, it is assumed that the percentage of tries is approximately the same in hard mode and all modes.

Additional assumptions are made to simplify analysis for individual sections. These assumptions will be discussed at the appropriate locations.

2.2 Notations

Some important mathematical notations used in this paper are listed in Table 1.

3 Analysis and Prediction of Number Based on ARIMA Model

3.1 Explanation of Changes in the Number of Reported Results

By analyzing the trend of the number of reports in Wordle in the data file, it has gone through three stages of rapid growth, extreme decline, and stable decrease, as shown in Figure 4. Since it has been assumed in the Assumptions section that the reported numbers can reflect the actual number of players, the explanation of the reasons for the change in the number of players also applies to the explanation of the change in the reported numbers.

Combined with the Product Life Cycle Theory proposed by Raymond Vernon[2], the following descriptions and explanations of the phenomenon of player number changes are given:

1. Rapid Growth Period (January 7, 2022-February 2, 2022)

In this stage, the number of players grows from 80,630 to 361,908, showing an explosive growth trend. The phenomenon mainly stems from the following three reasons:

Table 1: Notations

Symbol	Description
μ	The constant term
ϵ_t	The error term
γ_i	Autocorrelation coefficient
y_{t-i}	Historical reported data
θ_i	Correlation coefficient
ϵ_{t-i}	Historical reported data
y_{t-i}	Historical data error
n	Sample size
L	The great likelihood function
K	Number of model parameters
t	Number of tries
p_i	Percentage corresponding to the different number of tries
$\lambda_{\max}$	The maximum eigenvalue of the judgment matrix
n	The order of the judgment matrix
w_i	Weights of neuron
w_i	Weighted sum of the input signals
$\theta_{(i)}$	Threshold
f	Activation function
w_i	Weights of neuron
w_i	Weights of neuron

*There are some variables that are not listed here and will be discussed in detail in each section.

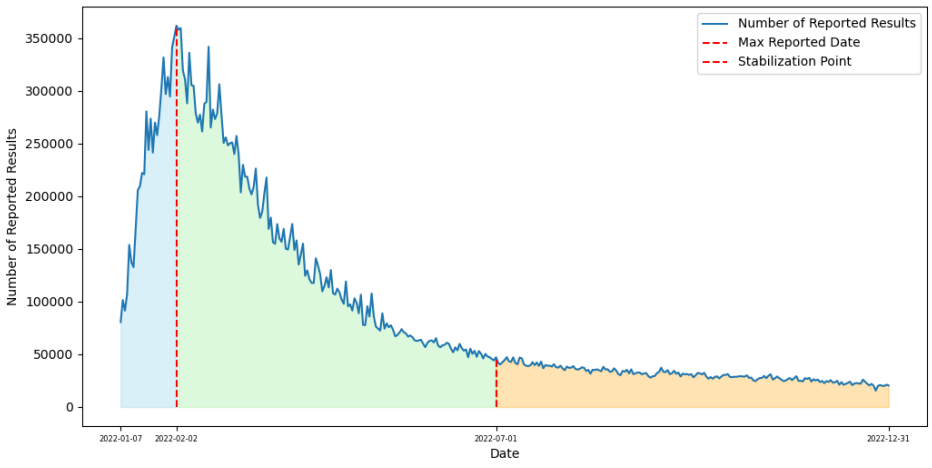

Figure 4: Changes in the number of results reported

- **Easy Entry** Wordle focuses on simplicity and accessibility in its game design, allowing people to quickly understand the rules and start playing. This ease of entry helps attract a large number of new players during the game's initial market launch.
- **Challenging** Wordle's once-a-day update frequency makes the game challenging. As a result, players are willing to keep trying to improve their game. The challenging design allows players to immerse themselves in the game and generate sustained interest.
- **Word of mouth and social media promotion** Many Wordle players are happy to share their gaming experiences and upload report scores on social media such as Twitter, thus creating a spreading effect. Sharing, commenting, and retweeting on social media can also expand the game's popularity in a short period and attract more players to participate.

2. **Extreme Decline Period (February 3, 2022-July 1, 2022)**

   In this stage, the number of players decreases from 361,908 to 47248, showing a precipitous decline. The reasons include the following two points:

   - **Loss of novelty** Due to the simple operation of the game, once the rules of the game are mastered and a certain level is reached, the sense of novelty is gradually lost after some time, and players may feel that the game is less attractive, leading to the loss of interest in the game.
   - **Emergence of competitors** After Wordle's popularity, a large number of imitations emerged in the app store, some charging as much as $30 for a subscription. In addition, the large number of similar games in the puzzle game market also distracted players.

3. **Stabilized Decrease Period (July 2, 2022-December 31, 2022)**

   In this stage, the number of Wordle players has decreased only slightly over a longer period of time and has gradually leveled off.

   This is the natural state of the game as it enters the maturity stage. At this stage, the game has attracted a major and relatively stable target user base. However, the number of players is still gradually decreasing due to factors such as time and changing interests. Social factors may no longer be the main driving force, while player satisfaction and game depth become the core factors for users to choose to stay.

The above analysis also applies to explain the reasons for the change in the number of Wordle report results.

3.2 ARIMA Model of First-order Difference

In this section, the prediction of the number of reported results on a given date is carried out using the ARIMA Model of First-order Difference, which is overviewed as follows:

3.2.1 Identification of the Model

The model uses data from 358 sets of reported results uploaded on Twitter from January 7, 2022, to December 31, 2022, by Wordle players. As can be seen from the

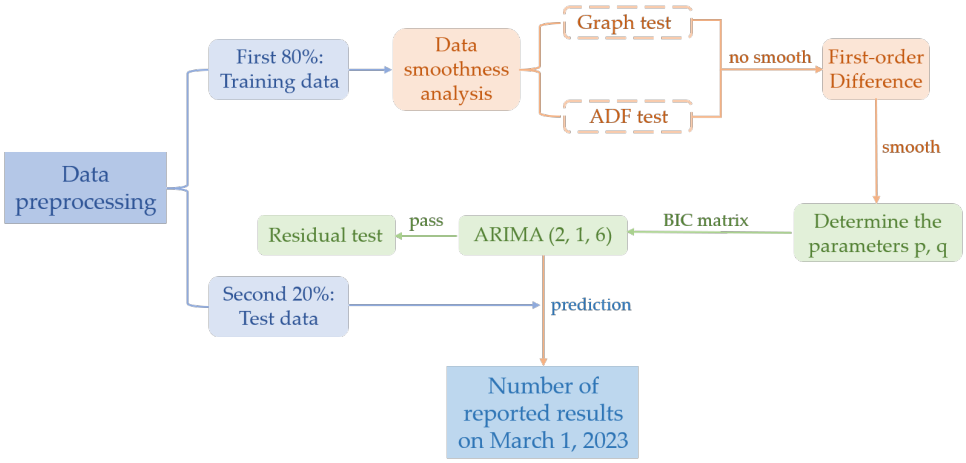

Figure 5: Flow chart of the model

data file, the number of reported results is in chronological order. The data is collected once a day and the sample data is large. Therefore, the sample data are characterized by discrete and equal intervals, which can constitute a period time series. Therefore Time Series Analysis Method is adopted to predict the number of results on March 1, 2023.

The Autoregressive Sliding Average Model (ARIMA) is one of the most commonly used forecasting models for smooth time series. If the following equation is satisfied:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{1}$$

where μ is the constant term, ϵ_t is the error term, γ_i is the autocorrelation coefficient, y_{t-i} is the historical reported data, θ_i is the correlation coefficient, and ϵ_{t-i} is the historical data error.

Then y_{t-i} is said to be an autoregressive sliding mixed average process, p is the order of the autoregressive process, and q is the order of the sliding average process, represented as ARIMA(p,q).

If the original time series is differentiated by d times to form a smooth ARIMA process, an ARIMA(p,d,q) model can be constructed.

3.2.2 Model Establishment and Solution

First, data smoothness analysis is performed. From the Figure 4, the time series is a non-smooth time series.

Therefore, the first-order difference is performed on the original time series to obtain the new time series. The ADF(Augmented Dickey-Fuller Test) test shows that the p-value, which is the probability value corresponding to the T-statistic, is 0.0049, which is less than 0.05. This indicates that the time series after the first-order difference is smooth, as shown in the figure below.:

Therefore, the time series is modeled using the ARIMA model with first-order difference.

Second, the BIC(bayesian information criterion) is used to determine the parameters p,q in the ARIMA(p,d,q) model.BIC is more objective and takes into account the

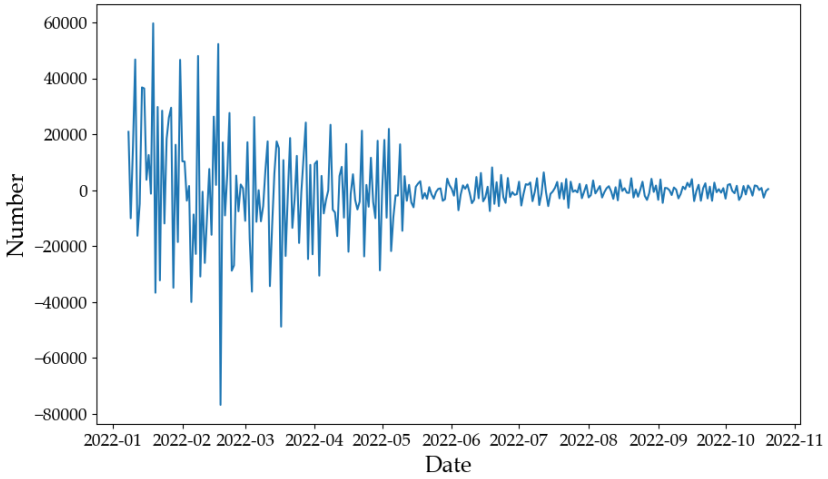

Figure 6: ADF test results

sample size. Thus the values of p,q can be determined more accurately.

The formula for calculating BIC is shown below:

$$\mathrm{BIC} = -2\ln(\mathrm{L}) + \mathrm{K}\ln(\mathrm{n}) \tag{2}$$

where n represents the sample size, L represents the great likelihood function of the model, and K represents the number of model parameters.

The BIC matrix obtained from this equation is shown below:

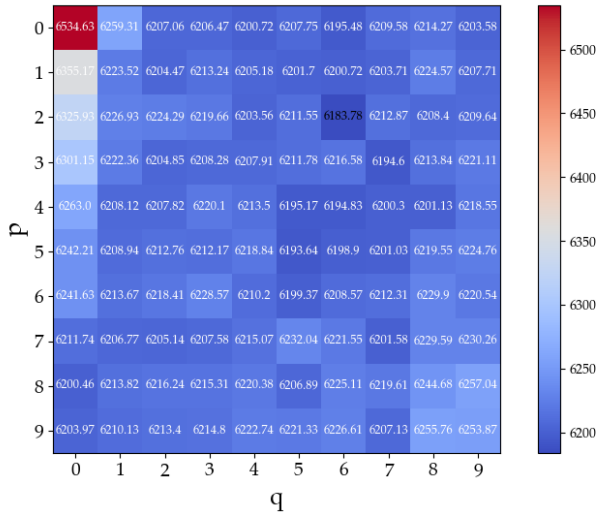

Figure 7: BIC matrix

In the Figure 7, each cell corresponds to a set of selected variables. The darker the blue color of the cell, the lower the BIC value, the better the model fit. From the Figure 7, it can be seen that when p=2, q=6, the BIC value is minimized and the ARIMA (p,q) model is optimal. Therefore, ARIMA (2, 1, 6) is chosen to build the time series model.

Finally, a residual test is performed to check the fit of the model. The residuals are equal to the actual values minus the estimated values. A correctly identified model has residuals that approximate the characteristics of white noise.

The Autocorrelation and Partial Autocorrelation plots of the residuals of the fitted model are shown below:

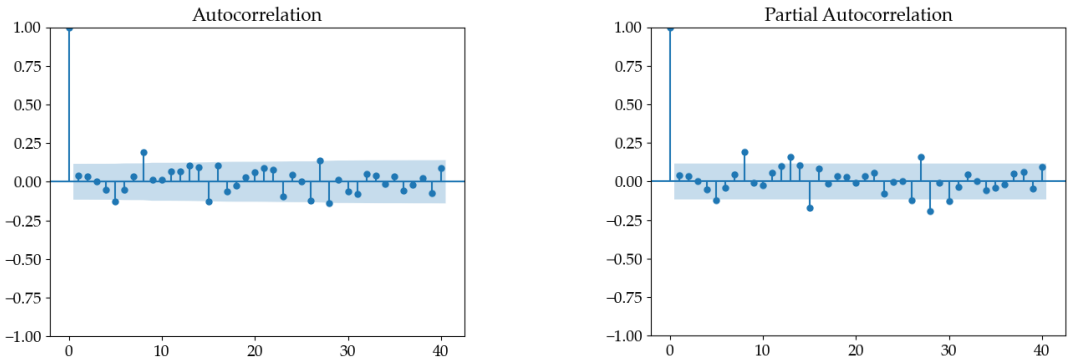

Figure 8: The Autocorrelation and Partial Autocorrelation plots

From the figure, it can be seen that most of the points in its Autocorrelation and partial Autocorrelation plots lie in the confidence intervals, indicating that the independence of the data columns is high. Therefore the residuals of the fitted model satisfy the properties of white noise, indicating that ARIMA (2, 1, 6) passes the residual test.

3.3 Prediction of results

After identifying the ARIMA (2, 1, 6) model, a prediction of the number of reported results of Wordle after December 31, 2022, is performed, and the prediction is shown in the following figure:

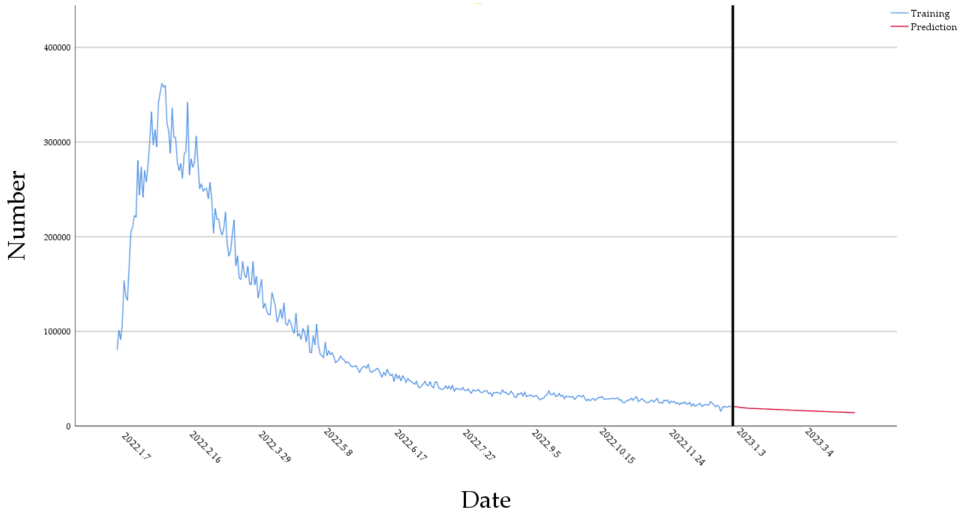

Figure 9: Prediction of results

From the Figure 9, the prediction for March 1, 2023 is 15,042 people. The prediction interval is [7128,28106].

4 Effect of Word Attributes on Score Percentage

4.1 Definition of $Score$

Assuming that the score is defined by the number of tries, the percentage of the score is determined by the percentage of the number of tries. Based on Assumption 5, the percentage of total tries is used to represent the percentage of tries in the hard

mode. To synthesize the percentages of different scores in the hard mode, the $Score$ in this section is defined by both the number of tries and its percentage.

In the data file, the percentages corresponding to one to six tries are easy to derive. However, when the number of tries is more than six, the guess is considered to have failed. Therefore the specific number of tries represented by 7(X) is not available and is not of practical significance. Moreover, when the percentage corresponding to the number of tries from one to six is determined, the percentage corresponding to 7(X) is subsequently determined. Therefore, 7(X) and its corresponding percentage are not considered.

Taking all these factors into account, the $Score$ is defined by the formula shown below:

$$Score = \sum_{i=1}^{6} t \cdot p_i \qquad i \in \{1, 2, \ldots, 6\} \tag{3}$$

where t represents the number of tries, and p_i represents the percentage corresponding to the different number of tries.

4.2 Selection of Word Attributes

Word attributes may affect $Scores$ in the hard mode. By reviewing the literature[3], five word attributes including word frequency, Sentiment Polarity, number of repeated letters, letter frequency, and word similarity are selected, as shown in Figure 10.

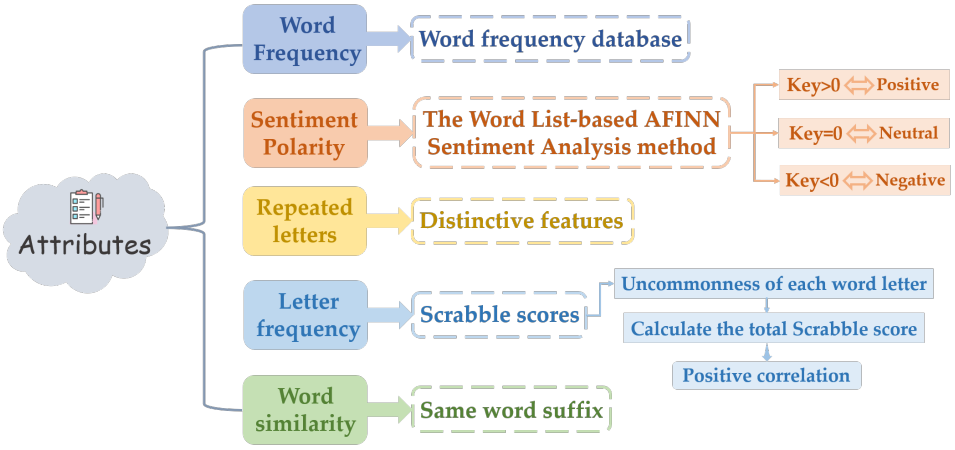

Figure 10: Five word attributes

By analyzing the correlation between the selected word attributes and the $scores$, it is possible to obtain the extent to which the word attributes influence the percentage of $Scores$ in the hard mode.

4.3 Analysis of Results

Import the database into Python and analyze the words and percent of tries in the Wordle data file. Using each attribute of the word as an independent variable and the $score$ as the dependent variable, scatter plots are drawn.

The Figure 11 shows that word frequency has no significant effect on $Score$.

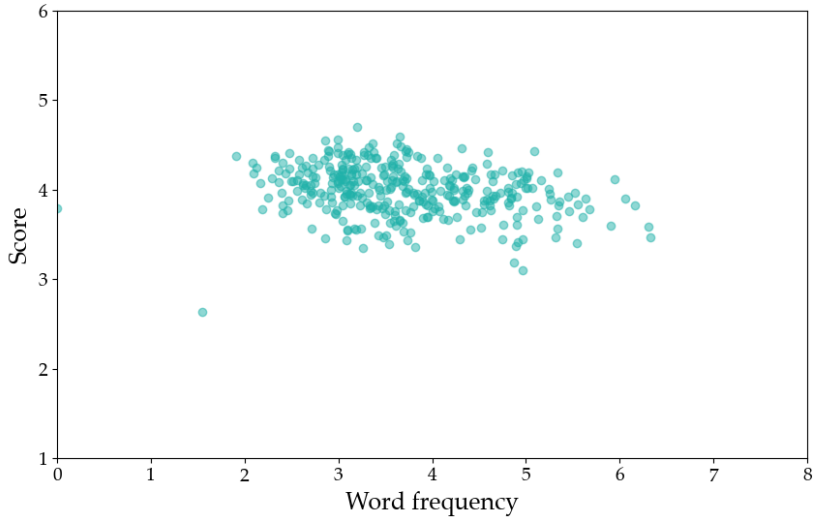

Figure 11: Word frequency

In addition, Sentiment Polarity, number of repeated letters, letter frequency, and word similarity all have varying degrees of effect on $Score$. The specific effects are as follows:

1. **Sentiment Polarity is partially and weakly positively correlated with $Score$.**

   From Figure 12, there is a tendency for the minimum value of the $Score$ to increase as the Polarity of the word increases, both positive and negative. However, this trend is not significant. Moreover, the word with the highest $Score$ has zero Polarity. Therefore,Sentiment Polarity affects the percentage of $Scores$ in the hard mode, but it is not significant.

2. **The number of repeated letters is positively correlated with $Score$.**

   From Figure 13, there is an upward trend in the $Scores$ when the number of repeated letters increases. When the number of repeated letters reaches three, the $Scores$ are all greater than four. Therefore, the number of repeated letters affects the percentage of $Scores$ in the hard mode.

3. **A significant negative correlation is found between letter frequency and $Score$.**

   From Figure 14, when the Scrabble score increases, there is a significant upward trend in the $Score$. There is a negative correlation between letter frequency and Scrabble score. Therefore, letter frequency and $Score$ are negatively correlated. From this, it can be learned that when the letter frequency increases, the number of tries decreases significantly, and guessing will be easier.

4. **Word similarity shows a significant positive correlation with $Score$.**

   From the Figure 15, when the value of similarity of the word is 1, it means that the specified common suffix is present in the word. When the similarity of the word increases, the $Score$ tends to be upward. Moreover, when the value of word similarity is 1, the score is generally higher. This shows that when a common suffix is present in the word, the number of tries increases significantly.

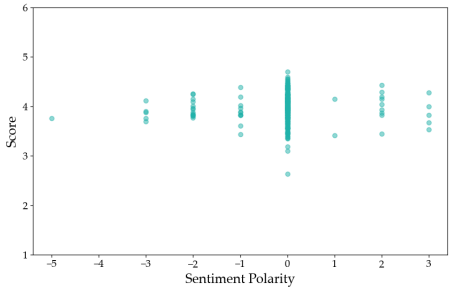

Figure 12: Sentiment Polarity

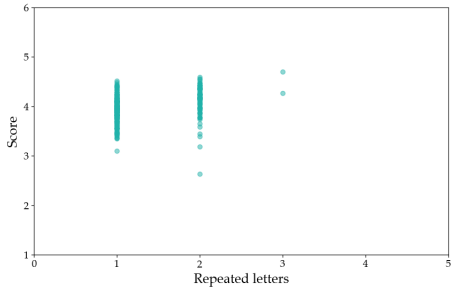

Figure 13: Repeated letters

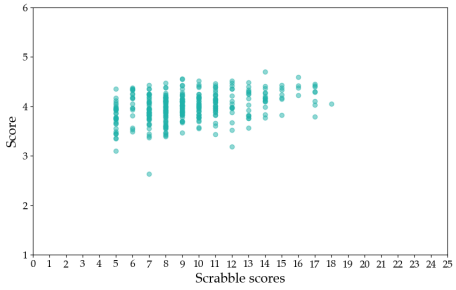

Figure 14: Letter frequency

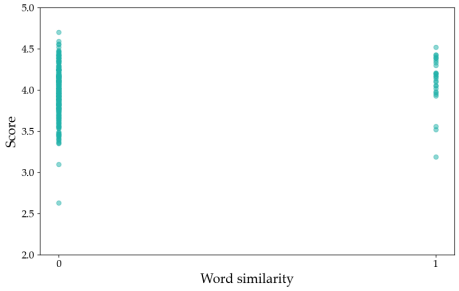

Figure 15: Word similarity

5 Word Difficulty Classification Model Based on AHP

5.1 Identification of word difficulty with AHP

When considering responses to word difficulty, different numbers of tries are of different importance. For instance, it is reasonable to consider the word easier when the number of tries is greater for 1 or 2 tries while the number of tries is less for 6 tries or for not solving the problem. Therefore, using AHP, a score is obtained through subjective calculations that serve as a basis for categorizing the difficulty of the word.

5.1.1 Hierarchical Model Building and Checking

Hierarchical Modeling

The main common evaluation-type methods are the coefficient of variation method and the analytic hierarchy process. The former is calculated objectively to obtain the weights assigned to each evaluation indicator, while the latter is calculated subjectively. When the importance of evaluation indicators is different, the latter is more applicable. The AHP constructs a two-by-two comparison matrix and finds the maximum eigenvector of the matrix, with which the importance of the indicators can be expressed.

Therefore, in this section, AHP is chosen to identify and evaluate word difficulty. The hierarchical model is established by taking the number of tries as the criterion layer and identifying the word difficulty as the object layer, as shown in the following figure:

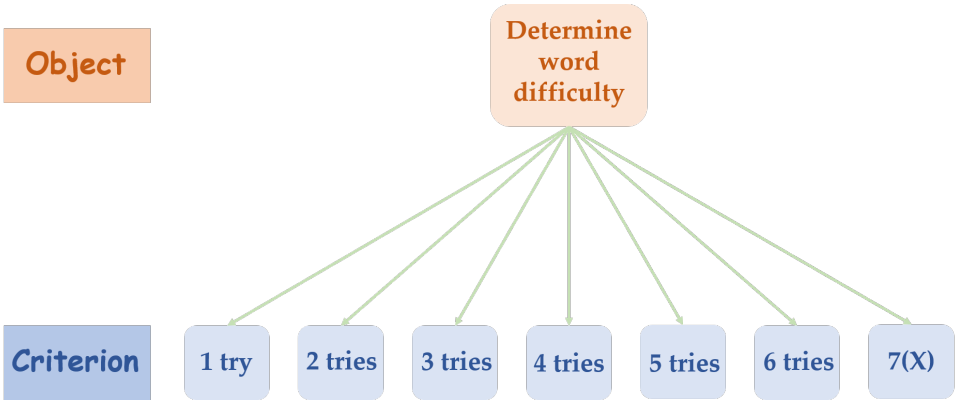

Figure 16: The hierarchical model

Construction of the Judgment Matrix

AHP finds the maximum eigenvector of the matrix by constructing a two-by-two comparison matrix. It can indicate the degree of importance of the indicator.

Through correlation analysis of different tries, it is found that the percentage of the number of people who tried once, twice, three times, and four times has a certain correlation, the percentage of the number of people who tried five, six and more times has a certain correlation, and the correlation of four and five times is extremely low. Therefore, the percentage of the number of one, two, three, and four tries is taken as a reflection of word ease, and five, six, and more tries as a reflection of word difficulty. When the proportion of fewer tries is large, the word is considered easy. When the proportion of more tries is large, the word is considered difficult.

In conclusion, the Judgment Matrix is determined as follows:

$$\begin{bmatrix} 1 & 3 & 5 & 7 & 7 & 3 & 1 \\ \frac{1}{3} & 1 & 3 & 5 & 5 & 1 & \frac{1}{3} \\ \frac{1}{5} & \frac{1}{3} & 1 & 3 & 3 & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{7} & \frac{1}{5} & \frac{1}{3} & 1 & 1 & \frac{1}{5} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{5} & \frac{1}{3} & 1 & 1 & \frac{1}{5} & \frac{1}{7} \\ \frac{1}{3} & 1 & 3 & 5 & 5 & 1 & \frac{1}{3} \\ 1 & 3 & 5 & 7 & 7 & 3 & 1 \end{bmatrix}$$

Hierarchical Ordering and Consistency Checks

Hierarchical ordering and consistency checks are performed to test whether the constructed judgment matrices are close to the consistency matrices.

First, the consistency index CI is calculated.The formula for CI is shown below:

$$CI = \frac{\lambda_{max} - n}{n - 1} \tag{4}$$

where λ_{max} is the maximum eigenvalue of the judgment matrix and n is the order of the judgment matrix. The result of the calculation is $CI = 0.034$.

Second, by consulting the average random one-time index table, when $n = 7$, the average random one-time index $RI = 1.32$.

Finally, the consistency ratio CR is calculated.The formula for CR is shown below:

$$CR = \frac{CI}{RI} \tag{5}$$

The result of the calculation is $CR = 0.026 < 0.1$. Therefore, the judgment matrix passes the consistency check.

5.1.2 Determination of the Weight Vector

The weight vector is a reflection of the importance of the indicator to the result. According to Saaty's research results, the maximum eigenvector of the judgment matrix can be used as the weight vector after normalization. To make the word difficulty quantifiable, the first four numbers of the weight vector are taken as opposite numbers to get the modified weight vector.

The original weight vectors and the modified weight vectors are shown in the table below:

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7(X) tries
Original value	0.29817735	0.1364467	0.06667492	0.03203849	0.03203849	0.1364467	0.29817735
Modified value	-0.29817735	-0.1364467	-0.06667492	-0.03203849	0.03203849	0.1364467	0.29817735

Table 2: Weight vectors before and after modification

By multiplying the weight vectors by the percentage of each number of tries, a quantitative metric is obtained. The metric is positively correlated with the word's difficulty. All words are ranked in descending order of the metric. The first 180 words are defined as "Easy", and the remaining words are defined as "Hard". Thus, the classification of word difficulty has been completed.

5.2 EERIE Difficulty Prediction Based on BP Neural Networks

5.2.1 Principle of BP Neural Network

Based on the word attributes, it can determine a word's difficulty classification. The artificial neuron model is presented, and the neural model of its basic unit is shown below:

where x_1,x_2,x_3,x_4..., x_p represent the input signals. By inputting the word attributes, the output value $y_{(i)}$ is obtained. From this, the difficulty level of the word can be obtained.

The artificial neuron model works on the basis of a set of weights to get the input signal cumulative value. Then using a nonlinear activation function, the cumulative value is mapped to a limited range. The working principle can be expressed mathematically as:

$$u_i = \sum_{j=1}^{k} w_{ij} x_j \tag{6}$$

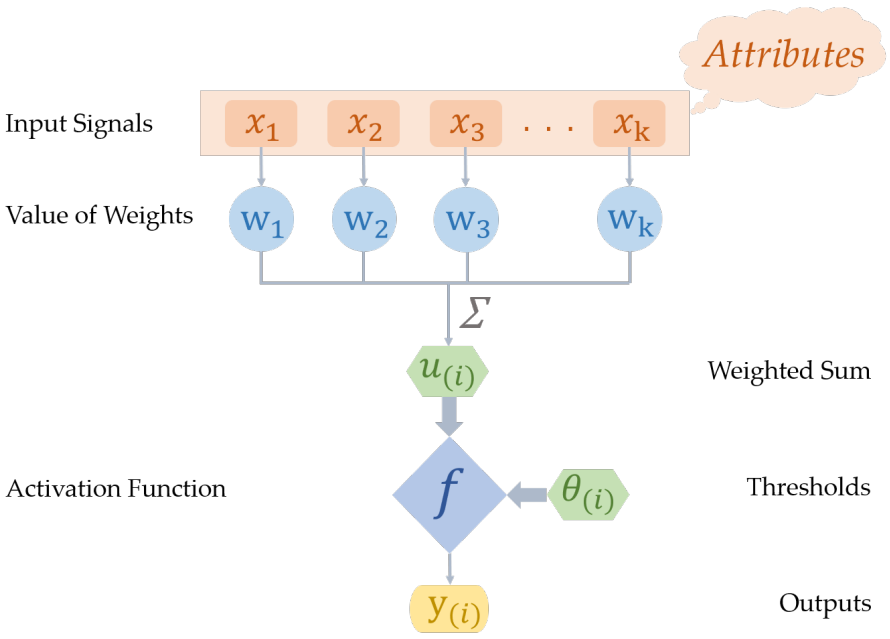

Figure 17: The basic unit of the artificial neuron model

$$v_i = u_i - \theta_i \tag{7}$$

$$y_i = f(v_i) \tag{8}$$

where w_1, w_2, w_3, w_4..., w_k represent the weights of neuron, $u_{(i)}$ represents the weighted sum of the input signals, $\theta_{(i)}$ represents the threshold, and f represents the activation function, which includes the segmented linear function, the step function, and the sigmoid function, etc. Each layer of the BP neural network consists of several neurons as shown in Figure 17.

The BP neural network learning algorithm works in the following steps: first, initialize the network and learn the parameters. Based on this, training patterns are provided to train the neural network.The BP neural network continuously modifies the weights by analyzing the error between the output value and the desired value. Finally, when the error meets the requirements, the results are output.

BP neural network is used to solve the difficulty classification problem of words due to its high fault tolerance and advantages such as self-learning, self-organization, and self-adaptation capabilities.

5.2.2 EERIE Difficulty Prediction Based on BP Neural Networks

The attributes of the words selected in the previous section: Sentiment Polarity, number of repeated letters, letter frequency, and word similarity are used as inputs (independent variables), and the level of difficulty is introduced into the neural network workbox as the target (dependent variable). The samples are divided into two classifications for processing: training set (80%), and test set (20%). The hidden layer is selected as the first layer. The activation function of the hidden layer is the hyperbolic tangent function.

In this model, the training pattern of the BP neural network model for four neurons with four input attributes is shown below:

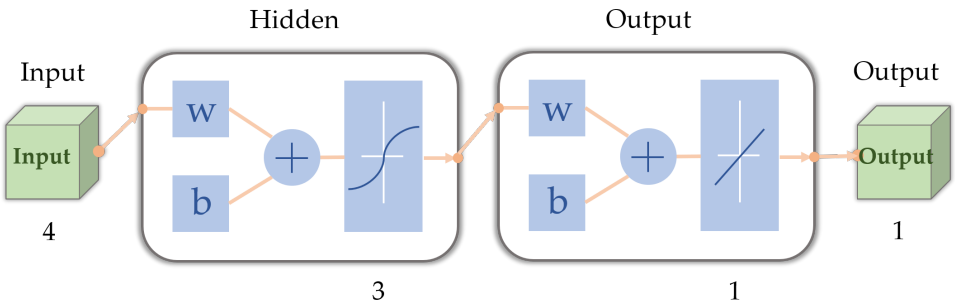

Figure 18: BP neural network model with four neurons

After training, the model ends up with greater than 70% correctness for both the training and test sets and can be used for prediction.

Enter the attributes of the word "EERIE", the likelihood that the difficulty classification of EERIE is " Hard" is 85%.

6 Prediction of the Distribution Based on Euclidean Distance

6.1 Modeling and Prediction

Word attributes are related to word difficulty, and word difficulty affects the percentage of tries in the reported results. Since the correlation between the four selected word attributes is weak and these attributes can be quantified, Euclidean distances between word attributes are used to build a distance model that characterizes the similarity of the word attributes and thus predicts the percentage. The formula for calculating the Euclidean distance between word attributes is shown below:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2} \tag{9}$$

where x_1, x_2, x_3, x_4 are quantifiers of the attributes of one word and y_1, y_2, y_3, y_4 are quantifiers of the attributes of another word.

The nearest neighbor interpolation of the two words closest to "EERIE" can be used to calculate the percentage of tries in the reported results for "EERIE". The words "MUMMY" and "FLUFF" are calculated to be closest to "EERIE".The results of the calculations are shown below:

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7(X) tries
MUMMY	0	1	4	14	27	37	18
FLUFF	0	0	4	25	44	23	4

Table 3: Percentage of tries of MUMMY and FLUFF

Therefore, the mean value of the percentage of tries in the reported results for these two words is selected for the interpolation fit.

The percentage of tries in the reported results of "EERIE" is calculated as (0, 1%, 4%, 19%, 35%, 30%, 11%). Hence, the prediction of the percentage of tries in the reported results of "EERIE" is completed.

6.2 Accuracy Analysis

The model is considered to be sufficiently rigorous and accurate in the following ways:

1. Word attributes are rigorously screened and analyzed for relevance. An attribute is filtered out when it is determined that there is a sufficient correlation between the attribute and the difficulty of the word.
2. The quantification of word attributes has a systematic mathematical approach and a rigorous process.
3. It is reasonable to characterize the similarity of attributes by calculating the Euclidean distance of the attributes. When the attributes of two words are close to each other, it is reasonable to assume that they have the same or similar difficulty.

However, the model also has limitations, which are summarized below:

1. Due to the possibility that some people did not upload their report results, there is an inaccuracy in the percentage of report results for MUMMY and FLUFF.
2. Word attributes are not entirely reasonable, and there may be some difficulty-related attributes that have not been discovered.

7 Other Characteristics of the Data

By analyzing the dataset, we also found some other interesting features, which are described as follows:

1. When the percentage of six or more tries for one word is greater than that for the other, usually the percentage of less than or equal to two tries for that word will be less than that for the other. In addition, the sum of the percentages of three and four tries for the word is less than the other, and the sum of the percentages of five and six tries is more than the other.
2. There is a monotonically increasing percentage of the number of results reported for one, two, three, and four tries, and a monotonically decreasing percentage of the number of results reported for five, six, and more tries.
3. The number of reports of hard mode as a percentage of the total number of reports has gradually increased over time.
4. The percentage of passes in a single try rarely exceeded 5%, and the percentage of reports of game failures rarely exceeded 10%. This suggests that the game is relatively simple but somewhat challenging.

8 Sensitivity Analysis

To test the sensitivity of the prediction model for EERIE difficulty prediction, sensitivity analysis is performed on the number of hidden layer neurons in the BP neural network. The number of hidden layer neurons varies from 3 to 14, and then the model

is trained. The trained models were used to predict the difficulty of EERIE separately. It is found that all the models predict the difficulty of EERIE with Hard. The corresponding Hard probabilities are shown below:

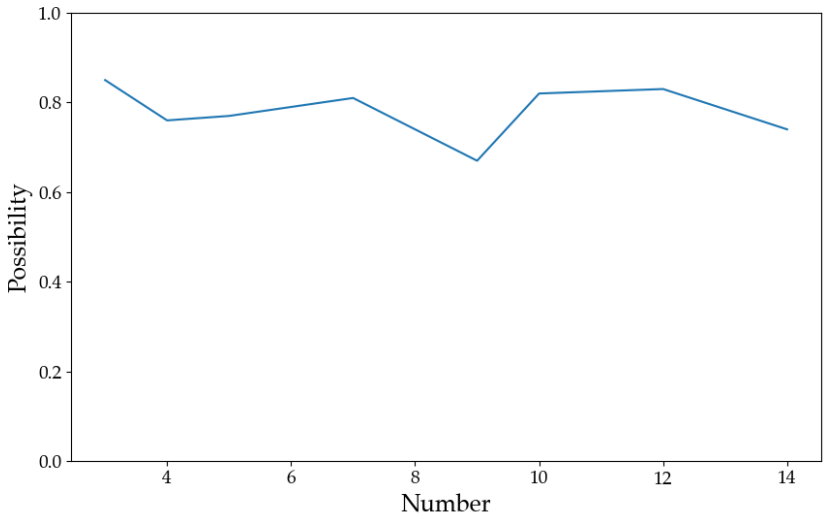

Figure 19: Probability of prediction as Hard

From the Figure 19, when the number of neurons in the hidden layer of the BP neural network is changed, the probability of EERIE being judged as Hard fluctuates but remains roughly stable. It indicates that changing the number of neurons in the hidden layer has little effect on the stability of the model. The output of the model is not particularly sensitive to the transformation of the system parameters.

9 Strengths and Weaknesses

9.1 Strengths

- The determination of word attributes has been rigorously screened and supported by sufficient literature and research. This lays a solid foundation for the analysis and prediction of our subsequent models.
- All the databases used in the prediction are from official sources, and the data volume is large and real, which makes the prediction effect of each model more accurate.
- The judgment matrix constructed in the AHP is quite close to the consistency matrix and passes the consistency check. This makes the model more accurate in predicting the number of outcomes reported at future dates.
- The trained BP neural network model has greater than 70% correctness for both the final training and test sets. This allowed the model to accurately predict the distribution of reported results for EERIE.

9.2 Weaknesses

- Since some players did not upload the reported results, there is some deviation between the model's predictions and the true values.
- The use of Euclidean distances to determine similarity in distance models is not robust in some cases.

10 A letter to the Puzzle Editor of the New York Times

Dear Puzzle Editor,

Hello! We are a competition team in the Mathematical Contest In Modeling. In the competition, we are going to investigate some patterns in the Wordle game based on the number of words reported results and the number of people who chose the Hard Mode as well as the percentage of reports of words tried once, twice, three times, four times, all the way up to six times, and the number of game failures, as provided to us by the official of the competition, for each day from January 7, 2022 to December 31, 2022.

The Wordle itself is both fun and challenging. Since we want to study the patterns of the game, we have to try to put ourselves in the perspective of the players: what kind of answers they would enter the first time they play the game, what kind of answers they would enter the next time, what kind of words would be easier for the players, and what kind of words would be more difficult for the players. This process is brain-wracking but necessary because we have to try to summarize the mathematical patterns behind this game. I'm writing now to inform you of some of our findings, and I hope it will be of some use to you.

First of all, according to the analysis of the data and in combination with the Life Cycle Theory, the number of reported results of the game will continue to level off after September 2022. Based on the time series analysis of the available data to predict the number of reports for the game in the future, we have found that the number of reported results in the future will generally decline steadily, and by March 1, 2023, the number of reported results will be in the range of [7128, 28106].

Then, we investigated the words' attributes. After analyzing the correlation between each word's attributes and its difficulty, we finally determined that the word's Sentiment Polarity, the letter frequency, the repeated number of letters, and the degree of similarity affect the word's difficulty. We found that the greater the absolute value of the word's Sentiment Polarity, the greater the number of identical letters, and the greater the degree of similarity, the greater the word's difficulty.

Next, we quantified the difficulty of each word through AHP and classified the word difficulty into two categories, Easy and Hard. In addition, we utilized the BP neural networks to associate word attributes with difficulty and used the model to predict the difficulty of solution words. The model predicts with more than 70% accuracy, which is our success.

Based on these findings, we predicted the difficulty and reported results for the word EERIE. First, we substituted its attributes: Sentiment Polarity index of -2, letter frequency of 5, number of repeated letters of 3, and similarity of 0 into the neural network to obtain its difficulty as Hard. Then, based on the similarity of the attributes,

we found that the most similar words to it are MUMMY and FLUFF by calculating the Euclidean Distance. Therefore, we predicted the percentage of each of the EERIE to be (0,1,4,19,35,30,11)%, using the method of averages. This result was found to be plausible by the test.

These are some of our work on Wordle game pattern summary. We hope that our work will be helpful for your newspaper. If you still want more detailed information, please contact us.

Sincerely,

Team # 2420494

References

[1] https://zh.moegirl.org.cn/Wordle

[2] Vernon, Raymond. “International Investment and International Trade in the Product Cycle.” The Quarterly Journal of Economics, vol. 80, no. 2, 1966, pp. 190–207. JSTOR, https://doi.org/10.2307/1880689. Accessed 24 Jan. 2024.

[3] https://www.waldrn.com/what-makes-a-wordle-word-hard/

[4] https://github.com/rspeer/wordfreq

[5] https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-111.txt

[6] Robyn Speer. (2022). rspeer/wordfreq: v3.0 (v3.0.2). Zenodo. https://doi.org/10.5281/zenodo.7199437