Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2301192

Crack the Wordle Puzzle: Word Attribute Analysis Approaches

Summary

In the past 600 days, a five-letter puzzle game called "Wordle" has been launching a blast of upsurge on Twitter. Wordle players' scores reports are crucial for managers as they provide valuable information for evaluating game difficulty, predicting player numbers and making timely adjustments. To better analyze the reports and provide game improvement suggestions, we conduct in-depth and close studies on this topic from multiple perspectives and levels.

Firstly, to explain the changes in the number of Wordle reports and make predictions, we use an analogy between playing Wordle and the spread of infectious diseases. We compare playing Wordle with infection, players with infected individuals, individuals who have not played Wordle for a long time with susceptible individuals, individuals who have become tired of the game with recovered individuals, sharing on Twitter with transmission, and quitting the game with recovery. Based on these assumptions, we use the SIRS model to fit the curve and explain the overall trend. We also use the Prophet model to insert breakpoints to explain data oscillations and provide a prediction interval for future data. Model evaluation results show that our model has high interpretability and accuracy.

Next, we extract various word attributes from a word database containing a large amount of corpus information and use multiple linear regression to investigate whether there is a relationship between word attributes and Hard-Mode scores. We then test the significance of the model based on the F-statistic. The result shows no significant correlation between these two factors.

Besides, we construct a BP neural network model based on the previously extracted word attributes to predict the distribution of the number of word guesses. The evaluation results show that the model has high prediction accuracy and efficiency, laying a solid foundation for next step analysis.

Furthermore, we use K-means++ clustering algorithm to divide words into three categories: easy, medium, and hard. We analyze the relationship between word attributes and difficulty to classify solution words by difficulty. We find that the difficulty of a word is closely related to the number of different letters in the word, the sum of letter frequencies, and the breadth of usage of the word in different fields, but there is no significant evidence of a correlation between difficulty and word frequency. Combined with the previous BP neural network model, we can accurately classify words.

In addition, we find that common words such as "mummy" and "watch" have a higher guessing difficulty, and there is also a certain correlation between the first letter of a word and its guessing difficulty.

Finally, we present predictive data and improvement suggestions to the editors of The New York Times to assist in improving Wordle and boosting the appealing feature of the game.

Keywords: Prophet; SIRS; Multiple Linear Regression; BP Neural Network; K-Means++

Contents

1 Introduction

1.1 Background

Recently, Twitter has sparked a trend of sharing the Wordle report. Puzzle game developers in the past were often not very clear about the difficulty of their games for the public. Games that are too difficult can be frustrating, while too easy can be boring. With the development of information technology, using big data analysis to control the difficulty of puzzles has become the key to making puzzles more interesting. The New York Times' Wordle game has collected statistics on the number of tries by players and the number of reports on Twitter. This data can be used to evaluate the number of players and the difficulty of a particular word, maintain players' enthusiasm, and make the game more attractive.

1.2 Restatement of the Problem

The New York Times collected 359 days of Wordle player score reports, including report time, number, percentage of difficult mode reports, and number of attempts. To control for gameplay and estimate the number of players, it is necessary to analyze the trend of report numbers, mine information contained in word attributes, and measure the difficulty of words. To achieve these goals, we need to:

- Analyze the reasons for the changes in the number of reports on a large time scale (overall trend) and small time scale (data mutation).
- Collect and mine potential word attributes.
- Analyze whether the percentage of difficult mode reports is related to word attributes.
- Analyze the distribution of attempts and its potential relationship with word attributes.
- Identify the influence of word attributes on difficulty.
- Mine other information that could help improve Wordle.

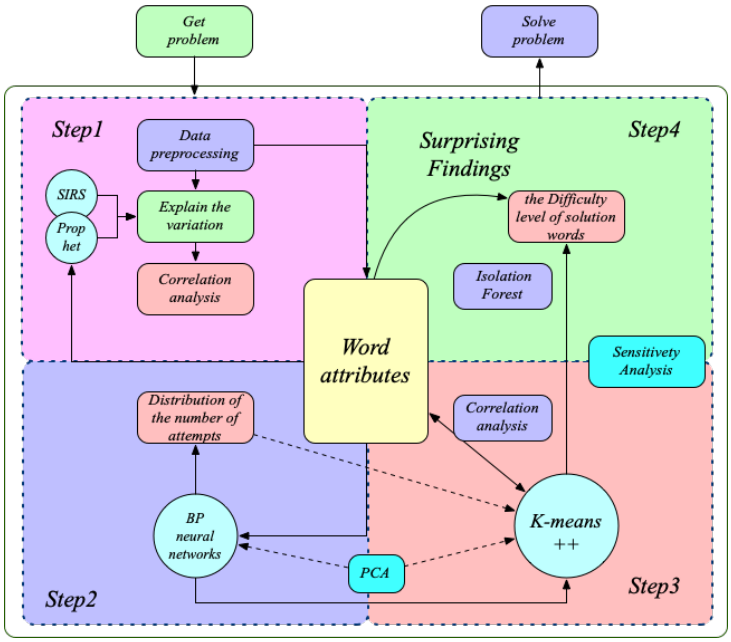

Figure 1: the Flow Chart in this Paper

2 Assumptions and Notations

2.1 Assumptions

To simplify the model, we have made several assumptions. However, we may need to relax some of these assumptions to optimize the model and increase its applicability in complex real-world environments.

• The number of Twitter users is essentially constant, and the probability of each user receiving information related to Wordle is equal.

• All Wordle players are Twitter users, and all Twitter users are potential Wordle players.

• The word for each day in Wordle is completely random and chosen from all five-letter words.

• Players who report their game results on Twitter are a random sample of all players.

• People may get tired of playing Wordle, but they may eventually want to play again after a long time.

2.2 Notations

Table 1: 18 Part-of-Speech Symbols

Symbols	Definition
NN	Noun, singular or mass
JJ	Adjective
RB	Adverb
VBP	Verb, non-3rd person singular present
VBD	Verb, past tense
NNS	Noun, plural
VBN	Verb, past participle
VB	Verb, base form
IN	Preposition or subordinating conjunction
VBZ	Verb, 3rd person singular present
VBG	Verb, gerund or present participle
MD	Modal
PRP	Possessive pronoun
RBR	Adverb, comparative
CC	Coordinating conjunction
JJR	Adjective, comparative
DT	Determiner
JJS	Adjective, superlative

Table 2: Notations of Word Attributes Used in the Paper

Symbols	Definition
Freq	Word Frequency
SLF	the Sum of Letter Frequencies
BU	the Breadth of Usage of a Word
NDLW	the Number of Different Letters in a Word
a-z	the Number of Letters from a to z in a Word

3 Model 1-Integration of Interpretation and Prediction Model based on Prophet and SIRS

3.1 Data Preprocessing and Exploratory Analysis

3.1.1 Data Collection and Pre-processing

In addressing task 1, it is dispensable to analyze the attributes of words related to the problem and collect relevant data.The possible factors include the frequency, the breadth of the usage in different fields, the number of different letters in words and parts of speech. In general Natural Language Processing (NLP), there are 36 commonly used parts of speech[2], of which we selected 18 types relevant to this task as shown in Table 1.

To process missing values, abnormal values and repeated observations in the original data set, we apply a series of data processing methods: data cleaning, **establishment of dummy variables for discrete variables**, **logarithmic transformation of the number of reports** and set-up of new attributes. The four steps enable the elimination of extraneous information and facilitate the identification and extraction of relevant information from the dataset.

Step 1: In the stage of data cleaning, we use Python to check for missing, outlier and duplicate values. By measuring length of words, we check for empty or unusually long values. We find that there are no empty values but three outliers: ”tash”, ”clen” and ”rprobe”. After searching and comparing online, we correct those words as ”trash”, ”clean” and ”probe”. Furthermore, using the ”duplicate()” method, we check for duplicate values with no duplicate value found.

Step 2: To make the discrete variable of part-of-speech easier to be processed by the model, we construct 17 dummy variables to convert the discrete variable into binary variables.

Step 3: We plan to use a time series model to predict the number of reports on March 1, 2023. In these types of models, it is crucial to eliminate heteroscedasticity in the data. Taking the logarithm of the data does not change its nature or correlation, but it compresses the scale of the variable. By shrinking the absolute values of the data, it is easier to eliminate the problem of heteroscedasticity. Therefore, we logarithmically transform the reported quantity.

Step 4: To comprehensively explore the influence of various word attributes on reported Hard-Mode-played scores, we further extract the attributes of words and establish several new variables. This will be elaborated in Section 3.4.

3.1.2 Data Description and Exploratory Analysis

The data is visualized to dig into the inherent rules, which is helpful for modeling. Figure

2 depicts the correlation between the variables, while Figure 3 presents the distribution of the number of attempts in a histogram. Figure 4 displays the changing curve of the total number of reports and the proportion of reports in difficult mode over time.

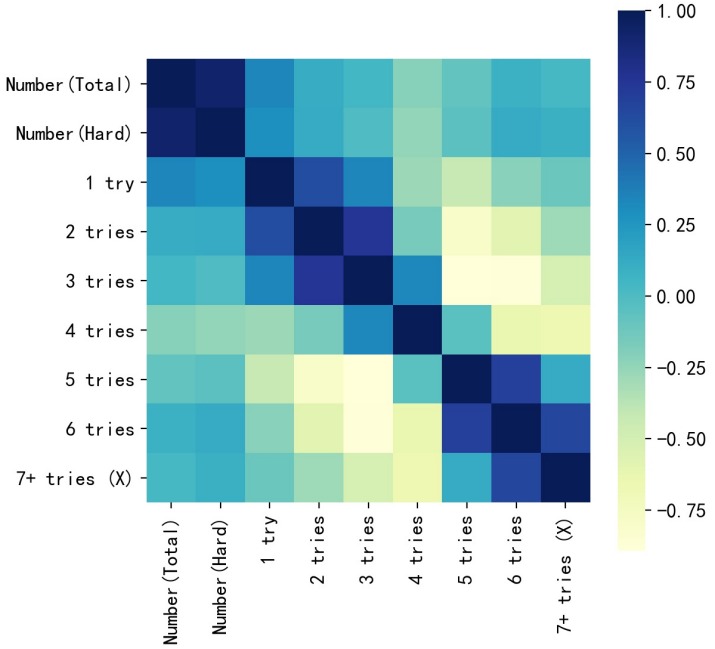

Figure 2: Correlation Matrix

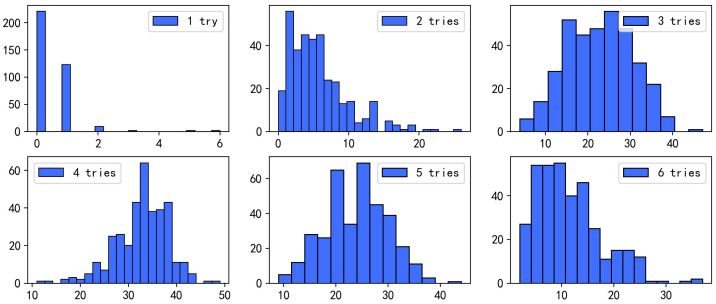

Figure 3: Distribution Histogram

We can see that the correlation between variables is generally weak, and the distribution of the number of attempts shows a state of low at both ends and high in the middle. The trend of the quantity curve is somewhat similar to the infection curve, which will be analyzed in detail in the following steps.

3.2 Prophet Model

The Prophet algorithm provided by Facebook[3] can not only handle time series data with some outliers but also deal with partially missing values. It can almost automatically predict the future trends of time series. Based on time series decomposition and machine learning fitting, it uses the open-source tool pyStan to fit the model, so it can obtain the predicted results quickly.

After performing a logarithmic transformation on the data(elaborated in Section 3.1.1), we use Prophet to establish a multiplication model with the parameters listed in Table 3, where τ is a

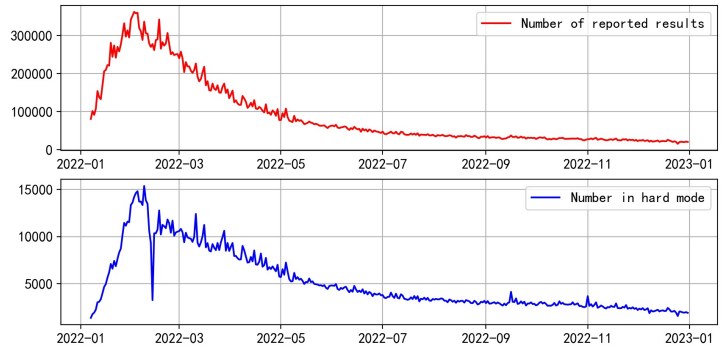

Figure 4: Quantity Curve

parameter that controls the slope of the linear function at the breakpoint.For the rate of change at a changepoint, denoted as Δ, it follows that $\Delta \sim Laplace(0, \tau)$. As τ decreases, Δ approaches 0. Therefore, increasing τ will broaden the upper and lower limits of the predicted values. The trend term uses the default piecewise linear function. Setting more changepoints and increasing the range of breakpoints makes the model more sensitive to changes in time series data, which improves the fitting effect.

Table 3: Prophet Model Parameter Setting

the Number of Changepoints	60	
τ	0.8	
the Range of Changepoints	0.9	
Holidays	Valentine	2022/02/14
	Easter	2022/04/24
	Halloween	2022/10/31
	Thanksgiving	2022/11/24
	Christmas	2022/12/25

A Prophet model typically consists of a trend term $g(t)$, a seasonal term $s(t)$, a holiday effect term $h(t)$ and an residual term $\varepsilon(t)$. $g(t)$ is a piecewise linear function that satisfies:

$$g(t) = (k + a(t)\Delta)t + (m + a(t)^{\intercal}\gamma) \tag{1}$$

where k represents the growth rate, Δ represents the change in growth rate, and m represents the offset parameter. $s(t)$ contains the weekly periodic changes:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right) \tag{2}$$

where P is the period time, and $(a_n, b_n), (n = 1 \ldots N)$ follow a normal distribution. $h(t)$ illustrates the potential impact of holidays on the outcome:

$$h(t) = \sum_{i=1}^{L} k_i * l_{\{t \in D_i\}} \tag{3}$$

where $k_i, i = 1 \ldots L$ follow a normal distribution. Based on the parameters and functions above, a multiplicative model is established:

$$y(t) = g(t) * \ s(t) * \ h(t) * \varepsilon(t) \tag{4}$$

We use the data from 2022-01-07 to 2022-11-21 as the training set and the data from 2022-11-21 to 2022-12-31 as the test set. The fitting result is shown in Figure 5, where the red vertical line represents the breakpoint we set.

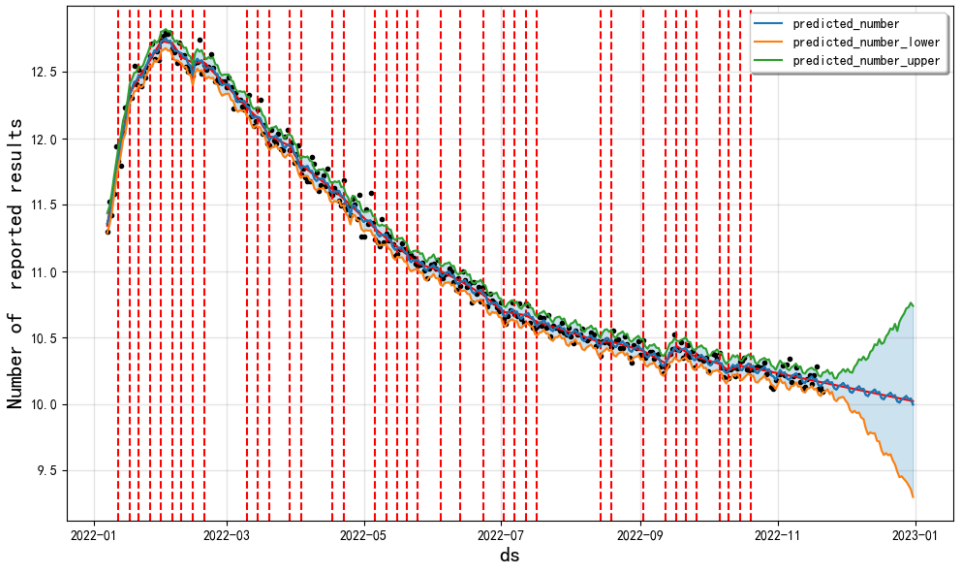

Figure 5: Prophet Forecasting

We evaluate the effectiveness of the model using four metrics: R-squared, MSE, RMSE, and MAPE, which are as follows:

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \\ RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ MAPE &= \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \end{aligned} \tag{5}$$

where $\hat{y}_i$ represents the fitted value and y_i represents the actual value.The results are shown in Table 4. The R-squared value is close to 1, indicating an excellent fit of the model. As our final results are obtained by taking the exponential of the log-transformed data, the small RMSE and MSE can be considered. The MAPE of 4.8% indicates a small average absolute percentage error. Overall, the established model is suitable for prediction.

Based on the data above, we reduce τ to increase the precision of the prediction interval. We then reestablish the model and predict that the number of reported results on March 1, 2023

Table 4: Evaluation of the Prophet

R-squared	MSE	RMSE	MAPE
0.9924	60340502	7767.9149	4.8002

is 14534, with a prediction interval of **(13175, 16128)(95% confidence level)**. These prediction results indicate that the popularity of Wordle is decreasing over time.

3.3 Explanation of the Changes in the Number of Reports

The changes in the number of reports can be decomposed into trend, seasonal, and holiday components as shown in Figure 6. We will explain the changes in report numbers from these three aspects.

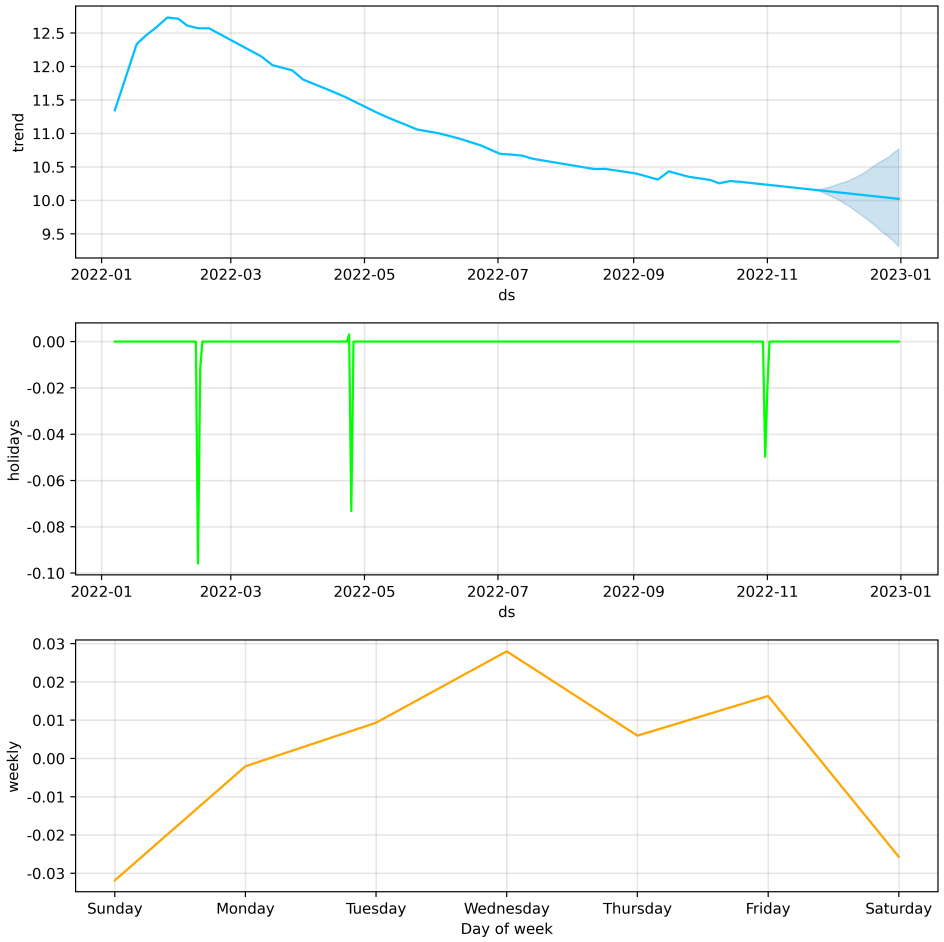

Figure 6: Time Series Decomposition Plot

Seasonal and Holiday Effects:

Holidays cause a decrease in the number of reports, such as a slight dip in the number of reports around Valentine's Day as seen in the linear trend chart. In the weekly effect, the number of reports increases from Sunday to Wednesday and decreases from Wednesday to Saturday (with a rebound on Friday). This suggests that people tend to play Wordle as a pastime on workdays, and have less interest on holidays.

Explanation of Overall Variation:

The SIRS infectious disease model can explain changes in the trend component well. Our assumptions are as follows:

Assumption 1: All Twitter users $A(t)$ can be divided into three groups:

(1) Ordinary Twitter users $S(t)$. They may be influenced by seeing some Wordle player's score reports on Twitter and may be motivated to become Wordle players. They correspond to "susceptible individuals";

(2) Wordle players $I(t)$. Some players will post reports on Twitter, which will attract others to become Wordle players. They correspond to "infected individuals";

(3) Tired players $R(t)$. They will not play Wordle for a period of time, but may start playing again after this time. They correspond to "recovered individuals".

Assumption 2: Ordinary players S may have a probability of λ of being infected; in players I, they have a probability of β of getting tired of playing Wordle and not playing for a period of time; in tired players R, there is a probability of η of being influenced by external factors and starting to play Wordle again. Ordinary players S, players I, and tired players R may all have a probability of natural removal of a certain θ.

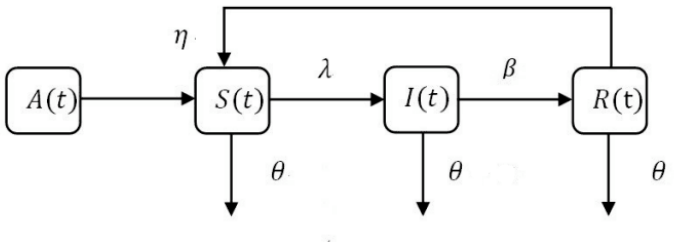

Figure 7: Player State Transitions

Based on the above assumptions, after setting the parameters, the number of players can be fitted by solving the differential equations, and then multiplied by a certain proportion to calculate the number of score reports on Twitter. The corresponding fitting curve of the report quantity is shown in Figure 8, which conforms to the trend curve of Prophet. Therefore, the SIRS model can be used to explain the overall trend of the change. Wordle became popular from January 2022, and the number of players reached its peak around February (the number of reports also reached its peak). After that, the game gradually cooled down, the number of players decreased, and the number of reports also decreased.

3.4 Extracting the Attributes of Words

To investigate the impact of the attributes of words on the proportion of reports of challenging patterns, we first need to extract various useful properties of words.

1. **the Number of Different Letters in a Word(NDLW)**

   In general, the fewer different letters a word has, the lower the probability of guessing a letter in the trial, and the more difficult the puzzle becomes. We count the distribution of words with different numbers of letters and the average proportion of people who made 5+ tries, and the results are shown in Table 5. As can be seen from the table, the fewer

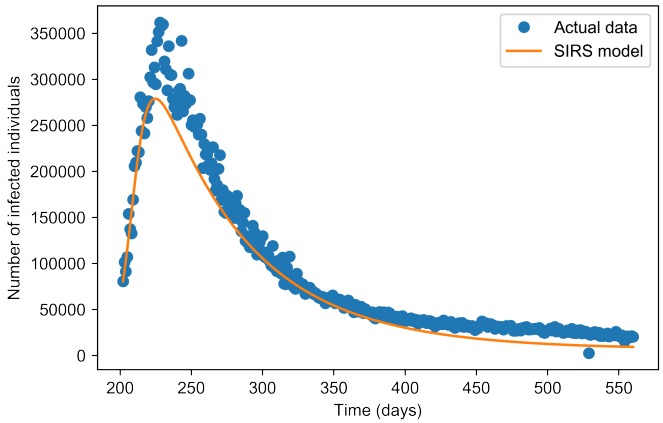

Figure 8: Fitted Curve of the SIRS

different letters a word has, the higher the proportion of people who made 5+ tries, indicating that the puzzle is more difficult. Therefore, the number of different letters in a word is an important attribute of the word.

Table 5: Varieties of Letters in Words and Proportion of 5+ Tries

Different Letters	Number of Words	Proportion of 5+ Tries
3	6	62.50%
4	94	45.10%
5	259	34.90%

2. the Frequency of Word Usage in Daily Life(Freq)

In general, the more frequently a word is used in daily life, the more familiar people are with it, and vice versa. The more unfamiliar a word in a puzzle is, the more difficult the puzzle becomes. Therefore, the frequency of word usage in daily life is also an essential attribute. We use the word frequency data from Wolfram[4], which is calculated from the Google Books dataset.

3. the Breadth of Word Usage in Different Fields(BU)

The more widespread the usage of a word, the more people are familiar with it, and vice versa. The less familiar people are with the words in a puzzle, the more difficult the puzzle becomes. The prevalence of a word is defined as the number of corpora in which the word appears among 100 corpora (data from "Word Frequencies in Written and Spoken English").

4. the Sum of Letter Usage Frequency(SLF)

When playing the Wordle game, players usually try words that contain more common letters to gain more information. Therefore, whether the letters in a word are common or not is also an important attribute to measure the difficulty of a word. We define the SLF to describe this attribute of a word:

$$SLF = \sum_{i=1}^{5} frequency_i \tag{6}$$

where $frequency_i$ represents the frequency of the i^{th} letter in the word. The letter-frequency data is obtained from the website Algoritmy[1].

5. **the Sum of a Letter in a Word**

   The sum of a letter in a word is also an attribute of the word, as the puzzles consist of five letters that can be the same or different.

6. **Part-of-Speech of a Word**

   The part-of-speech of a word is one of the most common attributes of a word.

3.5 Impact of Word Attributes on the Proportion of Hard-Mode Reports

The proportion of reports in the hard mode is defined as follows:

$$percentage_{hard} = \frac{number_{hard}}{number_{reported}} \tag{7}$$

We establish a multiple linear regression model based on the least squares method and use the significance test of the regression equation (i.e., F-test) to study whether word attributes have an impact on the proportion of Hard-Mode reports.

3.5.1 Model Establishment

Multiple linear regression describes the relationship of the dependent variable y with independent variables $x_1, x_2, \ldots, x_m$ by the following equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{8}$$

where β_0 is the constant term, β_k is the regression coefficient of the kth independent variable, and ε is the random error term.We perform a multiple linear regression with $Freq$, SLF, $NDLW$,BU, the Sum of a Letter in a Word, Part-of-Speech of a Word as independent variables, and $percentage_{hard}$ as the dependent variable. Due to the length of the obtained regression equation, it is included in **Appendix A**.

3.5.2 Significance Test of Regression Equation

1. **Hypothesis Formulation:**

   Null hypothesis: $H_0 : \beta_0 = \beta_1 = \cdots \beta_m = 0$;

   Alternative hypothesis: $H_1 : \beta_0, \beta_1, \cdots, \beta_m$ are not all equal to 0.

2. **Calculate F-statistic:**

$$F = \frac{SSR/m}{SSE/(n-m-1)} \sim F(m, n-m-1) \tag{9}$$

   where $SSR = \sum_{i=1}^{n} (\hat{y}i - \bar{y})^2$ represents the regression sum of squares, and $SSE = \sum i = 1^n (y_i - \hat{y}_i)^2$ represents the residual sum of squares.

3. **Making Decisions** The rejection region of the test is $F > F_\alpha(m, n-m-1)$, based on the given significance level of $\alpha = 0.05$. We establish a multiple linear regression model with word attributes as independent variables and the proportion of reports in hard mode as the dependent variable.

   The F-statistic of the regression equation is 1.058 with a corresponding P-value of 0.379 (> $\alpha = 0.05$), indicating that the regression equation does not exhibit statistical significance. Therefore, we conclude that word attributes **do not** have a significant impact on the proportion of reports in hard mode.

4 Model 2-Distribution Prediction Model based on BP Neural Network

4.1 Model Building of BP

We first preprocess the data by combining pretraining with Global Vectors model (GloVe) and dimensionality reduction with PCA. Word embedding is a technique that maps words to real-valued vectors and is a fundamental application in natural language processing. GloVe model is one of the word embedding models, which adopts squared loss and fits the word vectors to the global statistical information calculated based on the entire dataset. We use pre-trained word vectors from the GloVe model as features.

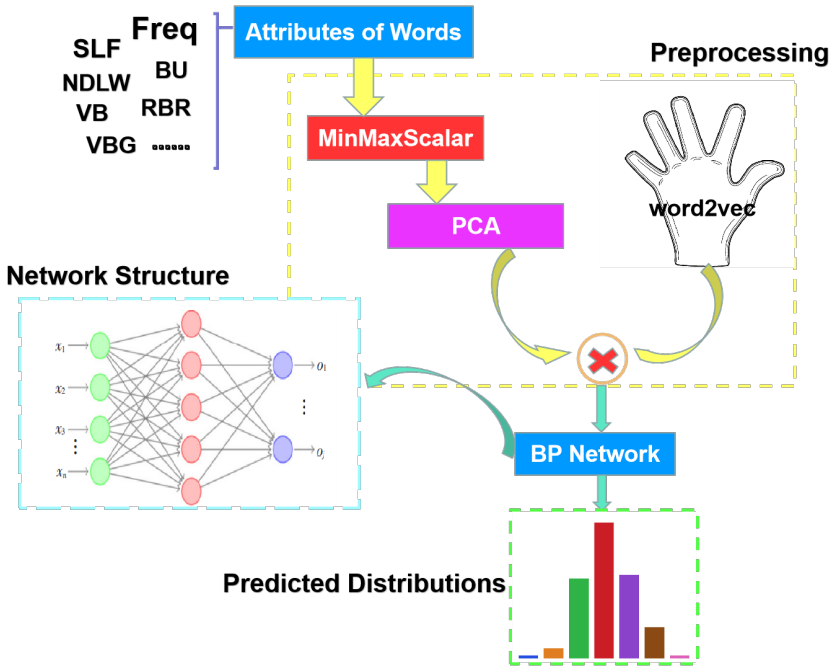

Figure 9: Implementation Process

In addressing task 2, we use the word properties obtained from the first sub-question as part of the features, normalize them, and extract the principal components using PCA. We combine the extracted principal components with the word vectors pre-trained using the GloVe model and use them as the input features for the BP neural network. In the BP neural network, we input the word features and use the percentage distribution of each word as the label. We select 80% of the data as the training set and 20% as the test set to train the neural network and test its

performance. Since the amount of data given is small, we choose to establish a low-complexity network, which includes an input layer, a single hidden layer, and an output layer. The hidden layer contains 1024 hidden units, and the ReLU function is used as the activation function. Dropout is applied during training to drop 50% of the network units to counter overfitting. The L2-norm is selected as the loss function, the Adam optimizer is used for gradient optimization during backpropagation, and the learning rate is set to 0.05. Xavier random initialization is used.

4.2 Model Uncertainty of BP

Neural networks have considerable randomness, and the initialization parameters in the Xavier method of the neural network are sampled from a uniform distribution. Additionally, dropout randomly drops neurons in the hidden layer. This means that the training results of the neural network may vary each time. To address this issue, we try to train the model multiple times and select the best model.

The model output may be negative, and to address this issue, we choose to adjust the negative values to 0.

Since the output values cannot be directly used as percentages, as their sum may exceed or be less than 100, we divide each output by the total sum to obtain the final predicted percentages.

4.3 Model Evaluation of BP

The evaluation results on the test set are as follows:

Table 6: Evaluation on Test Set

MAE	MSE	MAPE
3.2302	21.857	32.9194

On the test set, the mean absolute error (MAE) of the neural network is around 4, indicating that the average absolute difference between predicted values and true values is 4, which indicates a high accuracy of the model. Other metrics also support this conclusion.

4.4 Model Prediction of BP

We predict the distribution of people trying different times and the result is as follows. We are confident that the error is within 3%.

5 Model 3-Difficulty Classification based on K-Means++

5.1 Clustering Analysis based on K-Means++.

The K-Means algorithm is an unsupervised learning method and a clustering algorithm based on partitioning. It usually uses Euclidean distance as the metric to measure the similarity between data objects, and the similarity is inversely proportional to the distance between data objects. The larger the similarity, the smaller the distance. The algorithm requires a predetermined initial number of clusters k and k initial cluster centers. Based on the similarity between the data object and the cluster center, the algorithm continuously updates the position of the cluster center and reduces the sum of squared errors (SSE) of the clusters. When SSE no

Try Times	Percentage(%)
1	0
2	6
3	18
4	29
5	27
6	15
7+	5

Table 7: Predicted Result of ”EERIE”

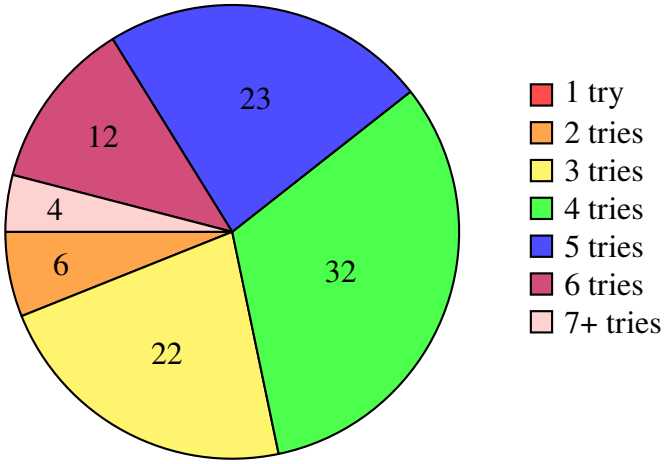

Figure 10: Predicted Result of ”EERIE”

longer changes or the objective function converges, the clustering ends, and the final result is obtained.

The categories ”1 try”, ”2 tries”, up to ”7+ tries” well reflect the difficulty of the puzzles. We use these variables as inputs and employ the K-Means++ algorithm to classify the difficulty of the words. The specific process is as follows:

Step 1: Determine the number of clusters k and initialize k cluster centers.

Step 2: Calculate the Euclidean distance between the data points and the k initial cluster centers, and cluster partition based on the minimum distance, resulting in k regions.

Step 3: Calculate the center position of each cluster obtained in the previous step and use it as the next iteration's cluster center.

Step 4: Repeat the above steps until the change between the last two clustering results meets the accuracy requirements.

The elbow rule chart in Figure 11 is used along with our experience in differentiating game difficulty to determine the number of clusters. We choose the number of clusters to be k=3, representing three levels of difficulty: hard, medium, and easy.

Finally, we obtain the clustering results, which show that cluster 1 contains 135 words, cluster 2 contains 156 words, and cluster 3 contains 68 words. The statistical results of the mean and standard deviation of each property in the clusters are shown in Table 8. By calculating the average proportion of tries with 5+ tries in each cluster, we obtain Table 9. By observing Table 8 and Table 9, we categorize the words in Cluster 1, Cluster 2, and Cluster 3 as easy, medium, and hard, respectively.

5.2 Relationship between Word Attributes and Difficulty Levels

5.2.1 Relationship Between Difficulty Levels and NDLW

The distribution of word difficulty levels for words with different NDLW is shown in Figure 12, and the proportion of NDLW of different difficulty levels is shown in Table 10. We found that the proportion of words with fewer different letters increases with increasing difficulty level. In the dataset provided for this study, there are six words that have only three different letters.

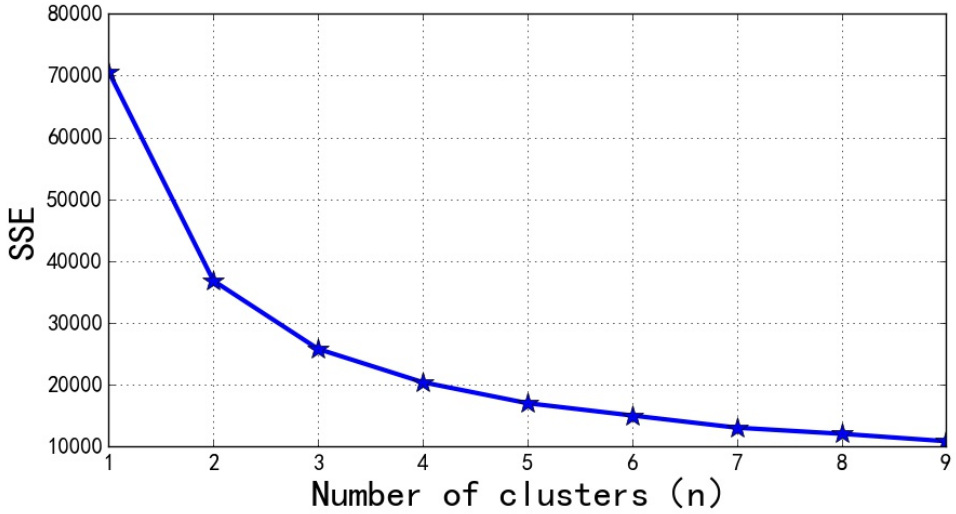

Figure 11: K-value Optimization

Table 8: Results and Significance Tests of K-means Clusters for Different Categories

	Cluster Categories (means ± sd)				
	1 (n=135)	2 (n=156)	3 (n=68)	F-value	P-value
1 try	0.8 ± 1.057	0.269 ± 0.459	0.279 ± 0.452	21.307	**0.000*****
2 tries	9.459 ± 4.168	4.013 ± 1.749	2.868 ± 1.962	168.146	**0.000*****
3 tries	30.748 ± 3.81	20.212 ± 3.516	12.574 ± 4.108	594.862	**0.000*****
4 tries	33.637 ± 3.824	35.487 ± 3.819	25.647 ± 4.485	150.08	**0.000*****
5 tries	17.807 ± 3.159	26.436 ± 3.115	28.794 ± 5.732	268.933	**0.000*****
6 tries	6.43 ± 2.261	11.596 ± 3.05	21.662 ± 4.188	565.646	**0.000*****
7+ tries (X)	1.081 ± 0.931	1.974 ± 1.169	8.132 ± 7.035	119.123	**0.000*****

Table 9: Average Proportion of 5+ tries for Different Categories

Cluster categories	Average Proportion of 5+ tries
Cluster 1	25.35%
Cluster 2	39.92%
Cluster 3	58.59%

Table 10: Proportion of NDLW Across Difficulty Levels

Cluster Categories	Proportion of NDLW		
	5	4	3
Easy	90.78%	9.22%	0%
Medium	63.13%	35.63%	1.24%
Hard	52.94%	41.18%	5.88%

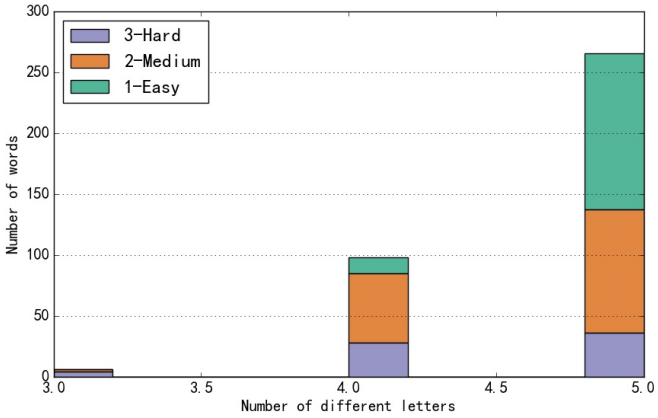

Figure 12: Relationship between Difficulty Levels and NDLW

Two of them are classified as medium difficulty, namely ”motto” and ”madam”, while the other four are classified as difficult, namely ”fluff”, ”mummy”, ”cacao”, and ”vivid”. Based on the above analysis, we have sufficient evidence to suggest that **the fewer different letters a word contains, the more difficult it is to guess.**

5.2.2 Relationship between Difficulty Levels and SLF

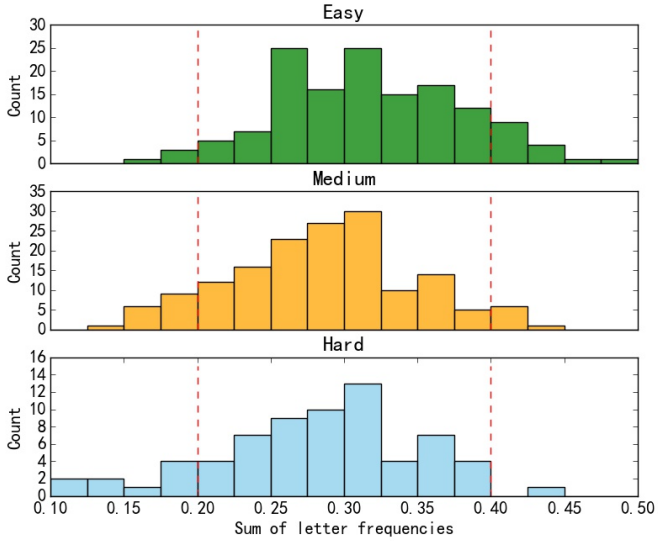

Figure 13: Relationship between Difficulty Levels and NDLW

The distribution of the SLF for words of different difficulty levels is shown in Figure 13 and Table 11. We focus mainly on the part where the SLF is less than 0.2 and greater than 0.4 because it represents the relationship between the difficulty level and the majority of letter frequency sum. It can be observed that as the difficulty level increases, the proportion of whose SLF less than 0.2 will increase, while the proportion of whose SLF greater than 0.4 will decrease. This means that **the more commonly used letters in a word, the easier it is to be guessed, and vice versa.**

5.2.3 Relationship between Difficulty Levels and BU and Freq

Table 12 shows the distribution of word breadth for different levels of difficulty. The breadth of a word is defined as the number of corpora in which the word appears out of a total of 100

Table 11: Distribution of the SLF Across Difficulty Levels

Cluster Categories	Proportion of SLF		
	< 0.2	0.2-0.4	> 0.4
Easy	2.84%	86.52%	10.64%
Medium	10%	85.63%	4.37%
Hard	13.24%	85.29%	1.47%

corpora, and it takes integer values between 0 and 100. The table indicates that **the more widely a word is used in different fields, the easier the corresponding puzzle, and vice versa.**

At the same time, we attempt to find a relationship between the difficulty level of a word and its frequency of use in everyday life. Although there is a difference in the mean frequency of different levels of difficulty, the mean value is sensitive to outliers. Therefore, we first sort the words by frequency and then use a histogram to show their distribution, as shown in Figure 14. Ultimately, we find that **there is no significant relationship between the difficulty level and the frequency of use in everyday life, as the distribution of word frequencies for different levels of difficulty is fairly uniform**.

Table 12: Distribution of BU Across Difficulty Levels

Cluster Categories	BU		
	0-33.33	33.33-66.66	66.66-100
Easy	9.93%	12.77%	77.30%
Medium	19.38%	21.87%	58.75%
Hard	22.06%	26.47%	51.47%

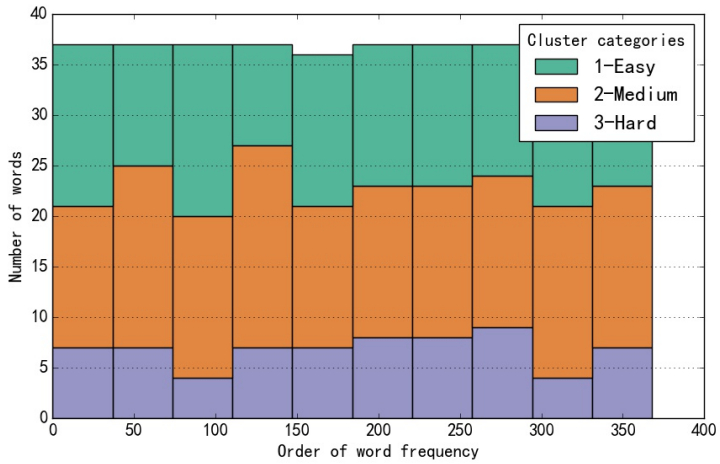

Figure 14: Relationship between Difficulty Levels and Freq

5.3 PCA Discussion on the Accuracy of Model Classification

The discussion of model classification accuracy can be divided into two parts. The first part is whether there is a significant difficulty difference among the words in each cluster, and the second part is the effectiveness of K-Means clustering. As mentioned earlier, in the first part,

the average proportion of tries with 5+ tries for each cluster's words is 25.35%, 39.92%, and 58.59%, respectively, which shows a significant difficulty difference among the clusters.

Regarding the second part, we conduct principal component analysis (PCA) on seven features, including "1 try", "2 tries" to "7+ tries". We find that the variance explained by the first two principal components reaches 82.88% ($> 80\%$). Therefore, we take the first two principal components to create a scatter plot, as shown in Figure 15. From this scatter plot, it is evident that the words are well differentiated, and the K-Means clustering effect is good.

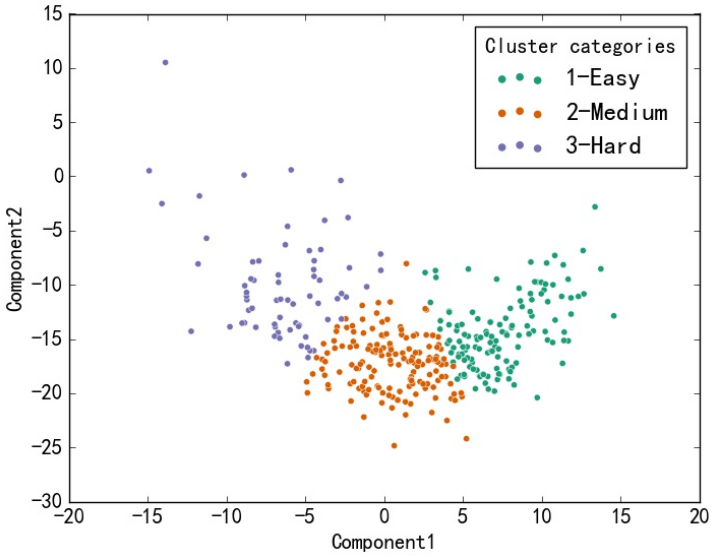

Figure 15: PCA on the Accuracy of Model Classification

From the perspective of evaluation metrics, we compared K-Means++ with other clustering methods like the Partitioning Around Medoids(PAM) and the Gaussian Mixture Model(GMM) and the results are shown in Table 13. From the comparison results, it can be seen that K-Means++ has a better clustering effect.

Table 13: Model Evaluation

Model	Silhouette Coefficient	CH Score
K-Means++	0.372	309.326
PAM	0.347	303.049
GMM	0.347	291.637

5.4 Determining the Difficulty Level of "EERIE"

According to the results of the second question, we obtain the distribution of EERIE's relevant percentages (1, 2, 3, 4, 5, 6, X), which is [0, 6, 18, 29, 27, 15, 5]. By inputting this distribution into our model, we determine that the difficulty level of "EERIE" is "medium".

6 Interesting Surprise

6.1 Are These Words Really that Difficult?

After categorizing words by their difficulty levels, we find that the difficulty level of the word "mummy" is "difficult". When we calculate the proportion of words that have been attempted "5 times or more" and sort them in descending order, we are surprised to find that "mummy" is ranked second in difficulty, with 82% of people attempting it five or more times. As we have been familiar with the word "mummy" since childhood, we subjectively thought it was an easy word, making the opposite result all the more surprising. Additionally, other similar words include "watch", "catch", "prize", etc. In fact, this confirms our view that "there is no significant relationship between word frequency and difficulty level."

6.2 Which Initial Letter Poses the Greatest Challenge for Solution Words?

We use hypothesis testing to identify which initial letter makes for the most difficult word puzzles. Firstly, we tally the frequency of each initial letter among all the words. Next, we define the difficulty coefficient of a word as the sum of the "5 tries", "6 tries", and "7 or more tries (X)" categories, and calculate the difficulty coefficient of all the words. After sorting the difficulty coefficients and selecting the top 20% most difficult words, we tally the initial letters of these words. We then calculate the probability of each letter being the initial letter of a word that enters the top 20% most difficult list. Some of the results are shown in Table 14.

Assuming that the probability of a word starting with a certain letter entering the top 20% most difficult list is 0.2, that is

$$H_0 : \quad p = 0.2 \tag{10}$$

Let n be the total number of times the letter appears as the first letter and k be the number of times the letter appears as the first letter in a word that enters the top 20% most difficult list. Then we have

$$k \sim B(n, 0.2) \tag{11}$$

Since both n and k are known, we can calculate the probability P of this situation occurring. Taking the significance level α as 0.05, when $P < \alpha = 0.05$, we reject the null hypothesis and believe that p is not equal to 0.2. We have calculated the P values for all the letters and sorted them in ascending order. Some of the results are shown in Table 15.

According to Table 15, for the letters e, s, f, w, and a, their corresponding P values are less

Table 14: the Relationship of First Letters and Difficulty

the First Letter	Total	Hard(20%)	Easy(20%)	Hard Rate	Easy Rate
a	28	2	5	0.071	0.179
b	20	4	3	0.200	0.150
c	33	7	10	0.212	0.303
d	12	4	5	0.333	0.418
e	10	6	0	0.600	0.000
f	22	8	2	0.364	0.091
g	17	5	1	0.294	0.059
h	11	4	2	0.364	0.182
...	...	...	...	...	...

Table 15: Statistics of Initial Letters of Difficult Words

the First Letter	Total	Hard(20%)	Hard rate	P value
e	10	6	0.6	0.006
s	51	4	0.078	0.011
f	22	8	0.364	0.036
w	11	5	0.455	0.039
a	28	2	0.071	0.046
r	13	0	0	0.055
t	30	3	0.1	0.079
...	...	...	...	...

than 0.05, so $p \neq 0.2$ is significant. The Hard rates of words with initial letters e, f, and w are greater than 0.2. Therefore, we have sufficient reason to believe that words with initial letters e, f, and w are difficult, with words starting with e being the most difficult. Likewise, we can find that the letter t corresponds to the easiest words when used as the first letter. Specific data can be found in Table 16.

Table 16: Statistics of Initial Letters of Easy Words

The first letter	Total	Easy(20%)	Easy rate	P value
t	30	16	0.533	0
d	12	5	0.417	0.053
c	33	10	0.303	0.056
s	51	13	0.255	0.081
...	...	...	...	...

6.3 What Words Can Make Wordle Continue to be Popular?

An upward mutation point indicates a sudden increase in a trend that was supposed to decline. This may mean that the word corresponding to the upward mutation point is more likely to stimulate communication and spread among people, which led to an increase in the number of reports on that day. To detect upward mutation points, we first used the Isolation Forest algorithm to detect all mutation points, and then used the positivity or negativity of the first-order difference to search for upward mutation points. At the same time, we calculated the mutation rate and selected mutation points with a mutation rate greater than 5% as the final points. The results are shown in the table below.

Date	Word	Date	Word	Date	Word
2022/2/8	frame	2022/3/18	saute	2022/4/18	flair
2022/2/15	aroma	2022/3/21	their	2022/4/22	plant
2022/2/17	shake	2022/3/24	chest	2022/4/26	heist
2022/2/19	swill	2022/3/27	nymph	2022/4/29	tarsh
2022/2/22	thorn	2022/3/30	stove	2022/5/2	story
2022/3/2	nasty	2022/4/1	snout	2022/5/4	train
2022/3/5	brine	2022/4/2	trope	2022/5/9	shine
2022/3/11	watch	2022/4/8	scare	2022/5/11	farce
2022/3/15	tease	2022/4/13	chunk	2022/12/26	judge

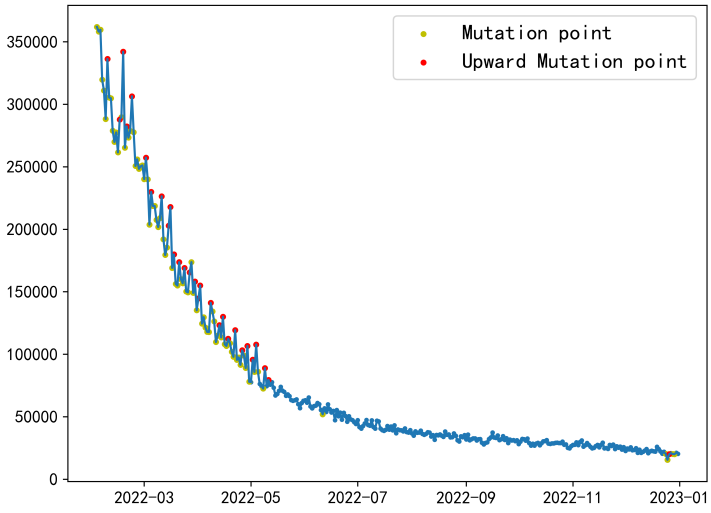

Figure 16: Sensitivety Analysis

The New York Times may be able to learn from the characteristics of these upsurge words to enhance the fun of Wordle.

7 Sensitivety Analysis

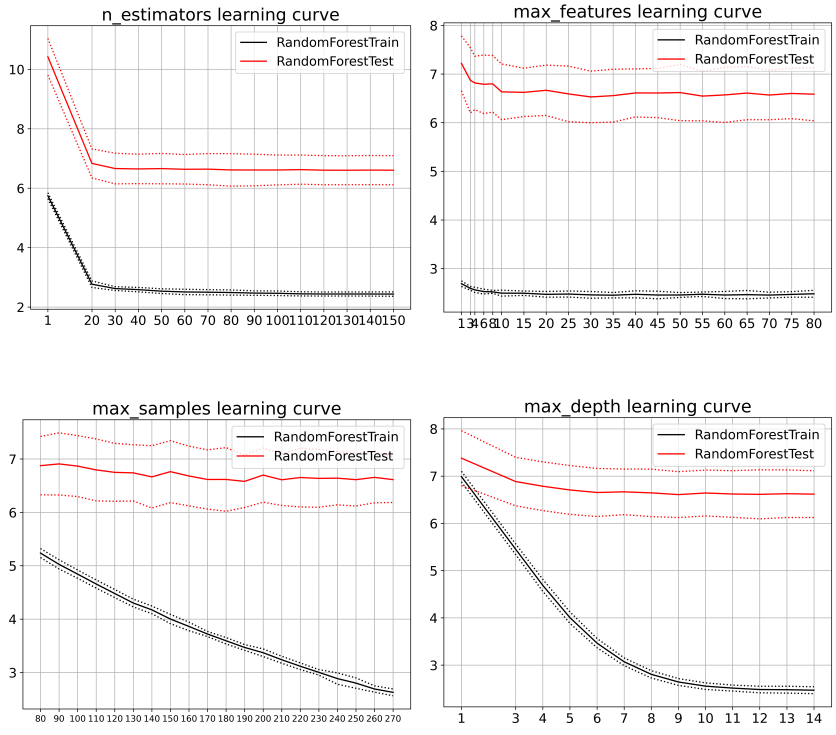

Figure 17: Sensitivety Analysis

In improving the BP neural network model, we used a random forest model to perform regression on the results of three attempts. We selected parameters $n_{estimators}$(the number of weak evaluators), $max_{features}$(the maximum number of sampling features),$max_{samples}$(the maximum sample size for random sampling), and max_{depth}(the maximum depth of the tree) for sensitivity analysis. In each parameter analysis, we only varied that parameter while keeping the other pa-

rameters at their default values. We performed 5-fold cross-validation on the dataset, and took the average of the *RMSE* on the training and test sets as the result for each parameter value. As shown in Figure 17, the *RMSE* on the training and test sets gradually decreases and stabilizes as the parameter values increase. The *RMSE* under various parameter changes is stable at around 6.60. This indicates that the model is not sensitive to parameter changes and is relatively stable.

8 Model Assessment

8.1 Strengths

1. The Prophet model takes into account the influence of holidays on time series and has stronger interpretability of the model parameters, which helps us understand the changes in the number of reports.

2. The SIRS model has excellent explanatory power for the trend term of the Prophet model.

3. The BP neural network has the ability to output multiple targets and does not require separate models for each target.

4. The K-Means++ model has the characteristics of simplicity and practicality and is suitable for the dataset of this question.

8.2 Weaknesses

1. The Prophet model is somewhat inadequate for long-term forecasting, similar to traditional time series models to some extent.

2. The parameters of the BP neural network have randomness during initialization, and the trained model contains this randomness, which means that in some cases, the output results of the model are not always a unique value.

3. Although we are working hard to find word attributes related to the difficulty of the puzzle, there may still be some word attributes that we have overlooked.

References

[1] Algoritmy.net. Letter frequency english, Accessed on 2023-02-18.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[3] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, pages 0–0, 2017.

[4] Wolfram.com. Wordfrequencydata, Accessed on 2023-02-18.

Dear Sir or Madam,

We are honored to present our analysis on the number of reports and word difficulty of Wordle after data analysis and modeling. We are confident that our interesting findings will be beneficial to you. The following are some of our theoretical analyses and numerical predictions.

1. The number of reports will decrease slightly, but will rise again. The process of playing games is like an infection, people will always tire of it after liking, and like it again after being tired of. **The number of players is just like an infection curve,** which increases over time, reaches a peak, and then gradually decreases. **Furthermore, the number of reports tends to decrease during holidays and weekends, resulting in small oscillations in the curve.** According to our model, the number of reports will drop to the range of (10,452, 21,454) on March 1, 2023(95% confidence level). Nonetheless, it will increase later.

2. No word attribute affects the proportion of scores reported that were played in Hard Mode. The multiple linear regression model obtained by fitting shows an R-squared value of only 0.129. Correspondingly, the p-value is 0.379, which is much greater than 0.05. This indicates that there is almost no correlation between word attributes and the proportion of Hard-Mode reports. This is reasonable because no one knows what the word is before playing, and therefore, a word's attributes do not affect whether one plays in difficult mode or not.

3. We can predict the distribution of the answer tries more accurately based on the attributes of words. By training a BP neural network model, we can grasp the answer rate of words based on their properties. For example, for the word "EERIE", the distribution of the number of people'tries should be as follows: 0% passed in one try, 6% passed in two tries, 22% passed in three tries, 32% passed in four tries, 23% passed in five tries, 12% passed in six tries, and 4% failed to pass. We are confident that the margin of error is within 3%.

4. The attributes that contribute to the difficulty of guessing words may be different from what you imagine. Through K-means++ clustering, we divided word difficulty into three levels based on the proportion of successful tries and analyzed the relationship between difficulty and word attributes. Combining the neural network model mentioned earlier, we can directly judge the difficulty of a word based on its attributes. People may think that the difficulty of guessing a word is related to its frequency of use, but in fact, this is incorrect. **In Wordle, the difficulty of guessing a word is related to the variety of letters in the word, the sum of the frequencies of use of each letter, and the breadth of its usage in different fields.** Based on our analysis, the word "EERIE" should be classified as having medium difficulty.K-means++ model performed better than other similar models, thereby increasing the credibility of our results.

Next, we will introduce **some interesting findings**: We bet you never thought "**mummy**" is the second most difficult word to guess! Up to 82% of players need to try five or more times to solve it. Other similar words include "watch", "catch", "prize", and so on. This is because the difficulty of guessing a word is related to what letter it starts with, for example, **words starting with "e", "s", "f", "w", and "a" are more difficult to guess, while starting with "t" easier to guess.** You can try to use these interesting characteristics to design games.

These are all suggestions and strategies our team has provided to your company. Thank you for precious time. Hope that our models and these conclusion can be helpful to you!

Sincerely,

MCM Team Members

Appendices

Appendix A Regression Equation

Here is the regression equation refered in 3.5.1

$$\begin{aligned}\hat{y} = & \underset{(0.000)}{0.045} - \underset{(0.365)}{8.696}\,Freq + \underset{(0.000)}{0.007}\,SLF - \underset{(0.250)}{0.003}\,NDLW + \underset{(0.003)}{0.008}\,a \\ & + \underset{(0.133)}{0.007}\,b + \underset{(0.045)}{0.006}\,c + \underset{(0.013)}{0.009}\,d + \underset{(0.000)}{0.010}\,e + \underset{(0.011)}{0.010}\,f + \underset{(0.014)}{0.008}\,g \\ & + \underset{(0.040)}{0.007}\,h + \underset{(0.003)}{0.010}\,i + \underset{(0.018)}{0.027}\,j + \underset{(0.430)}{0.003}\,k + \underset{(0.011)}{0.006}\,l + \underset{(0.037)}{0.007}\,m \\ & + \underset{(0.069)}{0.006}\,n + \underset{(0.048)}{0.005}\,o + \underset{(0.052)}{0.007}\,p + \underset{(0.515)}{0.007}\,q + \underset{(0.123)}{0.004}\,r + \underset{(0.144)}{0.004}\,s \\ & + \underset{(0.004)}{0.008}\,t + \underset{(0.011)}{0.009}\,u + \underset{(0.047)}{0.010}\,v + \underset{(0.309)}{0.005}\,w + \underset{(0.227)}{0.011}\,x + \underset{(0.003)}{0.010}\,y \\ & + \underset{(0.032)}{0.022}\,z + \underset{(0.320)}{0.016}\,VBG - \underset{(0.062)}{0.020}\,VB - \underset{(0.477)}{0.016}\,CC + \underset{(0.079)}{0.008}\,JJ \\ & + \underset{(0.156)}{0.019}\,VBZ + \underset{(0.118)}{0.019}\,VBN + \underset{(0.052)}{0.016}\,VBD - \underset{(0.525)}{0.010}\,MD \\ & + \underset{(0.066)}{0.008}\,NN + \underset{(0.348)}{0.010}\,NNS + \underset{(0.042)}{0.014}\,RBR + \underset{(0.102)}{0.012}\,VBP \\ & - \underset{(0.700)}{0.008}\,JJS - \underset{(0.742)}{0.007}\,JJR + \underset{(0.918)}{0.003}\,PRP - \underset{(0.377)}{0.019}\,DT \\ & R^2 = 0.129, N = 359\end{aligned} \tag{12}$$