Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2307432

Cracking the Code: Twitter Data Insight in the Wordle Game

Abstract

Wordle is a daily word puzzle game which gives players six chances to guess a randomly selected five-letter word. In this paper, we develop three models to predict the total reports number and corresponding scores distribution by utilizing the daily reports generated on Twitter.

For **Problem 1**, we establish the **Improved Prophet Model Based On GBRT** to explain the daily variation of total reports number and perform prediction. We first replace the outliers and abnormal words of the dataset, and perform normalization on the distribution data. Then, we summarize and quantify the word difficulty indicators in our feature selection process. We list out the possible attributes of words and test their correlation with difficulty. We use 5 quantitative indicators to represent the valid attributes, which are **word frequency**, **letter frequency**, **bigram frequency**, **duplicate letters** and **the similarity factor**. After that, we propose the **Two-group Players Model** based on the assumption that the number of common players decreases at a fixed rate and the number of hardcore players remains the same. Deducing that the total reports number fits the general **Negative Exponential Model**, we fit the general trend and obtain the subtraction sequence. As it is a time series influenced by human factors, we choose the **Prophet Forecasting Model** to capture the seasonal effects and the **Gradient Boosting Regression Tree model** to model the residual. The prediction interval of our model is **17974 to 26628** on March 1, 2023. Finally, we determine that the attributes of the word **hardly affect** the percentage by performing the *Spearman* correlation test and calculating the correlation coefficient based on the five indicators selected.

For **Problem 2**, we notice that the distribution is compositional data under the unit-sum constraint, we develop the **Distribution Prediction Model Based on Dirichlet Regression** to handle the flexible correlation structure among the components. We set the time factor and the 5 difficulty indicators as variables to perform Dirichlet regression and our prediction distribution for the word "EERIE" is $(0.67\%, 8.49\%, 23.08\%, 28.86\%, 20.70\%, 12.31\%, 5.90\%)$. Uncertainties of our model may come from unexpected social factors on a specific date, the inputs data quality, the training set ratio and the prediction error due to the unit-sum constraint. For model validation on our testing set, the **chi-square test** shows that 69 in 72 words are in the 95% confidence interval and 50 words are in the 99% confidence interval, denoting the high confidence of our model.

For **Problem 3**, we propose the **Difficulty Classification Model Based on K-Means** and use the **Multivariate Logistic Regression Model** to explain the corresponding word's attributes by analyzing the coefficients. We take the distribution data of each word as the input of the K-Means Clustering Model and classify the words into three groups based on the elbow rule. The number of words in "easy", "normal" and "hard" groups are 134, 152, and 73. The word EERIE is classified into the "HARD" group. To test the accuracy of our model, we use **Rank Sum Ratio (RSR)** method and **Systematic Clustering** and get 305 and 298 the same classification result, respectively. By adding a certain ratio of disturbance into our input distribution data, we analyze the sensitivity of our classification model and get a high *F1-score* over 0.875 with a disturbance under 6%.

For **Problem 4**, we consider the trend of total numbers, the percentage of people choosing the hard mode and the difficulty level of the words and find out some interesting features of this data set such as holiday effects and high-reel effects. Finally, we summarize the models and prediction results established in the former sections and write to the Puzzle Editor of the New York Times for development and extension.

Keywords: **Dirichlet Regression** **Prophet** **XGBoost** **K-Means**

Contents

1 Introduction

1.1 Background

Wordle is a web-based word game created and developed by Welsh software engineer Josh Wardle and owned and published by the New York Times Company since 2022. Every day, a five-letter word is chosen from an allowed list, which players aim to guess within six tries. After every guess, each letter is marked as either green, yellow, or gray: green indicates that the letter is correct and in the correct position, yellow means it is in the answer but not in the right position, while gray indicates it is not in the answer at all. Players can also choose to play in a regular mode or "Hard Mode" via setting. Wordle's Hard Mode requires players to include letters marked as green and yellow in subsequent guesses. Near the end of 2021, Wordle was released and quickly rose in popularity while the numbers calculate that Wordle has dropped in popularity by a staggering 51 percent in a little over five months. Different versions of the game are now available in over 60 languages.

1.2 Restatement of the Problem

Attachment 1 is a file of daily Wordle results obtained by mining Twitter from January 7, 2022, to December 31, 2022. The file includes the date, contest number, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage that guessed the word in one try, two tries, three tries, four tries, five tries, six tries, or could not solve the puzzle (indicated by X).

We need to develop a model to determine and solve the following problems:

- **Problem 1:** Develop a model to explain the variation of the number of reported results and predict the number on March 1, 2023 with a prediction interval. Secondly, determine whether the attributes of the word affect the percentage of scores reported that were played in Hard Mode and give explanations.
- **Problem 2:** Given a future date and the corresponding word, develop a model to predict the associated percentages of players reporting 7 different tries. Then, analyze the uncertainties of the model and prediction. Finally, apply the model on the word EERIE on March 1, 2023 and give distribution. Analyze the confidence level of the model's prediction.
- **Problem 3:** Develop and summarize a classification model based on difficulty. After that, identify the attributes of a given word that are associated with each classification. Apply the model on the word "EERIE" and discuss the accuracy of the classification model.
- **Problem 4:** Find out other interesting facts of the data set.

Finally, we need to summarize our results and write a letter to the Puzzle Editor of the New York Times.

1.3 Analyze of the Problem

First of all, we need to do the data preprocessing: replace the outliers and abnormal words. After that, we need to perform the feature selection: we need to fully consider all possible attributes of words which are associated with the game, and use quantitative indicators to present them.

For **Problem 1**, note that the daily reports can be considered as time series with human factors. Also, the falling trend of the total number indicates the decreasing rule of the social event. Then, we need to study whether the correlation between the attributes listed above and the Hard Mode percentage exists.

For **Problem 2**, we mainly need to establish a word-based model to solve the distribution of the percentage. After indicator selection, we need to select an appropriate prediction model to deal with the distribution of data. As we are dealing with compositional data under the unit-sum constraint, the usual multivariate statistical methods are not available. We need to develop a corresponding prediction model to predict the distribution of the percentages. The uncertainties of the model and prediction can be analyzed through the process of data input, variable selection, parameter estimation, and prediction. Also, the confidence of our model can be measured by the confidence level in chi-test, which can be obtained by performing a goodness-of-fit test. Model assessment index and also be utilized for model validation.

For **Problem 3**, Since we don't know exactly how difficult the words are, the problem is actually an unsupervised classification problem. And we also need to figure out the correlations between attributes of the word and corresponding categories to give explanations.

1.4 Literature Review

The prediction problem that this paper aims at is based on compositional data under the unit-sum constraint. Aitchison[1] , and Beardah et al. [6] investigate different methods to handle it. Connor and Mosimann [7] originally proposed the Dirichlet distribution and more recently Gueorguieva, Rosenheck, and Zelterman [10] have applied the Dirichlet component regression to the assessment of schizophrenia symptoms.

Moreover, there are also many studies about the **Characteristics of Hard Wordle Words**. The reasons include the obscurity of the word, uncommon spelling patterns[2] word usage, letter frequencies,the word structure and so on[3, 4]. Besides the reason for hard words, people also investigate the features of hard mode and list some tricks to play[5].

1.5 Our Work

In this paper, we develop three models to predict the total reports number and corresponding scores distribution utilizing the daily reports generated on Twitter. Our work mainly includes the following 4 parts and the mind map is shown in Fig 1.

The first model is the **Improved Prophet Model Based On GBRT**, which explains the daily variation of total reports number and We give the prediction interval of **17974 to 26628** on March 1, 2023. Our prediction model contains three main parts: A negative exponential curve to fit the general decreasing trend after June, a **Prophet Forecasting model** to capture the seasonal effects and the **Gradient Boosting Regression Tree** model the residual of the time series.

Then, we refer to the research online to summarize the possible attributes of words and use our data to test the correlation between attributes and the difficulty level. We use 5 quantitative indicators to represent the valid attributes, which are word frequency score f_{word}, letter frequency score f_{letter}, bigram frequency score f_{bigram}, duplicate letters $repetition$, and the similarity factor $adjacency$. Based on the five indicators selected, we perform the *Spearman* correlation test and calculate the correlation coefficient to prove that the attributes of the word hardly affect the percentage.

The second model we develop is the **Distribution Prediction Model Based on Dirichlet Regression**. Noticing that the distribution is compositional data under the unit-sum constraint, we need a model to handle the flexible correlation structure among the components. We select the Dirichlet Regression model and set the time factor and the 5 difficulty indicators as variables, and our prediction distribution for the word "EERIE" is $(0.67\%, 8.49\%, 23.08\%, 28.86\%, 20.70\%, 12.31\%, 5.90\%)$. For model validation, we use the **chi-square test**, the model assessment indexes MAE and MSE, and indicators BIC and AIC, all denoting the high confidence of our model.

The third model we propose is the **Difficulty Classification Model Based On K-Means**. We take the distribution data of each word as input and classify the words into three groups "easy", " normal" and "hard", and the word EERIE is classified into the "HARD" group. To test the accuracy of our model, we use the **Rank Sum Ratio (RSR)** method and **Systematic Clustering**, and over 70% of the classification results are the same. By adding a certain ratio of disturbance to our input distribution data, we analyze the sensitivity of our classification model and get a high *F1-score* over 0.875 with a disturbance under 6%.

We also find out some interesting features of this data set and finally, we summarize the results in the letter to the Puzzle Editor of the New York Times.

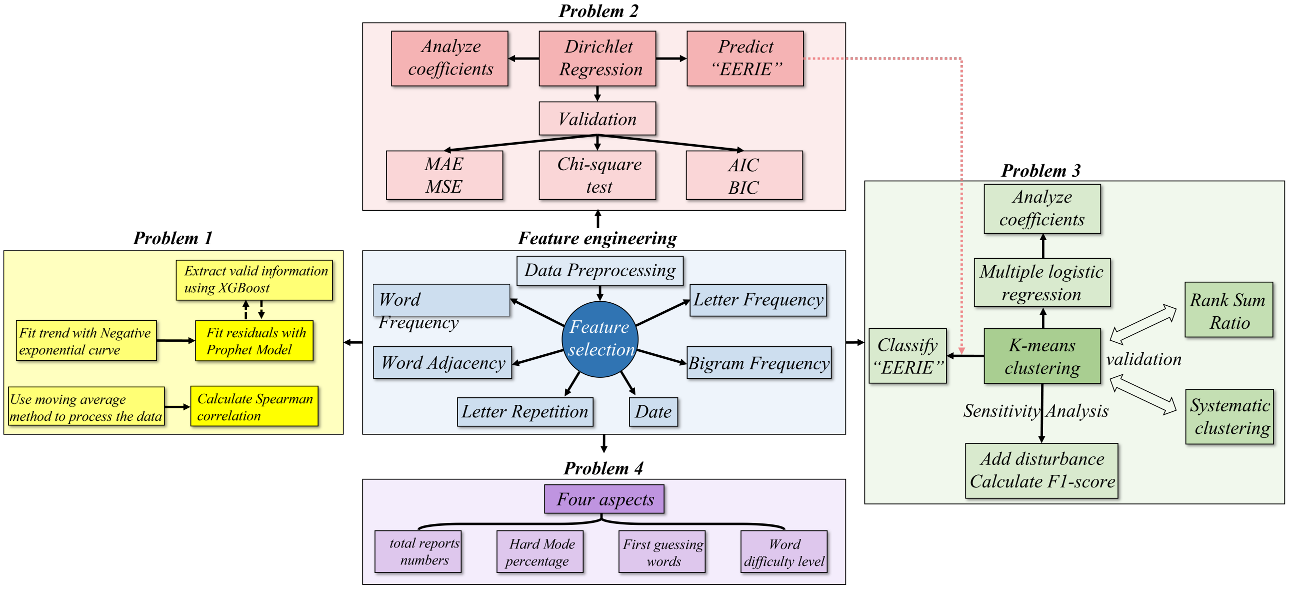

Figure 1: Overview of our work

2 Assumptions and Justifications

- **Assumption 1:** The word is randomly selected from the word list every day.
- **Assumption 2:** The number of common players $C(t)$ decreased at a fixed speed λ.

Table 1: Notations used in this literature

Symbol	Description
$N(t)$	Players number at t time
$C(t)$	Common players number at t time
h	Hard core players number
λ	Loss ratio of common players number per unit time
$\text{freq}(x)$	Frequency of x
f_{word}	Frequency score of the word
f_{letter}	Frequency score of letters in the word
f_{bigram}	Frequency score of bigrams in the word
$repetition$	Number of duplicate letters in the word
$adjacency$	Number of words with the edit distance one
$x_1, \cdots, x_6$	Percentage of reports for each number of tries
k	Ratio for normalization

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

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

4 Data Preprocessing

Since the reports number are generated daily, we consider it a time series problem and sort the data according to time by reversing the sequence for further investigation. Then, we calculate the percentage of players choosing Hard Mode because we only focus on the relative proportion for Hard Mode. The data preprocessing process is mainly made up of **outliers detection** and **normalization**.

4.1 Outliers Detection and Replacement

Notice that the number of reported results of the 529th contest is 2569, which is significantly different from the number of its closer dates at over 24000 reports in total. So we replace the outlier using the results of three spline interpolation. After that, we also found that there are some abnormal words due to errors in mining Twitter. So we calculate the edit distance between each abnormal word and all the words in the solution list, and we find the optimal replacement word taking word frequency into consideration. For example, we replace "clen" and "rprobe" with "clean" and "probe" respectively.

4.2 Normalization

Compositional data are proportions or percentages of disjoint categories adding to one. we apply an normalization method by dividing the data by the sum of all seven percentages, and the equation for the normalization is as follows:

$$C(x_1, x_2, \cdots, x_6, x_7) = (\frac{k \cdot x_i}{\sum_{i=1}^{7} x_i}, \frac{k \cdot x_2}{\sum_{i=1}^{7} x_i}, \cdots, \frac{k \cdot x_7}{\sum_{i=1}^{7} x_i}) \tag{1}$$

where $x_1, \cdots, x_6$ stands for the percentage of reports for each number of tries, and x_7 stands for the percentage of reports that could not solve the puzzle. k is the ratio for normalization under the unit-sum constraint. We successfully dealt with the outlier "nymph" with the total percentage of 120%.

After preprocessing, we draw line graphs of the total reports number as follows:

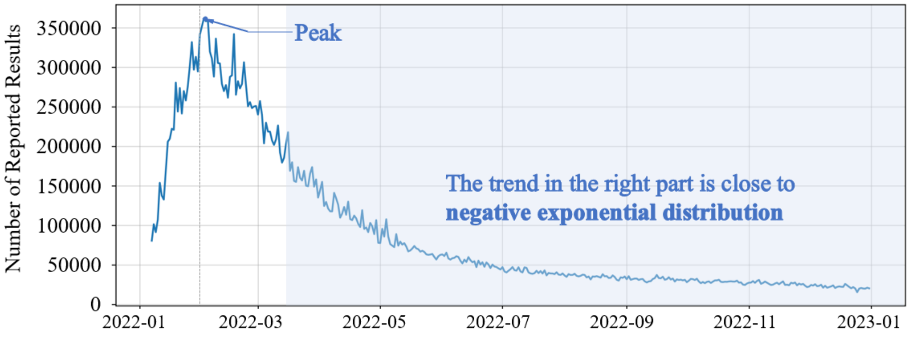

Figure 2: Line graph of total reports number on Twitter

From Fig 2, we can see that there is a dramatic rise in the total reports number in the first month, reaching a peak at 361908 reports on February 2, 2022. Then, the number generally continued to decline until the last day of the year. We also notice that the falling trend become milder over time, and finally it fluctuates around 21000 reports.

5 Improved Prophet Model Based on GBRT and Negative Exponential Model

Analysis: We notice that in Fig 2, a certain proportion of players were lost every day because of the decline of Wordle's popularity. Suppose that the number of all the players are $N(t)$. While dedicated fans of Wordle will still remain, other common players may lose interest. We divide the players into two groups. One group is called the **hardcore players** who will keep playing Wordle every day. The other group is called **common players**. The number of the two groups of players are $C(t)$ and $H(t)$ respectively, and we assume that the number of common players decrease daily at a certain speed $\lambda(t)$ per day. Through assumptions and derivation, we found that the variation of $C(t)$ fit the Negative Exponential Model. So we use the general **Negative Exponential Model** to fit the curve to model the general trend of our total reports. Consider that the daily total reports number is a time series affected by human factors, we select **Prophet Forecasting Model** for prediction. Since we found that the residual of Prophet model doesn't satisfy the normal distribution, we use the ensemble model **Gradient Boosting Regression Tree** to model the residual term. The overall prediction flow chart is shown in Fig 3:

5.1 Negative Exponential Model

Model Construction

We have the equation (2) to indicate the variation of two types of players.To simplify the model, we set $H(t) = h$, which means hard core players love the wordle game so much that they play it every

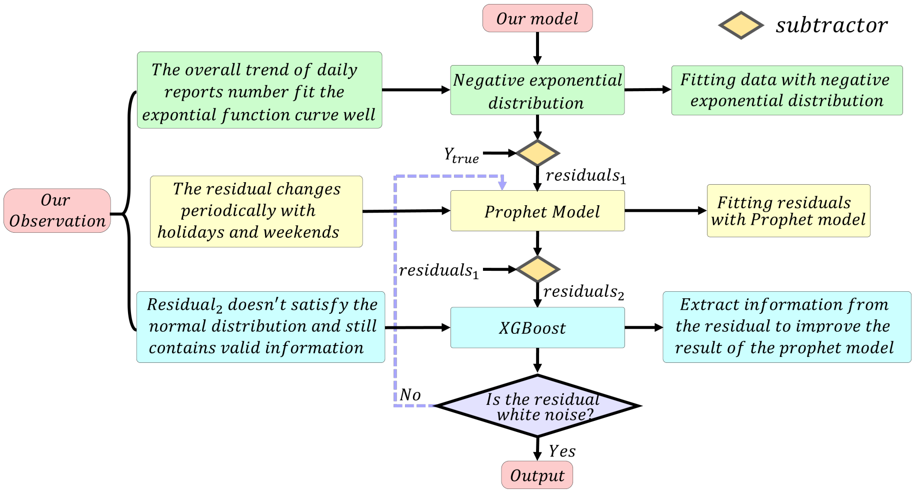
Our model
subtractor
The overall trend of daily reports number fit the expontial function curve well
Negative exponential distribution
Fitting data with negative exponential distribution
Y_{true}
$residuals_1$
Our Observation
The residual changes periodically with holidays and weekends
Prophet Model
Fitting residuals with Prophet model
$residuals_1$
$residuals_2$
Residual$_2$ doesn't satisfy the normal distribution and still contains valid information
XGBoost
Extract information from the residual to improve the result of the prophet model
No
Is the residual white noise?
Yes
Output

Figure 3: Flowchart of our time-series prediction model

day and never get bored. We also set $\lambda(t) = \lambda$, which fix the decreasing rate of common players.

$$\begin{cases} N(t) = C(t) + h, \quad t \geq 0 \\ \dfrac{\mathrm{d}C(t)}{\mathrm{d}t} = -\lambda C(t) \\ C(0) = C_0 \end{cases} \tag{2}$$

From equation (2), we notice that $C(t)$ can be presented as follows:

$$C(t) = C_0 e^{-\lambda} \quad t \geq 0 \tag{3}$$

which satisfies the equation for the Negative Exponential Model. Therefore, the total number $N(t)$ satisfy the Negative Exponential Model, and we use the **Negative Exponential Model** to fit the total number $N(t)$:

$$N(t) = C_0 \mathrm{e}^{-\lambda t} + h \tag{4}$$

Model Solution

First, we use only the data after June to fit the negative exponential curve (equation (4)) to model the general declining trend of the total number. We set the x coordinate value to 1 on June 1, 2022, and we got the parameter estimation of the equation (4) :

$$C_0 = 3.83058 \times 10^4, \lambda = 1.2856 \times 10^{-2}, h = 2.1021 \times 10^4$$

and we get the fitted function below:

$$N(t) = 3.83058 \times 10^4 e^{-1.2856 \times 10^{-2} t} + 2.1021 \times 10^4 \tag{5}$$

which indicates that there are about 2.1021×10^4 hardcore players and about 3.83058×10^4 common players on June 1, 2022.

We then use the model to predict the number of reported results on March 1, 2023 and got the number of 22288 .

5.2 Prophet Forecasting Model

To model the daily variation accurately, we use the original total reports number minus the data fitted by the Negative Exponential Model as a new time series, indicating the effects of the potential fluctuation factors. Considering that the publish of Wordle game is a social event and the popularity of the game may be associated with human factors such as holidays and weekends, so potential variation factors may be weekly or monthly effects and holiday effects. So we choose Prophet to model the fluctuations in daily variation.

5.2.1 Model Construction

Prophet is a decomposable time series model for forecasting based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. The nonlinear trend denotes the regularity of year, week and day and the holiday effect. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. The model mainly consists of the following parts:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{6}$$

where $g(t)$ is a trend function that represents the value of non-periodic changes, $s(t)$ represents the weekly, seasonal and annual cyclical changes, and $h(t)$ represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term ϵ_t represents any idiosyncratic changes which are not accommodated by the model, and we make the parametric assumption that it is normally distributed.

After applying the negative exponential model on the data to get the general daily variation trend, we set the significance level to 0.05 and use the Prophet to perform forecasting, obtaining an interval with the confidence level of 95%. We perform the Pearson chi-square test on the residual term of Prophet model and found that it doesn't obey the normal distribution. Therefore, we try to model the residual term to obtain more information of the time series, which will be demonstrated in the next subsection 5.3.

5.3 Improve the Model

To model the residual term of Prophet, we select the Gradient Boosted Regression Trees (GBRT), which is a classification regression tree model in XGBoost to improve the prediction accuracy of our Prophet model.

5.3.1 Gradient Boosted Regression Trees (GBRT)

Gradient Boosted Regression Trees (GBRT) is a flexible non-parametric statistical learning technique for classification and regression. For the gradient boosting regression tree, the prediction result of each sample can be expressed as the weighted sum of the results on all trees:

$$\hat{y}_i^{(k)} = \sum_{k}^{K} \gamma_k h_k \left(x_i\right) \tag{7}$$

where K is the total number of trees, k represents the $k - th$ tree, γ_k is the weight of the tree, and $h_k$represents the predetermined result on the tree.

GBRT uses a particular model ensembling technique called gradient boosting. If we let $g_t(x) = \sum_{i=0}^{t-1} f_i(x)$ be the classifier trained at iteration t , and $(y_i, g(x_i))$ be the empirical loss function, at each iteration we will move g_t towards the negative gradient direction $-\left.\frac{\partial L}{\partial g}\right|_{g=g_t}$ by η amount. Hence, f_t is chosen to be

$$f_t = \arg\min_{f} \sum_{i=1}^{N} \left[\left.\frac{\partial L\left(y_i, g\left(x_i\right)\right)}{\partial g\left(x_i\right)}\right|_{g=g_t} - f\left(x_i\right) \right]^2 \tag{8}$$

and the algorithm sets $g_{t+1} = g_t + \eta f_t$. For regression problems with squared loss function, and $\frac{\partial L(y_i, g(x_i))}{\partial g(x_i)}$ is simply $y_i - g(x_i)$. The algorithm simply fit a new decision tree to the residual at each iteration.

5.3.2 Prediction Interval

In this way, we can perform the forecast based on the general trend predicted by Negative Exponential model and the fluctuation term by the improved Prophet Forecasting Model. The mind map for prediction is shown below in Fig 4:

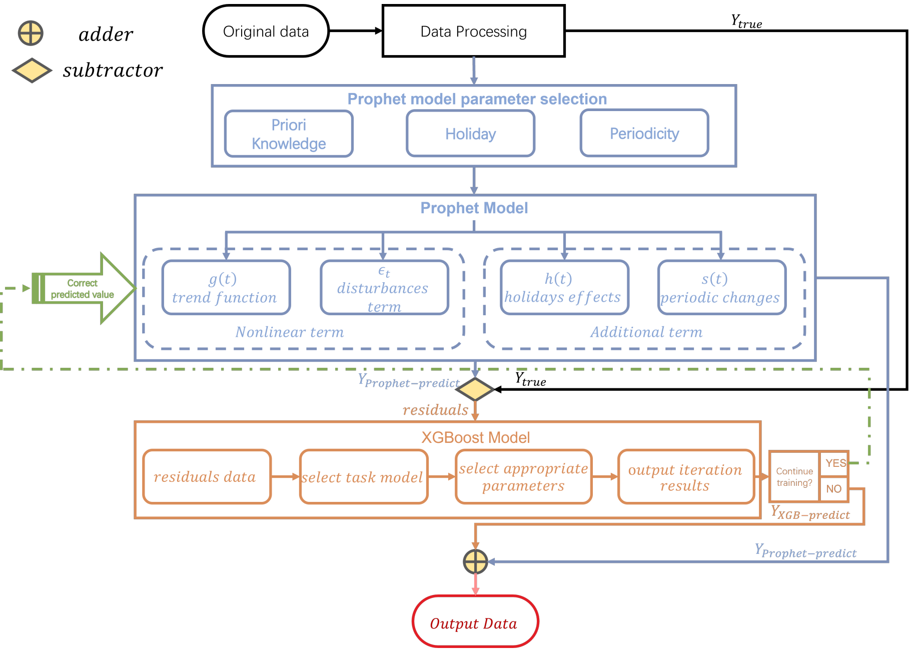

Figure 4: Mindmap of reports number prediction model

Finally, we add up the two terms and got the prediction interval. The prediction interval for the number of reported results is drawn in Fig 5 as follows:

We predict that the lower boundary of reports number is 17974 and the upper boundary is 26628.

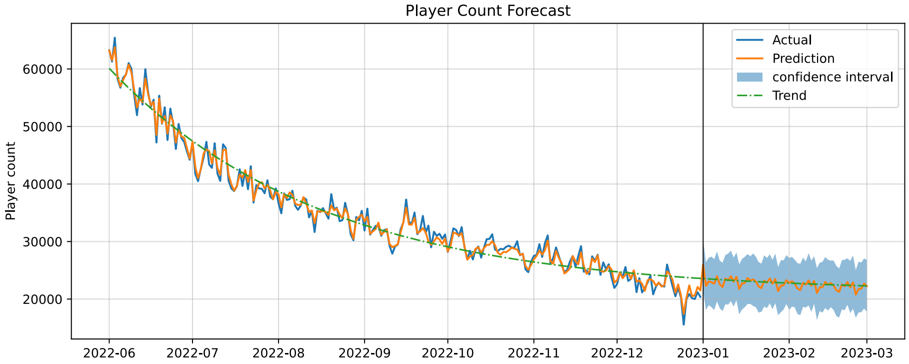

Figure 5: Prediction for the number of reported results

5.4 Model Validation

In order to test the accuracy and reliability of the model, we collect the total reports number from Twitter as actual data, and we use the single Prophet model to predict the value directly to make comparison. We plot the actual data, prediction of our model and the single Prophet model in Fig 6 to validate our prediction model.

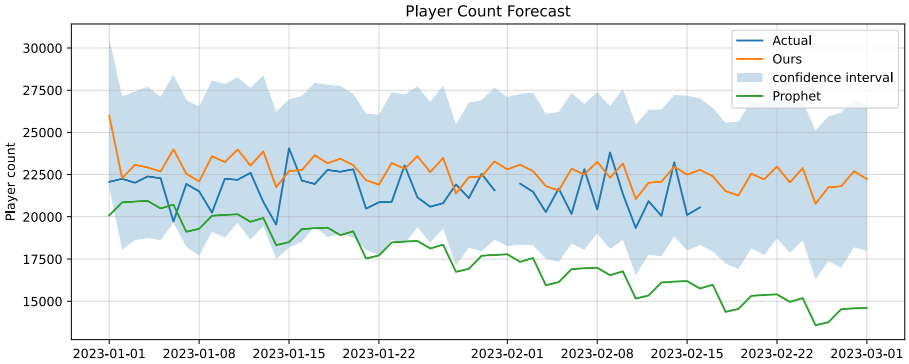

Figure 6: Prediction contrast between Prophet and Our Model

The green line represents the prediction from the single Prophet model while the blue line and orange line represents the actual data and our prediction respectively. While the deviation between the orange line and the green line increases with time to about 8000, our prediction model has a much better performance with a deviation around 1000 along the time.

6 Feature Selection

6.1 Difficulty Indicator Selection

The prediction problem and classification problem that this paper aims at is mainly based on how difficult the word is to guess. We list the possible attributes of words that may be associated with difficulty. For each of the attribute, we search for corresponding quantitative factors and indicators to represent them. The main 5 difficulty indicators are shown in Fig 7.

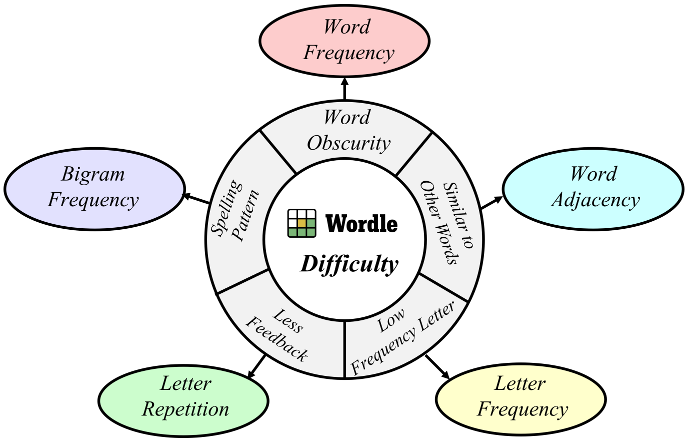

Figure 7: Wordle Difficulty indicators

1. Word obscurity level:

   It is obvious that common words are more easy to come up with compared to obscure words in the Wordle game. The best indicator for word obscurity level is the word frequency, which can include commonness and readability. To get the frequency of each word, We use the python package "wordfreq" [14, 16]. Then we define the score of the word frequency as:

   $$f_{word}(w) = \log(\text{freq}(w)) \tag{9}$$

   where w denotes the word.

   As for why we should apply logarithmic transformation to the actual frequency in Equation (9), (10) and (11), this can be found in Section 6.2.

2. Low frequency letter:

   People often try different popular letters to maximize the probability of guessing the right letters. First, we use the weighted average method to calculate the frequency of each letter in the 26 letters. Based on the Wordle's allowed word, list[15] we can add up the frequency of its five letters and the word-based variable for letter frequency score can be calculated as follows:

   $$f_{letter}(w) = \sum_{i=1}^{5} \log(\text{freq}(x_i)) \tag{10}$$

   where $x_1, x_2, \cdots, x_5$ denote the 5 letters the word w contained.

3. The spelling pattern of the word:

   A word with uncommon spelling pattern may be difficult to guess due to uncommon syllables, which can be presented by bigrams and trigrams. As trigrams are rare in a 5-letter word and they may contain some common bigrams as well, we set another variable f_{bigram} to denote the score of the bigrams[11, 12, 15].

   $$f_{bigram}(w) = \sum_{i=1}^{4} \log(\text{freq}(x_i)) \tag{11}$$

   where $x_1, x_2, \cdots, x_4$ denote the 4 bigrams the word w contained.

4. Less feedback causing by duplicate letters

   According to research, the word with duplicate letters are harder to guess because of the feedback confused the player. We set the variable $repetition$ to represent whether the word contains duplicate letters.

5. Adjacency to other words

   If players have more options when filling the blanks, the words may be harder to guess. We use the variable $adjacency$ to denote the number of words with the edit distance 1, which can be the word's indicator of similarity to other words.

There are other possible factors such as the number of syllables, the number of vowels, the number of consonants...However, we found that they have little correlation with the distribution by performing the *Spearman* correlation test

Therefore, we have 5 indicators of the word difficult level, which are word frequency f_{word}, letter frequency f_{letter}, bigram frequency factor f_{bigram}, whether it contains duplicate letters $repetition$, and the similarity factor $adjacency$.

6.2 Variables Transformation

1. Logarithmic transformation:

   We notice that when modeling variables with non-linear relationships, the chances of producing errors may also be skewed negatively. And the absolute value of frequency is not necessary the best indicator for describing how common the word is. For example, the frequency of the word "argue" and "world" is 3×10^{-5} and 5×10^{-4} respectively. While they have difference of ten times, they are both common words that people can easily come up with when playing Wordle. **Logarithmic transformation** [17] is a convenient means of transforming a highly skewed variable into a more normalized dataset. So we perform the **logarithmic transformation** on three frequency score variables f_{word}, f_{letter}, f_{bigram}.

2. Normalization:

   To eliminate the impact of dimensional gap, we use the **Min-Max scaler function** to perform Normalization on the 5 variables.

7 Hard Mode Correlation Analysis

Because Hard Mode requires players to include letters marked as green and yellow in subsequent guesses, and players actually don't know the words in advance, so we suppose that the attributes of the word itself is the major factor in the correlation. And we select the 5 key indicators in chapter 6 for correlation analysis.

7.1 Eliminate the Trend Effect

After calculating the percentage data in the preprocessing process, we try to eliminate the effect of the overall trend so that we can exclude other unrelated factors. We process the percentage data using move average function and obtain the sequence after subtraction.

7.2 Correlation Detection

We perform the *Spearman* correlation test between the 5 indicators and the percentage fluctuation. We found only the p-value of "$adjacency$" is over 0.05, However, the correlation coefficient between the Hard Mode percentage and the indicator $adjacency$ is only -0.0792, indicating the low correlation.

8 Distribution Prediction Model Based on Dirchlet Regression

Analysis: The Dirichlet regression is a powerful method for modeling **compositional data**, which can explain the different trends and covariance structures and it does not require the independence of the data. It is also flexible in terms of the choice of the link function and the prior distribution, which allows for modeling different types of relationships and incorporating prior knowledge about the data. So we think it is an appropriate prediction model to deal with distribution data in our prediction problem. Despite the attributes of the word itself, we consider that the distribution may be associated with time because there are more tricks and hints available and core players tend to remain playing the game. So we include the time factor, and use the same 5 difficulty indicators in chapter 6.1 as key 6 variables to perform Dirichlet regression.

8.1 Model Construction and Results Analysis

Dirichlet Regression Models

The Dirichlet model with constant parameters is a model that can accommodate certain shapes of compositional data while the Dirichlet regression model is more flexible. In Dirichlet density function, data should be incorporated in proportion-form only. This means that, for $c > 0$, given two vectors, x and $x' = cx$, both will be treated as $y = C(x) = C(x')$ where C is the closure operation. Therefore, the Dirichlet models have scale-invariance property.

Let $\mathbf{x} = (x_1, \ldots, x_D)$ be a $1 \times D$ positive vector having Dirichlet distribution with positive parameters $(\lambda_1, \ldots, \lambda_D)$with density function

$$f(\mathbf{x}) = \left(\Gamma(\lambda) / \prod_{i=1}^{D} \Gamma(\lambda_i)\right) \prod_{i=1}^{D} x_i^{\lambda_i - 1} \tag{12}$$

where $\sum_{i=1}^{D} x_i = 1$ and $\lambda = \sum_{i=1}^{D} \lambda_i$.

A Dirichlet regression model is readily obtained by allowing the parameters of a Dirichlet distribution to change with a co-variate. For a given covariates, the parameters of a Dirichlet distribution $\mathcal{D}(\lambda_1, \ldots, \lambda_D)$ can be written as positive-valued functions $g_j(s)$ of the covariant s. Besides the exponential family, the family of polynomials is a suitable candidate in this situation where it exhibits the desired positivity on a certain range. A different Dirichlet distribution is modelled for every value of the co-variate, resulting in a conditional Dirichlet distribution for $\mathbf{x} \mid s$ which is $\mathcal{D}(g_1(s), \ldots, g_D(s))$.

For details on the estimation of Dirichlet parameters and the selection of starting values see Ronning [11], Narayanan[12, 13].

Model Construction and Prediction

We split our dataset into 8 : 2 for training and testing respectively, which is 72 words for testing and 287 words for training. Then, we use the model to predict the word ERRIE. The percentage of each group

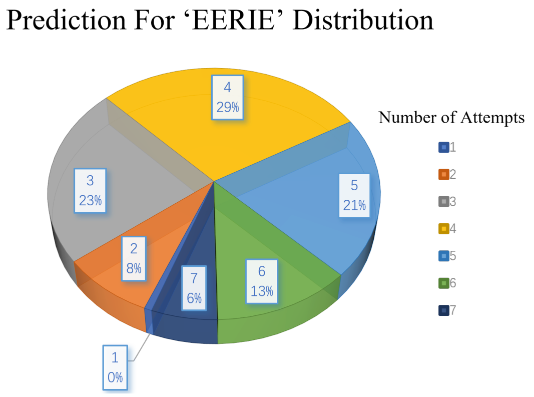

Figure 8: Distribution Pie Chart for the word EERIE

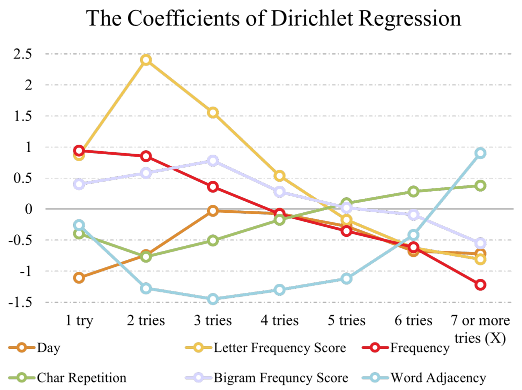

Figure 9: Coefficients of Dirichlet Regression

$(1, 2, 3, 4, 5, 6, X)$ are $(0.6717\%, 8.48636\%, 23.07775\%, 28.859\%, 20.703\%, 12.30789\%, 5.899406\%)$, with the expectation guesses 4.2. The Distribution Pie Chart is shown in Fig 8:

Coefficients Analysis of Dirichlet Regression

After applying the model on the percentage of seven tries group respectively, we got the coefficients of Dirchlet regression for the 6 variables. The line chart is shown in Fig9.

As we can see, the line which represents the coefficient for the letter frequency, bigram frequency and word frequency shows a declining trend, while the coefficients of similarity factor and duplicate letters rise. Note that the coefficients for time factor, represented by the orange line Day, are all negative for all seven percentage with no apparent rule. This indicates low correlation between days and the distribution. The graph reveals that there is a negative trend between three frequency and the attempts the words need, and words with duplicate letters or similar with other words are more difficult to guess.

8.2 Uncertainties of The model and Prediction

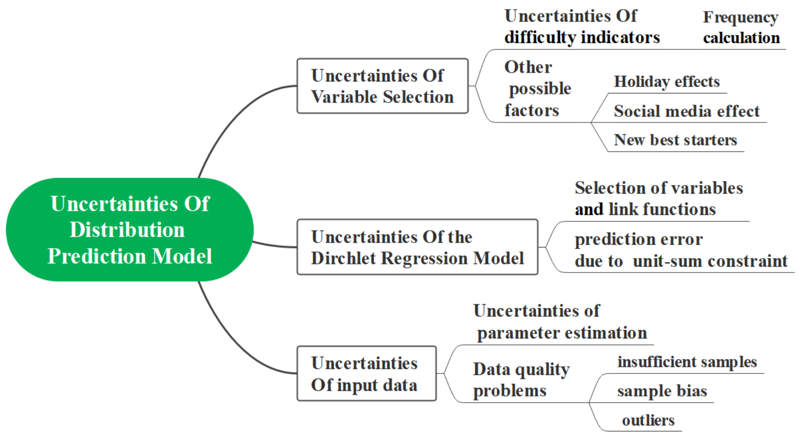
Uncertainties Of Distribution Prediction Model
Uncertainties Of Variable Selection
Uncertainties Of difficulty indicators
Frequency calculation
Other possible factors
Holiday effects
Social media effect
New best starters
Uncertainties Of the Dirchlet Regression Model
Selection of variables and link functions
prediction error due to unit-sum constraint
Uncertainties Of input data
Uncertainties of parameter estimation
Data quality problems
insufficient samples
sample bias
outliers

Figure 10: Uncertainties Of Distribution Prediction Model

Our Distribution Prediction Model and prediction results may have the following uncertainties :

1. Uncertainty of variable selection :

   In our prediction model, we select 5 attributes of words and the time factor as the indicators of the Dirichlet regression model. However, the word frequency and letter frequency are based on the allowed list of Wordle, which have certain randomness. Also, there may be other possible factors such as holiday effects and People's psychological factors in the distribution prediction. For example, "high-reel effect" and newly reported "best opening words" on social media. These factors are hard to quantify so are not included in our Dirchlet variables.

2. Uncertainty of the Dirchlet Regression Model:

   The selection of variables and appropriate link functions in Dirchlet Regression will significantly affect the accuracy of the model. In our prediction model, the rising percentage of Hard Mode players and the competence of remain players can hardly be fully represented by the time factor, so the selected variables have certain uncertainties. What's more, since the prediction of Dirichlet regression model is a probability distribution, the result have certain volatility under the unit-sum constraint.

3. Uncertainty of inputs data:

   The parameter estimation of Dirichlet regression model is usually carried out by maximum likelihood estimation or Bayesian inference. However, due to factors such as model complexity and noise in inputs data, there are certain errors in the process of parameter estimation and the accuracy of the prediction will also be affected. Moreover, data quality problems such as insufficient samples, sample bias and outliers, will also affect the results. These problems may lead to uncertainty and error of prediction results.

8.3 Model Validation

We perform three different validation method on the distribution of each word in the testing set, which contains 72 words.

1. Validation Based On Confidence Level

   We perform the **chi-square test** on two groups of data and get the corresponding p-value. Through calculation, we found that 69 words get a p-value higher than 0.95 and 50 words get a p-value higher than 0.99, which proves the high confidence level of our model.

2. Validation Based On MSE and MAE

   We calculate the MSE and MAE of our prediction on the training set and get 35.705 and 4.17 respectively, showing high accuracy and reliability of our model.

3. Validation Based On BIC and AIC

   We calculate the BIC and AIC of our prediction model and get -10672 and -12025 respectively. Since the smaller the value of the BIC and AIC, the more accurate the model is. Therefore, our model perform well on the prediction task.

9 Difficulty Classification Model Based on K-Means

Analysis: To solve the unsupervised classification problem, we need to select difficulty indicators as inputs to perform clustering on words for difficulty evaluation. According to the coefficients analysis of Dirchlet Regression model in section 8.3, we found that the distribution have little correlation with the time factor, which indicates that the distribution can fully describe a words' difficulty level. Therefore, we can utilize the word-based percentage distribution obtained in the former sector directly as the best difficulty indicators for clustering inputs. We select the K-Means Clustering model to do the unsupervised classification based on the compositional data.

To find the words' attributes for each classification groups for explanations, we need to investigate the correlation between 5 indicators and our classification results. As we have 3 categories as the targets for our data, we decide to use the **Multivariate logistic regression model** and use the coefficients for analysis. For model's accuracy discussion, we use two more methods, **Rank Sum Ratio (RSR)** and **Systematic Clustering** to complete the classification tasks and compare the classification results to validate our model.

9.1 Model Construction and Classification Results

K-Means Clustering Model

K-Means Clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest cluster centers (or cluster centroid). We have also tried the central logarithm to process the compositional data but still choose to use the origin data after preprocessing after comparing the K-Means Clustering result.

Determine the Number of Clusters in K-Means

By using the **elbow rule**, we find that the aggregation coefficient decreases significantly before the numbers reach 3 and tends to be flat after that, and remain mainly stable after that. Therefore, we set the number of clusters to 3, and named the 3 groups by "hard", "normal" and "easy".

Classification Result

After performing K-Means Clustering on the 359 words known, we predict that **the word EERIE is classified into the "HARD" group**. The number of 359 words in "easy", " normal" and "hard" group are 134, 152, and 73 respectively. We also used **principal component analysis(PCA)** to reduce the data dimensionality. We visualize the classification result and the expectation of attempts for each of the 359 words, which is shown in Fig 11.

As we can see in the left part of Fig 11, the blue points represent the easy words, the orange points represent the normal words, and the green points represent the hard ones. And as the right part of the graph shows, the more transparent the blue dot is, the lower the expected number of attempts has, which means the word is easier to be solved. Considering the expectation of attempts represents the difficulty of a word in some aspects, then we can see the correspondence between our model classifications and the difficulty of a word by comparing two graphs together. Therefore, the classification of our model well reflects the change of actual difficulty, meaning our difficulty classification model has high classification performance. Also, from Fig 12, we can find that there are strong differences in the distribution of the attempts number of different classification results. Most people guess easy word correct after only three or four attempts, normal words after four attempts, while in hard word guesses, the attempts number rises significantly, up to four or five or even six attempts.

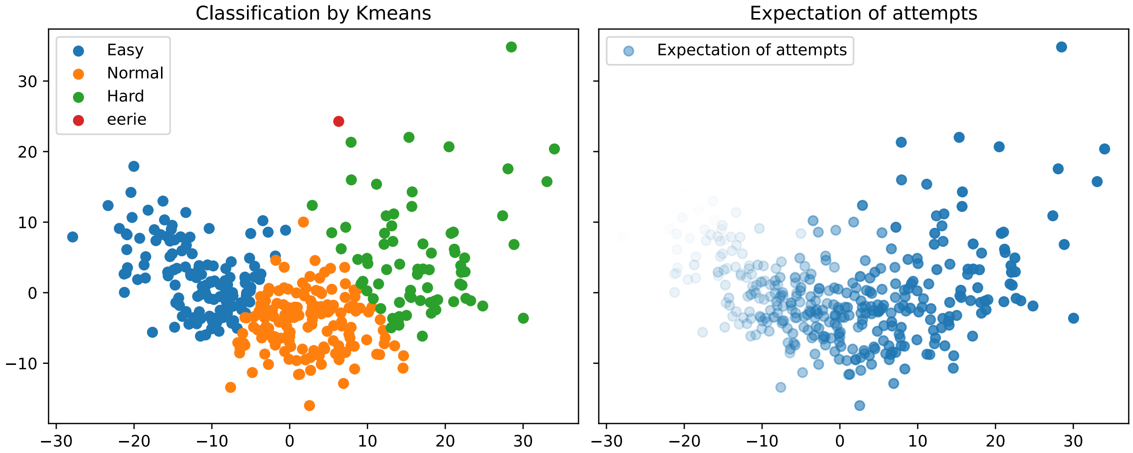

Figure 11: Scatter plot of classification result using PCA

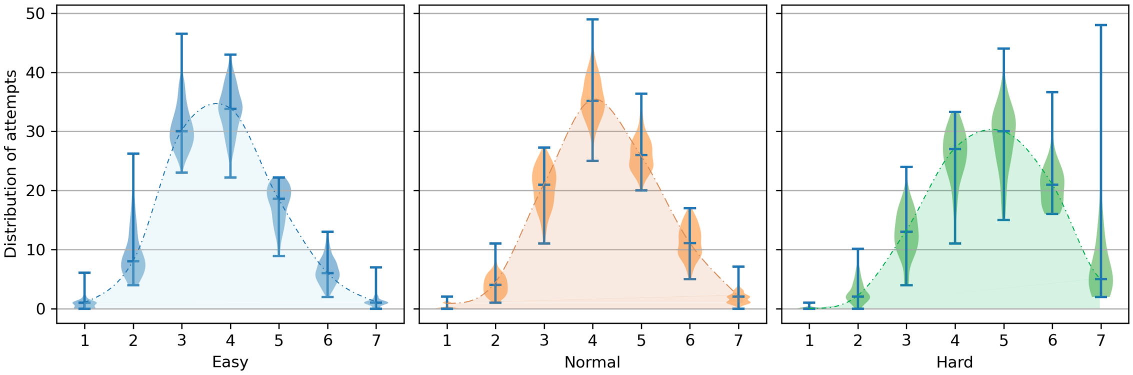

Figure 12: Violin plots of the distribution of classification results

9.2 Attributes of Words in Each Category

After applying the classification model on words, we need to find out the attributes corresponding to each category. We select the same 5 attributes we list in section 6.1 as difficulty indicators, which are f_{word}, f_{letter}, f_{bigram}, $repetition$, and $adjacency$.

9.2.1 Multivariate Logistic Regression Model

Multivariate logistic regression analysis is a formula used to predict the relationships between dependent and independent variables. It calculates the probability depending on multiple sets of variables. As we classified the words into 3 categories, which are “easy”, “normal”, and “hard”, indicating a tri-classification model.

The Multiple logistic regression model calculates the classification probability for each category and selects the one with the highest probability as the classification result.

Table 2: The weight of five difficulty indicators

Category	Letter Frequency	Letter Repetition	Word Frequency	Adjacency	Bigram Frequency
Easy	2.38548886	-1.34053858	1.34465633	-0.84567243	1.37055732
Normal	-0.39924048	0.34345807	-0.51728681	-1.12563792	-0.37230021
Hard	-1.98624838	0.99708052	-0.82736952	1.97131036	-0.99825711

9.2.2 Attributes Analysis Based on Regression Coefficients

We set the 5 indicators as variables and the three categories as targets and apply the Multiple logistic regression model on the 359 words. The regression coefficients are shown in Table 2. We can see that the coefficient of letter frequency is over 2.3 for the category “easy” while the ones for next two categories are negative, indicating that common words need less guesses. The similiar pattern exist in the coefficient of Word frequency and bigram frequency, showing that the common the letters and syllables in the word are, the easier for players to win the game. Also, the reverse trend exist in the other two indicators, which are duplicate letters and adjacency. It is clear that duplicate letters and similarity to other words make the Wordle game harder. This results are consistent with the coefficient analysis of our Dirchlet regression model in section 8.3, showing the reliability and accuracy of our model.

9.3 Model Validation

First we establish the difficulty evaluation model based on **Rank Sum Ratio (RSR)** and perform the classification based on the WRSR sorting. Secondly we perform **Systematic Clustering** with the 5 difficulty indicators as inputs data and classified the words into three categories.

9.3.1 Rank Sum Ratio (RSR) and Entropy Weight Method

RSR method is a statistical analysis technique fused the classical parameter estimation and modern non-parametric statistical methods. To eliminate human factors in our model, we use the **entropy weight method** to determine the weight of the 5 difficulty indicators.

9.3.2 Systematic Clustering

Systematic Clustering is a clustering model in which two types of data points are combined in pairs based on the minimum distance, and the merging process is repeated until all data points are classified into one type. We use the scores of 5 difficulty indicators as variables, and each word in the files represents an event for clustering.

9.3.3 Classification Results Comparision

Through calculation, our regression equation of Rank Sum Ratio (RSR) model is shown below:

$$RSR(WRSR) = a + b \times Probit \tag{13}$$

In our model, the parameter estimation for our WRSR is $a = 8.97, b = -28.97$. Applying the equation to our 359 words and we divided the words into 3 categories based on the same categories' ratio of our classification model. 305 words have the same classification results.We also compare the classification results of Systematic Clustering to ours and get 298 words the same.

These results shows high consistency between two difficulty classification model and prove the accuracy of Difficulty Classification Model Based On K-Means.

9.4 Sensitivity Analysis

Sensitivity analysis needs to study the influence of a certain change of relevant factors on one or a group of key indicators quantitatively. To test the robustness and reliability, we perform sensitivity analysis on our classification model by adding certain ratio of disturbance on our input distribution data. By studying the influence of the disturbance ratio on the classification results of the model, we can analyze the sensitivity of our classification model.

We add a p% of disturbance on the inputs distribution of our word-based compositional data and calculate the *F1-score* of the classification result in the 359 words. The graph in Fig 13 shows the relationship between the ratio i% and the corresponding *F1-score*. We can see that our *F1-score* remain over 0.875 with a disturbance under 6% and remain over 0.85 with a 9% of disturbance. This result indicates that our classification model is high in stability and sensitivity.

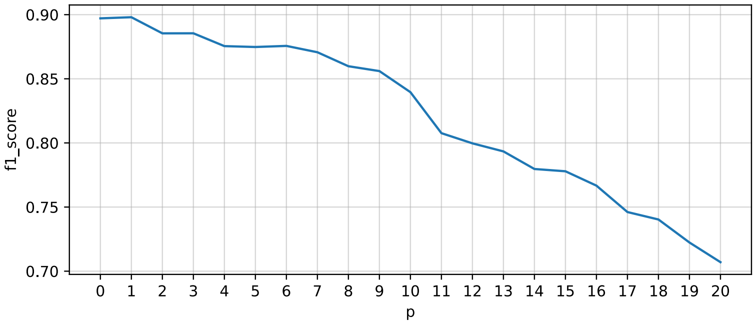

Figure 13: F1-score of classification model with disturbance

10 Interesting Features of the Data Set

10.1 Total Reports Number

During the process of model establishment, we notice that there are several interesting factors in the total reports number over time:

1. Total trend: The popularity of Wordle shot up rapidly in early 2022, peaking at February 20, 2022 while the decreasing trend fit the general Negative Exponential curve very well. The interest in Wordle over the course of 2022 so far mirrors the ups and downs in other viral games such as "Angry Birds" and "Among Us". The decreasing line may also be caused by other competitive word games since Wordle has inspired an innumerable collection of Wordle spinoffs and related games. The consistency of the trend shows the reliability of the model.

2. Hardcore players and common players As our model suggest in section 5, there are about 2.1021×10^4 hardcore players who remain engaged in the game. We also notice that as Fig 2 shows, the total number fluctuated around 21000 reports at the end of the year, which is approximately the number of hardcore players in our prediction. This means almost all the common players left and only fans will remain, which is a common rule in the viral games on websites. These reasonable explanation proves the reliability of our **two-group players** model.

3. Weekly trend: Also, we notice that there are weekly trend in the total numbers. More reports are generated during weekdays while weekend holds a smaller counts. We infer that people just use Wordle as a pastime on weekdays, and they like to have other recreation at weekends instead of challenging themselves to do the word puzzle.

4. Holiday effects: On Thanks Giving Day, there are surprisingly 6% people won the game with only one try. This is because the Wordle word “feast” is highly correlated with the holiday. Moreover, we also notice that on Christmas, the number of total reports drop by 23.3%(4727 people). The reason is obvious that people are busy with family dinners and receiving gifts.

5. Social effects: We found that almost all the words with more than 3% of “1 try” are the most common words like “world” and “dream”. However, the word “slate” holds 6% of “1 try” because “slate” is reported as the best opening words for Wordle just on the day before on social media.

10.2 Hard Mode Percentage

Interesting factors of the Hard Mode percentage over time:

1. Increasing trend: The percentage of reports choosing Hard Mode shows an increasing trend, indicating that maybe hardcore players who remain interested in Wordle tend to choose Hard Mode or maybe hardcore players play Wordle more now than before. This result has been proved by “WordFinderSurvey”[13], which declared that while some casual players may have abandoned the bandwagon at this point, more dedicated fans remain and are even more engaged with the game.

2. Social effect: According to WordleBot[5], The New York Times’ official Wordle analyst tool, roughly 20 to 25 percent of people play using hard mode. However, our data from Twitter shows that roughly less than 10% of people reports on Twitter are playing on Hard Mode. The huge gap indicates that hardcore players who tend to select Hard Mode seldom reports on Twitter.

10.3 Word Difficulty Level

Interesting factors in the difficulty level of the words:

1. **High-reel effect**:

   The graph in Fig14 shows the correlation between Twitter volume about the word and the average guesses of a word. It shows that the easier the words are, the more people will report it on Twitter, indicating that people are more likely to expose their perfect self in public which is called the “High-reel effect” on social media.

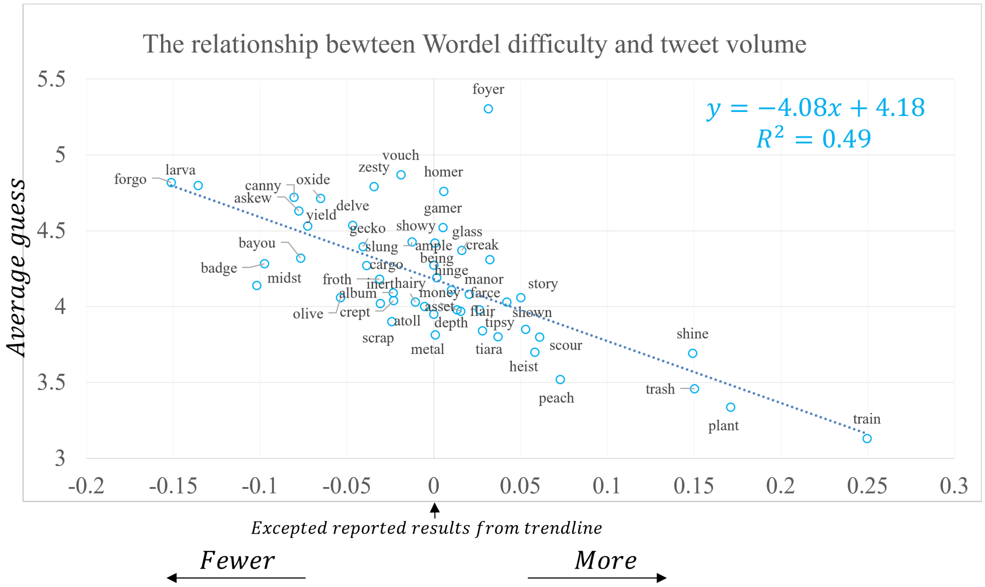

Figure 14: Wordle difficulty and tweet volume

2. **Is Wordle getting harder?**:

   According to the coefficients analysis of Dirchlet Regression model in section 8.3, we found that the distribution has little correlation with the time factor, which indicates that the difficulty level of words remain almost the same. However, as there are more tricks and hints available online and hardcore players are becoming the majority of the players online, Wordle should be "getting easier" over time. There are possibly two theories to explain the difficulty level, the disappearance of "High-reel effect" and the "hardcore player play first" theory, which is also mentioned in several websites[4].

10.4 The First Guessing Word

We conduct least squares regression between the word attribute and the correct rate of people's first guess, and analyze the multiple linear regression coefficients of each attribute of the word. We find that **letter frequency score** and **word frequency score** have a greater impact on the success rate of the first guess, indicating that people are more likely to use words with high frequency and more high-frequency letters as their first guess.

11 Model Evaluation and Further Discussion

11.1 Strengths

Our model offers the following strengths:

- **Comprehensive Feature selection**
  Based on the literature and websites, we comprehensively consider all the characteristics that

may affect the difficulty of words, and use correlation test to select the factors that have the greatest impact on the distribution and difficulty.

- **Model selection based on data distribution**
First, we fully consider the features of compositional data under the unit-sum constraint, we select the Dirchlet regression model, which can explain the different trends and covariance structures and does not require the independence of the data. It is also flexible in choosing the link function and incorporating prior knowledge about the data. Secondly, we select K-Means Clustering and Systematic Clustering respectively for distribution data inputs and 5 indicators inputs.

- **Reasonability and high interpretability**
Our derivation of the **Negative Exponential** trend based on reasonable assumption and our **two-group players** theory is highly interpretable and can provide perfect explanations.

- **Multiple methods for one question for optimal results** We use three methods to validate our Dirchlet regression model in Problem 2 and use 2 methods to prove the accuracy of our classification model in Problem 3. What's more we use three models in our prediction of the total reports number, combining advantages of Prophet and XGBoost, which is conducive to the optimal results.

11.2 Weaknesses

Our model has the following limitations and related improvements:

- **The difficulty of words is not considered in the total reports prediction model**
Due to the "Highlight Reel" Effect, people are more likely to reports their Wordle scores if they get the right words in less attempts. Therefore, the difficulty level of words are likely to affect the total reports numbers on Twitter. However, our model don't take it into consideration directly because we don't know the exact word as input in our prediction process.

- **Abstract factors are not considered in feature selection**
It is hard to quantify the readability of words and the influence of spoken and written language on the results. It is hardly possible to include social factors such as specific news on Twitter in the word guessing process.

11.3 Model Extension

- **The popularity decreasing model**
Our popularity prediction model based on Negative Exponential Model can be applied on other viral games. If given the word, we can also consider the correlations between reports numbers and the difficulty of the word based on the "high-reel effect" explained in problem 4.

12 Letter to the Puzzle Editor

To: Respected Editors of the New York Times
From: MCM Team #2307432
Subject: Cracking the Code: Twitter Data Insight in the Wordle Game
Date: Feburary, 21, 2023

With great honor to accept the assignments of your newspaper, we are writing to present our analysis result based on Twitter reports on Wordle data and give corresponding suggestions to help you attract more readers and promote the popularity of the puzzle game.

First, we analyze the number of daily reports on Twitter and propose the Two-group Players Model and a popularity prediction model. Based on the line graph above, We found that the popularity of Wordle shot up rapidly in early 2022, reaching a peak at 361908 reports on February 2 in 2022 and continued to decline. We divide the players into hardcore players and common players. While hardcore players will keep playing Wordle every day, the number of common players decreases at a fixed rate. After fitting the general trend to the negative exponential curve, we predict that there are about 21020 hardcore players and only them will remain engaged in the Wordle game in 2023. Then, we choose a Prophet Forecasting model to capture the seasonal effects and the Gradient Boosting Regression Tree to model the residual of the time series. We predict the total reports numbers for the next three months and give the prediction interval of 17900 to 26600 for the first season in 2023. Based on the declining trend of the negative exponential curve, We suggest that Wordle can cooperate with other puzzle games such as crossword to attract attention of common players. Wordle can set a rewarding daily challenge that allows players to guess as many words as possible within the allotted time, and the top players each day can earn rewards or bonus points. This may be helpful to increase daily attention.

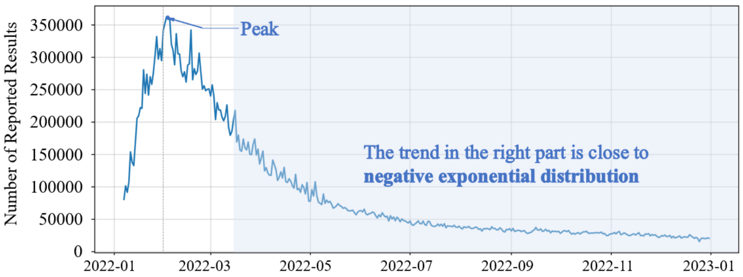

Figure 15: Line graph of total reports number on Twitter

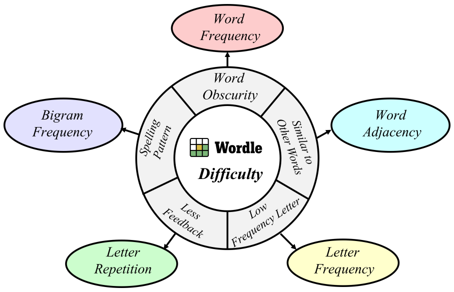

Figure 16: Wordle Difficulty indicators

Secondly, we notice that there are weekly trend and holiday effects in the total numbers. More reports are generated during weekdays while weekend holds a smaller counts. We infer that people like to have other recreation at weekends instead of challenging themselves to do the word puzzle. On Thanks Giving Day, there are surprisingly 6% people won the game with only one try due to the high correlation of the word “feast” with the holiday. The number of total reports drop by 23.3%(4727 people) on Christmas because people are busy with family dinners and receiving gifts. Therefore, we suggest that Wordle can set the weekend challenge and add some bonus on holidays to attract more attention.

Additionally, we find some social factors also have a huge impact on the Wordle data. Therefore,

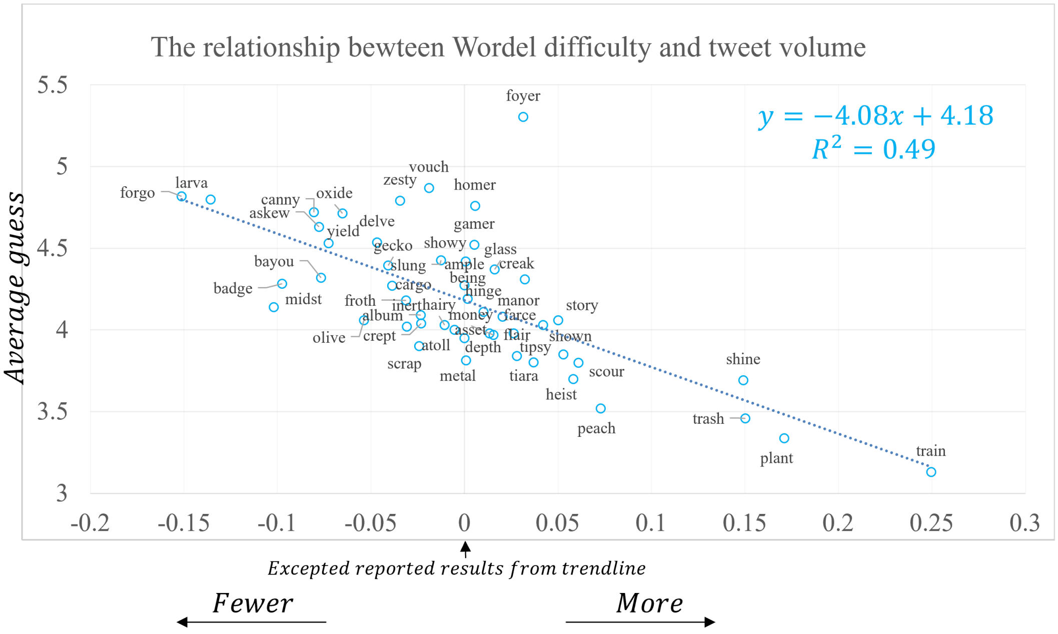

Figure 17: Wordle difficulty and tweet volume

promotion of the Wordle game using social media platforms, building a player community for communication and sharing may boost the social effects. The graph in Fig 16 shows the correlation between Twitter volume of the word and the corresponding average guesses. We compare the residuals (the difference between actual and expected values) to the average score for each word and find that the word difficulty level alone explains 49% of the variation in Twitter volume. It shows that the easier the words are, the more people will report it on Twitter, indicating the “High-reel effect” on social media. Utilizing the psychological factors, Wordle can create a points and rewards system, a leader board and achievement system that allow players to compete and challenge. Adding a cooperative mode so that players can team up to guess the word may increase interaction and sociality. Also, personal settings of the website including background and font color can be customized. All the measures above may simulate the players’ interest.

Then, we summarize the possible attributes of words and use our data to test their correlation with the difficulty level. We use 5 quantitative indicators to represent the valid attributes, which are word frequency score f_{word}, letter frequency score f_{letter}, bigram frequency score f_{bigram}, duplicate letters $repetition$, and the similarity factor $adjacency$. The difficulty level indicators are shown in Fig 17:

Based on the difficulty indicators, we develop the Difficulty Classification Model. We classify the words into three groups “easy”, “normal” and “hard” using K-means clustering. We then use the coefficients of Multivariate Logistic Regression Model to explain corresponding word’s attributes. Results indicate that common words need less guesses, words containing more common letters and syllables are easier but duplicate letters and adjacency have the reverse effects. Based on our Difficulty Classification Model, Wordle can set more difficulty levels, which can benefits people of different ages and competence.

Hopefully, the New York Times will take our suggestions into consideration, and we all believe Wordle will retain its animation in the future.

References

[1] J. Aitchison, (2003). The Statistical Analysis of Compositional Data, Caldwell, NJ: The Blackburn Press.

[2] https://wordfinder.yourdictionary.com/blog/the-hardest-wordle-puzzles-to-date-what-does-tacit-even-mean/

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

[4] https://observablehq.com/@rlesser/wordle-twitter-exploration

[5] https://wordfinder.yourdictionary.com/blog/wordle-hard-mode-what-is-it-who-plays-it-and-should-you-play-it-too/

[6] 2005, Baxter, M.J.; Beardah, C.C.; Freestone, I.C. (2005) Compositional analysis of archaeological glasses'. In CODAWORK'05. Eds. G. Mateu i Figueras and C Barceló i Vidal. Girona: La Universitat, 2005 .

[7] R. Connor and J. Mosimann, (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution, J. Amer . Statist. Assoc. 64, 194–206.

[8] G. Campbell and J. Mosimann, (1987b). Modeling continuous proportional data with the Dirichlet distribution, Unpublished manuscript.

[9] G. Campbell and J. Mosimann, (1987a). Multivariate methods for proportional shape, ASA Proceedings of the Section on Statistical Graphics, 10–17.

[10] R. Gueorguieva, R. Rosenheck and D. Zelterman, (2008). Dirichlet component regression and its applications to psychiatric data, Comput. Statist. Data Anal. 52, 5344– 5355.

[11] http://norvig.com/mayzner.html

[12] Solso, R.L., Barbuto, P.F. & Juel, C.L. Bigram and trigram frequencies and versatilities in the English language. Behavior Research Methods & Instrumentation 11, 475–484 (1979).

[13] https://wordfinder.yourdictionary.com/blog/is-wordle-dying-the-data-weighs-in/

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

[15] https://wordlestat.com/

[16] Roybyn Speer

[17] http://crr.ugent.be/archives/1352