Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2316012

Wordle Analytics: Leveraging Twitter Data to Decode the Puzzle Difficulty and Popularity

With the proliferation of online games, more and more players are sharing their results on social media. How to leverage the reported outcomes to better design the game becomes an increasingly important question for game developers, which helps attract more new users and retain old users.

To address this issue, we focus on a popular puzzle Wordle where players need to guess a five-letter word in six tries or less and conduct a series of analyses based on data collected from daily reported results from Twitter. We extract several attributes that may affect the difficulty of the solution word, including the word frequency in the corpus, the number of repeated letters, and whether the word has less-used letters. We also construct a new measure of word difficulty by considering the distance between the solution word and the optimal starting word suggested by prior studies. Based on these word attributes, we conduct Principal Component Analysis (PCA) and build a K-Means Clustering model to classify words into four difficulty levels, which suggests high accuracy and generalizability.

Moreover, to describe the pattern of the number of reported results, which serves as a proxy for the popularity of the puzzle, we draw on the product life cycle theory and develop Ordinary Differential Equations (ODE) to capture the fundamental trend of user participation. To further explain the time-series fluctuations, we also apply Auto-regressive Integrated Moving Average (ARIMA) model to achieve higher prediction accuracy. In terms of the proportion of players with hard mode, we build ARIMA with Explanatory Variables (ARIMAX) model to fit the trend of this proportion. We find that the number of reported results and word difficulty (e.g. word frequency) have significant impacts on the propensity to choose the hard mode. With harder solution words, players are more likely to report their game outcomes and exhibit a larger proportion of hard-mode games.

In addition, we develop a model based on Random Forest (RF) Regression to predict the distribution of the reported results, which reflects the performance of players given the solution word. Our proposed model can take the multi-output nature of the prediction task into account and renders an interval prediction of the percentage of each category of player tries.

Finally, we conduct several sensitivity analyses and validate the model's robustness. To sum up, our models provide important managerial implications for the design of online games.

Key Words: Online Puzzle, Time-Series Analysis, Ordinary Differential Equation, Principal Component Analysis, Random Forest Regression, K-Means Clustering

Contents

1 Introduction

1.1 Background

Wordle is a popular online word-guessing game that has gained immense popularity in recent times. The objective of the game is to guess a five-letter word chosen by the game's algorithm. Each round, the player enters a word and the game indicates how many letters are correct and in the correct position by displaying them in green. If a letter is correct but in the wrong position, it is displayed in yellow. The player has six attempts to guess the word correctly. In addition, players can also choose to play in Regular Mode or Hard Mode. Playing in Hard Mode on Wordle requires players to use any correct letter they've guessed in subsequent attempts, which increases the difficulty of the game. As shown in the right panel of Figure 1, the example was played in Hard Mode.

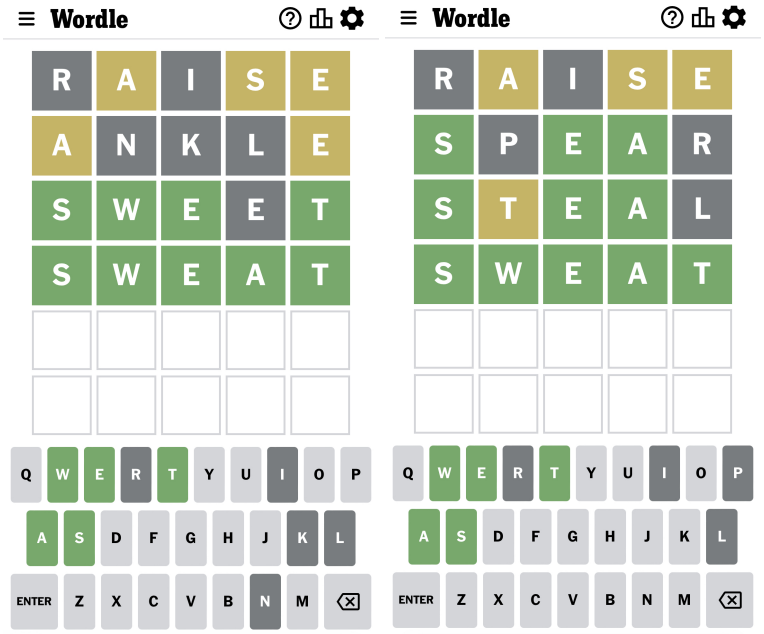

Figure 1: Wordle Game Example in Regular Mode (Left) and Hard Mode (Right)

Players can share the results to social networks after playing the game. In this analysis, we will use the reported information collected from Twitter to provide insight into the data.

1.2 Literature Review

Prior studies have analyzed what affects word difficulty and illustrated word puzzle strategy. Yang and Suyong (2018) proposed Natural Language Processing (NLP) to extract features that represent the difficulty level of words, followed by clustering and classification using Gaussian Mixture Model

(GMM) and Support Vector Machine (SVM). Hiebert et al. (2019) examined a different set of word features that contribute to the vocabulary difficulty in the aspect of educational learning. Basu et al. (2019) explored how to predict word difficulty by dividing words with determined difficulty into two classes using methods including Convolutional Neural Networks (CNN). With the prevalence of word puzzle games like Wordle in recent years, several relevant works have also been conducted. Anderson and Meyer (2022) presented a framework for discovering the optimal human strategy based on Reinforcement Learning (RL) for Wordle players. Li (2022) documented the relationship between the average success of Wordle players on Twitter and the linguistic characteristics of words.

However, despite various methods, they are not fully applicable to our study with a limited set of data and no predetermined difficulty level. For the specific Wordle reported results, more analysis is yet to be conducted (e.g. the distribution from 1 to 7 or more tries and the percentage of hard mode). In addition, there are more factors to consider in the context of our dataset (e.g., not everyone would report when they finish the game and the effect of time series should be taken into account). The innovation and contribution of our study are to model, analyze, predict, and verify the traits and difficulty of solution words, the number and distribution of reported data, and the consideration of several other factors from the perspective of game players using a variety of models.

1.3 Models Overview

Our main goal is to address three issues that need to be explored:

1. explain the variation in the reported results, create a prediction interval for the number of reported results on March 1, 2023, and investigate the attributes that affect the percentage of scores reported that were played in Hard Mode;

2. develop a model to predict the distribution of reported results for a given solution word on a future date, with associated uncertainties and provide a specific example for EERIE on March 1, 2023, and discuss the confidence in the prediction;

3. develop a model to classify solution words by difficulty and identify the attributes associated with each classification. Evaluate the difficulty of the word EERIE and discuss the accuracy of the model.

To solve these tasks, we build three models in general, i.e. the model of word difficulty classification, the model of user participation, and the model of reported results distribution. In section 2, we give a detailed illustration of the data set, word attributes, and relevant assumptions. In section 3, we focus on the model of word difficulty classification, where we apply the method of Principal Component Analysis (PCA) and K-means, and thus solve task 3. In section 4, we focus on the model of user

participation, where we apply the method of Ordinary Differential Equation (ODE), Auto-regressive Integrated Moving Average (ARIMA), and ARIMA with Explanatory Variables (ARIMAX), and thus to solve the task 1. In section 5, we focus on the model of reported results distribution, where we apply the method of Random Forest (RF), and thus solve task 2. In sections 6 and 7, we further conduct a sensitivity analysis and model evaluation. The models' overview is shown below.

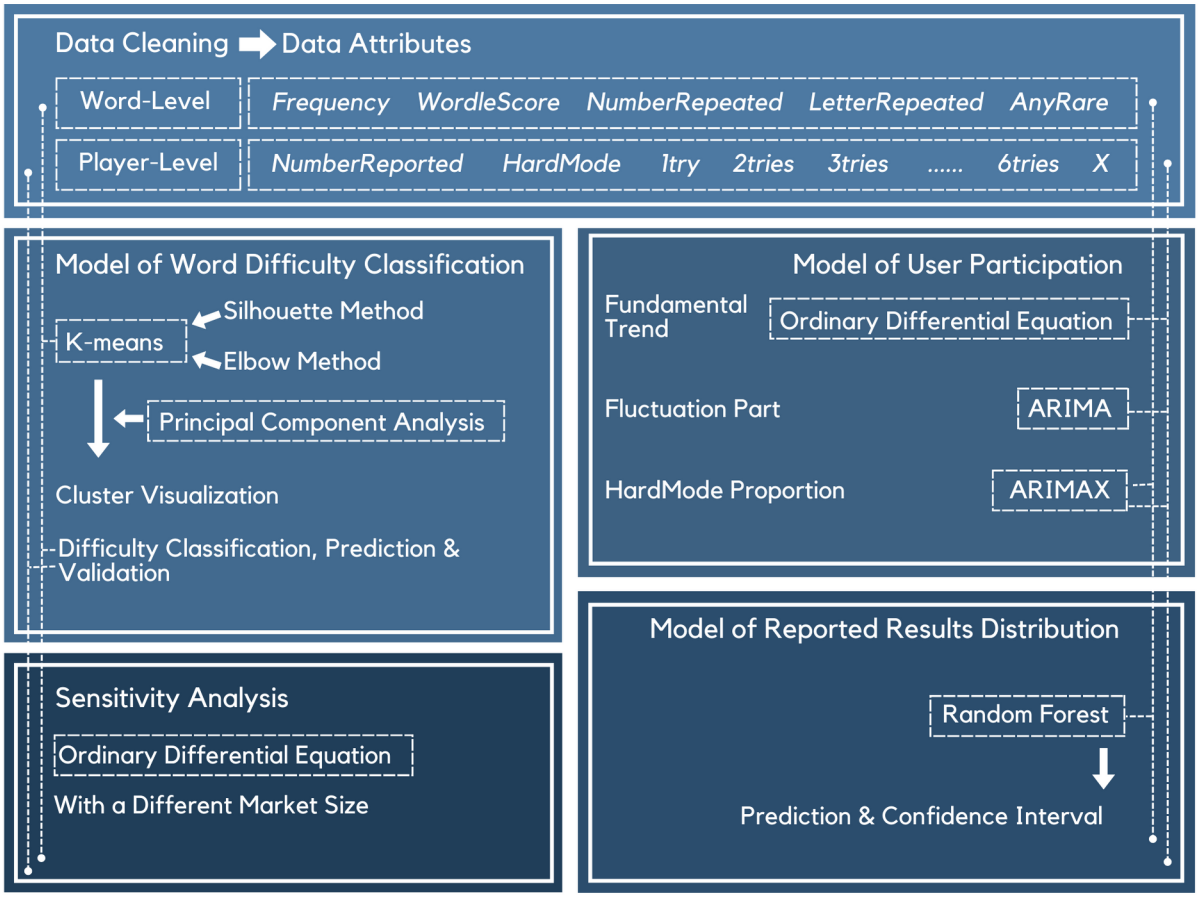

Figure 2: Models Overview

2 Variables and Assumptions

In this section, we present how we construct the variables for analysis and the assumptions for our models.

2.1 Game Features and Data Cleaning

The given dataset presents the features of the Wordle game gleaned from daily reports on Twitter, ranging from January 7, 2022 to December 31, 2022. To clean the data, We examine the presence

of outliers and wrong values. First, although all words are deemed to have five letters, several words are misspelled or misformatted, resulting in a smaller or larger string length. After checking relevant information, we correct "tash" to "trash", "clen" to "clean", "rprobe" to "probe", "naïve" to "naive", "marxh" to "marsh", and "favor " (extra space at the end) to "favor". In addition, the number of reported results on November 30 is an obvious outlier so we replace it with the mean value of the observation one day before and one day after. Furthermore, the percent in 1 to 7 or more tries on March 27 has a sum of 126%, which is significantly greater than 100% and we change the percent by normalizing the total percentage to 100%.

2.2 Word Attributes

We compile several word attributes that may potentially affect the difficulty of the solution word in Wordle. First, we acquire the word frequency from NLTK library, denoted as $Frequency$.[1] We select several packages from NLTK including webtext (a small collection of web text), nps_chat (a corpus of instant messaging chat sessions), brown (a million-word electronic corpus of English), gutenberg (a corpus of some 25,000 free electronic books) and reuters (a corpus of 10,788 news documents), which are representative of the whole corpus.

Second, to measure the difficulty of the solution words, we try to simulate the guessing procedure and assign different values to the information players can obtain. Anderson and Meyer (2022) show that some starting words could lead to a higher chance of successful guessing, including *later*, *soare*, *raise*, *roate* and *slate*. Therefore, we define a new variable to measure the average similarity between the optimal starting words and the targeted word based on Levenshtein Distance (Levenshtein 1966), denoted as $WordleScore$ (WS). The Wordle Score between two strings a, b (of length $|a|$ and $|b|$ respectively) is given by WS(a, b) as follows:

$$WS(a,b) = \begin{cases} 0 & \text{if } |b| = 0, \\ 0 & \text{if } |a| = 0, \\ 1 + WS(tail(a), tail(b)) & \text{if } a[0] = b[0] \text{ and } |a| = |b|, \\ 0.25 + WS(tail(a), tail(b)) & \text{if } a[0] = b[0] \text{ and } |a| \neq |b|, \\ \max \begin{cases} WS(tail(a), b), \\ WS(a, tail(b)), \\ WS(tail(a), tail(b)). \end{cases} & \text{otherwise.} \end{cases} \tag{1}$$

[1]Since this variable is highly skewed, we take natural logarithms after adding one to the variable.

where the tail of some string x is a string of all but the first character of x, and $x[n]$ is the nth letter of the string x. The original Levenshtein Distance measures the difference between two sequences by calculating the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other. The larger the distance, the greater the difference between the two words. But here instead of measuring the difference, we use $WordleScore$ to measure the similarity by assigning different values for those words containing the same letters in the right position or wrong position. For letters in the right position, i.e. the green notation in the Wordle game, we assign a value of 1 since we get the full information in this specific position; for letters in the wrong position, i.e. the yellow notation in the Wordle game, we assign a value of 0.25. We select this value as each position represents an information score of 1 and there are other four positions and a 25% chance to fill the letter to the right position. For example, the word *raise* and *there* have one same letter in the right position (e) and one same letter in the wrong position (r), so WS(raise, there) has a value of 1.25. Finally, we obtain $WordleScore$ for each word in our data set by taking the average of $WordleScore$ based on the five optimal starting words.

Third, we counted the total number of repeated letters in each word, denoted as $NumberRepeated$.[2] To be more precise, we also recorded the specific letter that repeats in the data set of words, denoted as $LetterRepeated$.[3] In addition, we also define a dummy variable $AnyRare$ to indicate whether the word has rare letters, including q, j, x, z, and v that have the lowest frequency in the corpus. Table 1 reports the definitions and summary statistics of our main variables.

2.3 Assumptions

To simplify our models and capture the main insights from data, we make the following main assumptions in this paper.

Assumption 1: The difficulty of guessing a certain target word in Wordle is independent of its meaning.

Assumption 2: Once a user leaves Wordle, he/she will never play it again. On the first day of our data (Jan 7th, 2022), no one had left.

Assumption 3: In general, the higher the frequency of Solution Words, the higher the Wordle Score, and the fewer repeated letters, the easier it is.

Assumption 4: The percentage of players solving the puzzle in more tries is positively correlated with word difficulty.

[2]We construct the variable by subtracting the number of unique letters from the total length of the word. Take the word *apple* for example, it has 4 unique letters and therefore the number of repeated letters is 1.

[3]This variable is a categorical variable and we convert it to several dummy variables for subsequent analysis.

Table 1: Definitions and Summary Statistics of Main Variables

Variable	Definition	Mean	Std. Dev.	Min	Max
Player-level variables					
$NumberReported$	Number of reported results	90976	89224	15554	361908
$HardMode$	Number of players with hard mode	5098	3167	1362	15369
$1try$	Percentage of players with one guess	0.47	0.78	0	6
$2tries$	Percentage of players with two guesses	5.84	4.08	0	26
$3tries$	Percentage of players with three guesses	22.73	7.78	4	47
$4tries$	Percentage of players with four guesses	32.93	5.35	11	49
$5tries$	Percentage of players with five guesses	23.64	5.95	9	44
$6tries$	Percentage of players with six guesses	11.56	6.21	2	37
X	Percentage of unsuccessful players	2.81	4.12	1	48
Word-level variables					
$Frequency$	Frequency in the corpus (logged)	3.22	2.12	0	9.45
$WordleScore$	Score derived from WS equation	0.80	0.44	0	2.65
$NumberRepeated$	Total number of repeated letters	0.30	0.49	0	2
$AnyRare$	Whether the word has rare letters	0.03	0.18	0	1

3 Model of Word Difficulty Classification

In this section, we try to classify the solution words by their difficulty based on the word attributes defined previously, i.e. $Frequency$, $WordleScore$, $NumberRepeated$, $LetterRepeated$ and $AnyRare$, where the K-Means clustering method is adopted. In operationalizing this, we normalize word features by subtracting the mean and dividing the standard deviation, where the distance between points is then defined as the Euclidean distance.

3.1 Principal Component Analysis

To determine the main components in our classification, we first apply Principal Component Analysis (PCA). It is a multivariate statistical method often used for reducing dimensionality. It transforms a group of variables with possible correlation into linearly unrelated ones through orthogonal transformation, and the converted variables are called principal components (Pearson 1901).

We apply PCA to analyze the data set. First, we standardize the $Frequency$ and $WordleScore$ by

$$X_1 = \frac{F - \mu_F}{\sigma_F}, X_2 = \frac{W - \mu_W}{\sigma_W} \tag{2}$$

where μ is the mean, σ is the standard deviation, and X is the standardized data. To be precise,

X_1 is the $Frequency$ and X_2 is the $WordleScore$. We also include other features, i.e. X_3 is the $NumberRepeated$, X_4 is the $AnyRare$.

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_P \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_P \\ \cdots \\ Z_P = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_P \end{cases} \tag{3}$$

In the above equation, for every i, we have the following conditions. (1) $c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2 = 1$, and $[c_{11}, c_{12}, \cdots, c_{1p}]$ maximizes the variance of Z_1; (2) $[c_{21}, c_{22}, \cdots, c_{2p}]$ is orthogonal to $[c_{11}, c_{12}, \cdots, c_{1p}]$, and it maximizes the variance of Z_2; (3) $[c_{31}, c_{32}, \cdots, c_{3p}]$ is orthogonal to $[c_{11}, c_{12}, \cdots, c_{1p}]$ and $[c_{21}, c_{22}, \cdots, c_{2p}]$, and it maximizes the variance of Z_3. And so on, we can derive the principle components.

Here we take the number of components of 3 to analyze, whose cumulative variance contribution reaches 95%. The PCA results are shown below.

Table 2: Results of Principal Component Analysis

PC	Explained Variance	Variance Contribution Rate	Cumulative Variance Contribution Rate
1	1.05	45.90%	42.90%
2	0.97	42.52%	85.42%
3	0.23	10.17%	95.59%

$$\begin{cases} Z_1 = 0.6993X_1 + 0.7057X_2 + -0.1132X_3 + -0.0086X_4 \\ Z_2 = -0.7106X_1 + 0.7036X_2 + -0.0031X_3 + -0.0042X_4 \\ Z_3 = 0.0773X_1 + 0.0824X_2 + 0.9932X_3 + -0.0268X_4 \end{cases} \tag{4}$$

Here from the results, it is clear that Z_1 and Z_2 explained most of the variance, about 85%. Also, X_1 and X_2 have a relatively larger contribution to Z_1 and Z_2, while X_3 contributes the most part of Z_3, with a coefficient of 0.9932. Compared with the other three variables, X_4 has little contribution. In other words, we can take $Frequency$, $WordleScore$, and $NumberRepeated$ as the main components.

3.2 Optimal Number of Clusters

To determine the optimal number of clusters, we use two metrics to compare the results with different numbers of clusters. First, we compute the Silhouette Score for each K, which is defined as

$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)} \tag{5}$$

where a represents the average within-cluster distance and b represents the average between-cluster distance. A Silhouette Score with a value near 1 means the data point is in the correct cluster. A Silhouette Score with a value near 0 means the data point might belong in some other cluster. And a Silhouette Score with a value near -1 means the data point is in the wrong cluster. We also calculate the within-cluster sum of squared error (SSE) as follows

$$\text{SSE} = \sum_{i}^{K} \sum_{x \in C_i} dist^2(m_i, x) \tag{6}$$

where x is a data point in cluster C_i and m_i is the center of cluster C_i.

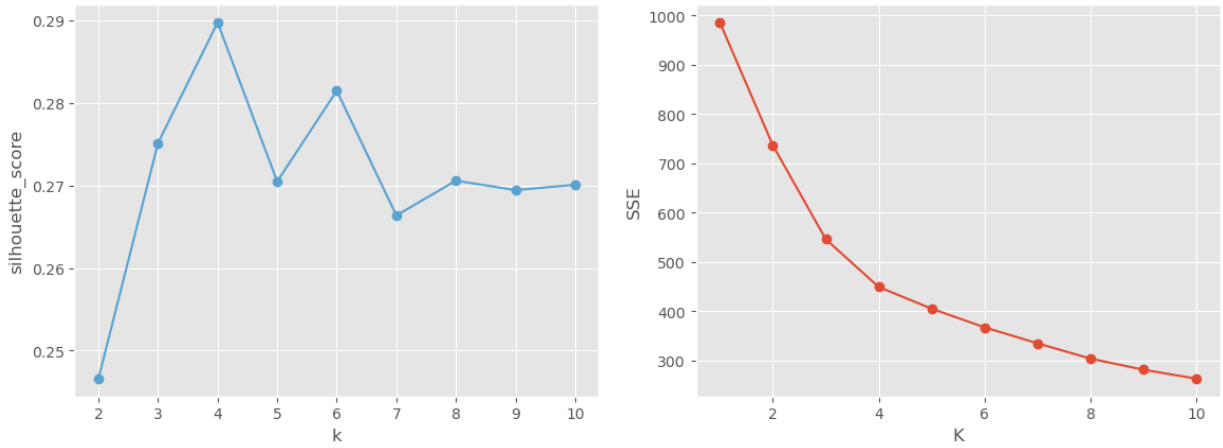

Figure 3: Selection of Optimal Number of Clusters

The results using different numbers of clusters are presented in Figure 3. In the left panel, it can be seen that the score rises with the number of clusters increasing and reaches the peak with 4 clusters, which indicates that the optimal K is 4. In the right panel, we further confirm that the elbow takes place at the point where $K = 4$.

3.3 Cluster Visualization and Difficulty Definition

From the results of Principal Component Analysis (PCA) in the previous section, we derive three features, i.e., $Frequency$, $NumberRepeated$, and $WordleScore$, that play a significant role in the first three principal components. To visualize the distributions of words with different difficulty levels, we plot the clusters of words where the three features serve as the coordinates in Figure 4.

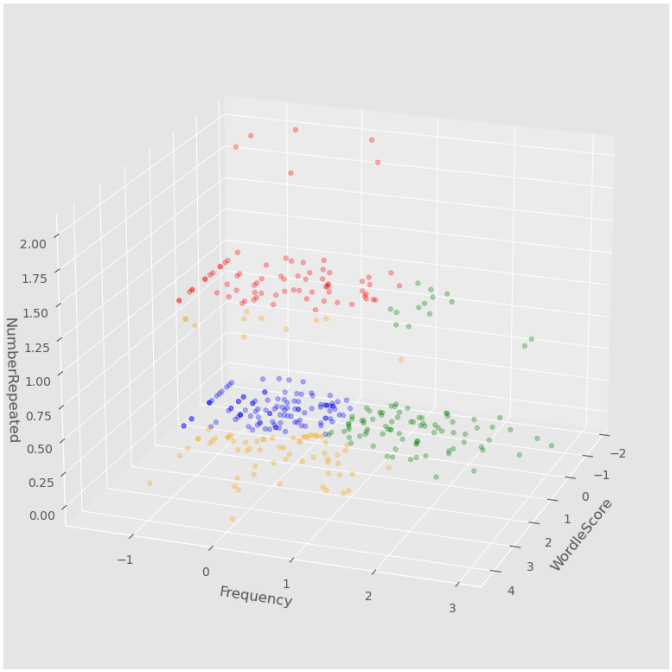

Figure 4: Clusters of Words

The scatter points in different colors represent the clustering results. It is clear that compared with the other 3 clusters, the green cluster exhibits an obviously larger $Frequency$. The yellow cluster presents a larger average $WordleScore$, while the rest 2 clusters are nearly the same. Despite the similar $Frequency$ and $WordleScore$ for the blue and red clusters, the latter cluster shows a significantly larger $NumberRepeated$. Using these significant features, we can classify the words into four difficulty levels. For Frequency, a higher frequency means that we are more likely to see these words in our daily lives and that they are more likely to be simple words, so we define a higher frequency as a lower difficulty level; for $WordleScore$, according to our previous definition, a higher score makes the target word easier to find, so we define a higher $WordleScore$ as a lower difficulty level; for $NumberRepeated$, because during the game, after knowing that a letter exists in the target word, players will be more inclined to try other letters and ignore that the letter may be repeated, thus making it more difficult to guess the word, so we define that the lower the number of repeated letters, the lower the difficulty level. Finally, we can classify the four clusters by their difficulty level. We

define the green cluster as difficulty level 1, which is the easiest; we then define the yellow cluster as difficulty level 2;[4] next, we define the blue cluster as difficulty level 3; and finally, we define the red cluster as difficulty level 4.

Table 3: Difficulty Level Classification

Difficulty Level	Cluster Color	*Frequency*	*WordleScore*	*NumberRepeated*
Level 1	Green	High	Low	Low
Level 2	Yellow	Low	High	Low
Level 3	Blue	Low	Low	Low
Level 4	Red	Low	Low	High

3.4 Difficulty Prediction and Evaluation

Applying the classification model, we can successfully derive the difficulty level of words. Take the word *EERIE* as an example, we take its attributes, i.e. $Frequency$, $WordleScore$, $NumberRepeated$, $LetterRepeated$, and $AnyRare$, as the input, and our model renders a predicted difficulty level of 4. Since the letter E appears in the word *EERIE* three times and this word is indeed rare, it's intuitive to understand why it has been assigned to the most difficult level, providing supportive evidence of our proposed classification model. In Table 4, we also present the attributes and predicted difficulty level of other words in the data set and the results are consistent with our intuition.

Table 4: Some Examples of Word Difficulty Classification

Word	*Frequency*	*WordleScore*	*NumberRepeated*	*LetterRepeated*	*AnyRare*	*Difficulty Level*
EERIE	1.95	1.1	2	*E*	0	4
FOYER	1.61	1	0	NA	0	3
SHAME	5.48	1.95	0	NA	0	2
LEAVE	6.80	1.8	1	*E*	1	1

Next, by taking the average of the features in each group, we can make a grouped bar chart and a 100% stacked bar chart, as shown in Figure 5. In the left panel, the $Frequency$, $WordleScore$, and $NumberRepeated$ are shown with regard to the difficulty level. It is clear that the $Frequency$ tends to decrease with the rise of the difficulty level, although the average $Frequency$ of level 4 is slightly larger than level 3. In addition, in terms of the $WordleScore$, we can observe that the values of levels 1 and 2 are obviously larger than those of levels 3 and 4. Furthermore, the $NumberRepeated$ of level 4

[4]Though the green cluster has a larger $Frequency$ and the yellow cluster has a larger $WordleScore$, the contribution of $Frequency$ we derived from PCA for clustering is greater than $WordleScore$, we believe that $Frequency$ is prioritized over $WordleScore$ in terms of classification criteria.

is greatly larger than the other three levels. In general, we can intuitively understand the main features of words at different difficulty levels. In the right panel, the pattern is rather obvious. We calculate the average proportion of the number of tries for each difficulty level and stack them with a scale of 100%. It is noticeable that, with the number of tries increasing from $1try$ to X, the proportion of the higher difficulty level is also rising. For example, the proportion of level 1 in $1try$ is nearly 40%. However, it ends up with a proportion of a mere 20% in X. But for level 4, while it only has a proportion of less than 10% in $1try$, the number increases a lot, with a value of about 35% in the end.

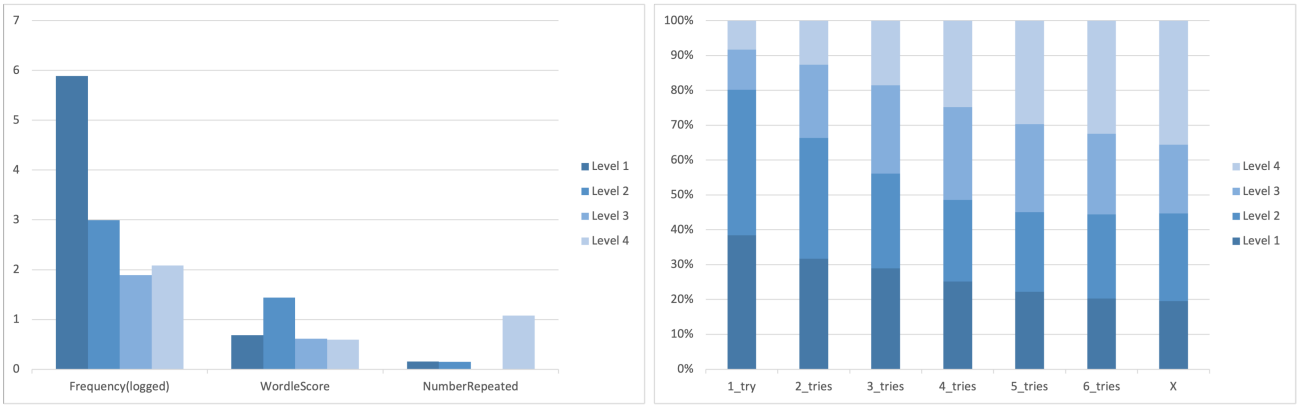

Figure 5: Features by Difficulty Levels

Initially, we categorized word difficulty using word-related features. Then we presented the features of the solution words in the data set according to our classification, as shown above. Interestingly, the two charts provide a quite good validation of our difficulty level classification, especially for the proportion of each number of tries shown in the right panel. Therefore, we believe that our classification has good accuracy in general.

4 Model of User Participation

4.1 Number of Reported Results

In the context of Wordle, we use the number of reported results as a proxy for the number of players every day, which can be decomposed into two parts: *fundamentals* (e.g. time trend of the popularity) and *fluctuations* caused by other factors (e.g. difficulty of the words).

$$R(t) = T(t) + f(t, ...) \tag{7}$$

To start with, we visualize the trend of the number of reported results in Figure 6. It can be seen that

there exists a peak of user participation, followed by a decline in the long term. To describe the pattern of user participation, we utilize the well-established product life cycle model (Bass model) (Kumar and Swaminathan 2003). Players get exposed to Wordle through advertising and word of mouth, which provides a channel for diffusion. Moreover, Wordle is different from other traditional products in the sense that it's an online game with little costs and monetary incentives. Therefore, we can develop a theoretical model to describe the trend of the popularity (number of reported results) of the game.

Suppose that there is a fixed maximal potential market size for Wordle, which is assumed to be 700,000. Based on the Bass model, we construct an Ordinary Differential Equations (ODE) model as follows:

$$\begin{cases} i(t) = \dfrac{\mathrm{d}I}{\mathrm{d}t} = (p + qI(t))S(t) - \mu I(t) = pS(t) + qI(t)S(t) - \mu I(t) \\ \\ s(t) = \dfrac{\mathrm{d}S}{\mathrm{d}t} = -(p + qI(t))S(t) \end{cases} \tag{8}$$

where $I(t)$ denotes proportion that is in Wordle at time t and $S(t)$ denotes that has never entered. To simplify the problem, we assume that no one had left at time $t = 0$, $S(0) + I(0) = 1$. For the parameters, p represents the impact of media influence, q represents the impact of cumulative fractions in Wordle, and μ represents the fixed proportion that leaves Wordle per day.

Since ODE has no analytical solution, we try numerical methods to get solutions of those time points. The parameters we finally adopt are $p = 1.13$, $q = 0.11$ and $\mu = 0.0011$. As can be seen from Figure 6, the estimated trend of the number of reported results using the ODE model is basically consistent with the actual trend, especially during the popularity growth stage before the peak. The discrepancy between the fitted curve and the actual curve can be attributed to some assumptions made. First, the potential maximal market size might be mis-specified, resulting in a later-coming maximal value of the estimated curve. Moreover, we may ignore some loyal players who continue to play the game throughout the study period, which causes a smaller minimal value in the fitted model. Nevertheless, the proposed ODE model can well explain the fundamental pattern under some reasonable assumptions.

After applying the ODE model to capture the fundamental trend of popularity, we still find an obvious difference between the estimated value and the true value. Therefore, we seek to further understand the impact of some fluctuating factors on the outcome to predict the number of reported results more precisely.

The time series of fluctuation is plotted in Figure 7. The series has distinct differences before and after the vertical dash line (May 15th, 2022), which is approximately where the ODE model curve and

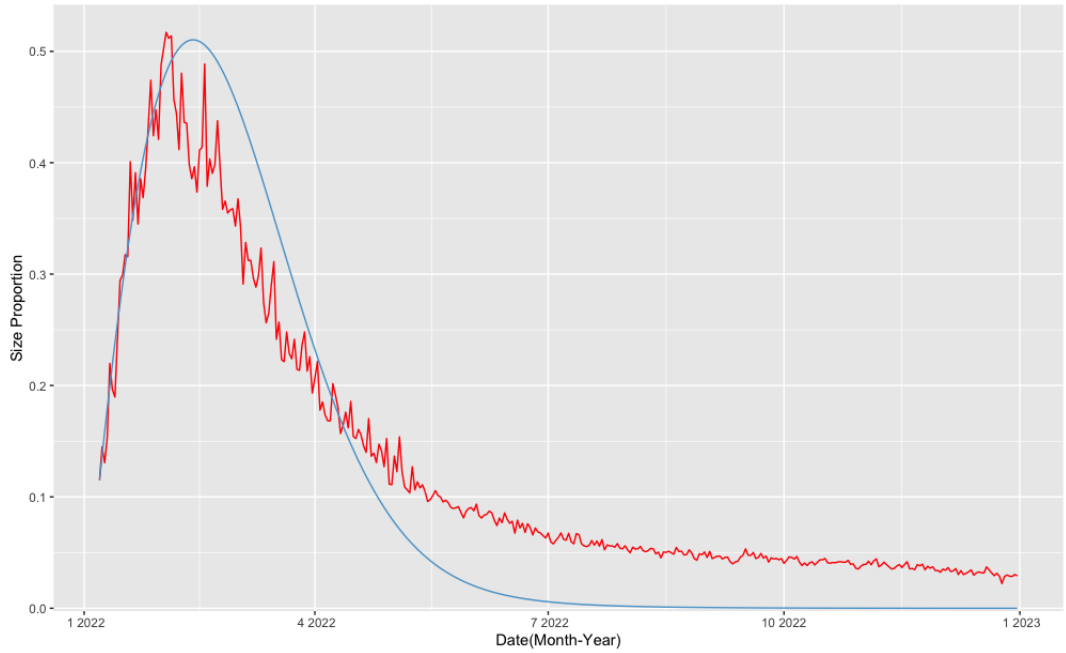

Figure 6: ODE Fitted Curve (Blue) and Actual Curve (Red)

actual value curve intersect. Consistent with the prediction of the ODE model and the trend that the actual value exhibits, it's more likely that the fluctuation will follow a pattern similar to the right-side pattern without any unexpected shocks. Therefore, our following prediction model only builds on the data after this cutoff.

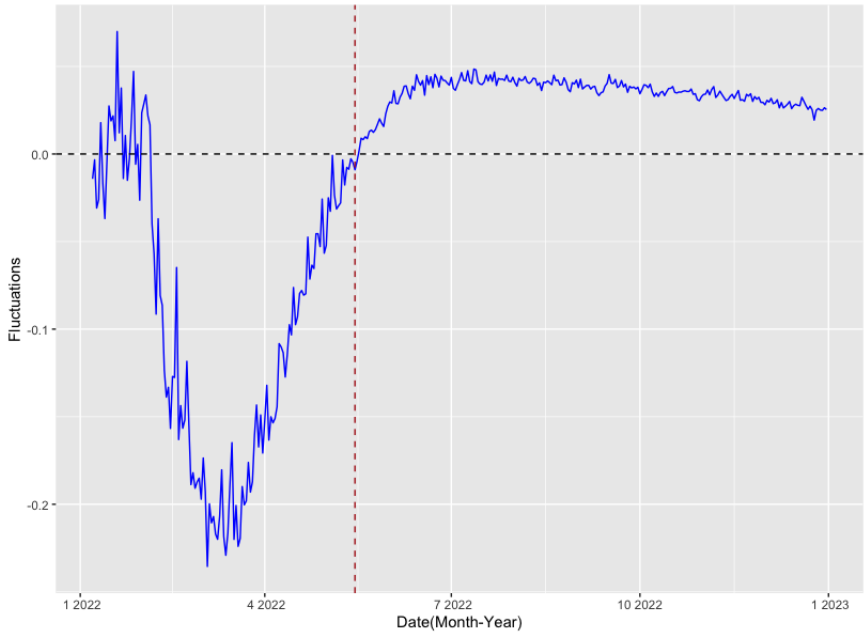

Figure 7: Fluctuation Time Series

Specifically, we adopt the Auto-regressive Integrated Moving Average (ARIMA) model to fit the up-and-down trend $f(t, ...)$. By conducting the Augmented Dickey-Fuller Test, the series is not stationary, so the parameter difference time is $d = 1$. By further checking the series auto-correlation (ACF) and partial auto-correlation (PACF), and comparing some models' AIC value, the optimal order of the model is $(1, 1, 4)$, and the model is as follows:

$$(1 - B)f_t = \phi_0 + \phi_1(1 - B)f_{t-1} + \varepsilon_t + \sum_{i=1}^{4} \theta_i \varepsilon_{t-i} \tag{9}$$

where B is the lag operator. By further conducting a Box-Ljung test on the residual of this model, the hypothesis of the residual ε_t being white noise cannot be rejected. Therefore, $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$ is a constant. The estimation result is reported in Table 5.

Table 5: Estimation Results of ARIMA for Fluctuation Series

Variable	**Estimate**	**Standard Error**
ϕ_1	0.9938^{***}	0.0088
θ_1	-1.5629^{***}	0.0673
θ_2	0.5166^{***}	0.1236
θ_3	-0.0546	0.1184
θ_4	0.1428^{**}	0.0627
σ^2	0.000000796	
Log-likelihood	1022.13	
AIC	−2034.36	

Notes. *** $p < 0.001$.

Shapiro-Wilk normality test on the residuals of the above model shows that the assumption of normality is valid. Therefore, we can further assert that $\varepsilon_t \sim N(0, \sigma^2)$. Based on the model, we can more precisely predict the reported number and its confidence interval on March 1st, 2023. The prediction for the future 60 days (i.e., until March 1st) is presented in Figure 8, in which the blue part of the curve represents predicted values, and the pink ribbon represents the confidence interval. To be more specific, the detailed prediction results are presented in Table 6.

Table 6: Detailed Prediction Results

Term	**Estimate**	**Lower Bound of CI**	**Upper Bound of CI**
Fundamental	75.11		
Fluctuation	11081.77	−35980.67	58144.21
Total	11156.88	−35980.67	58144.21

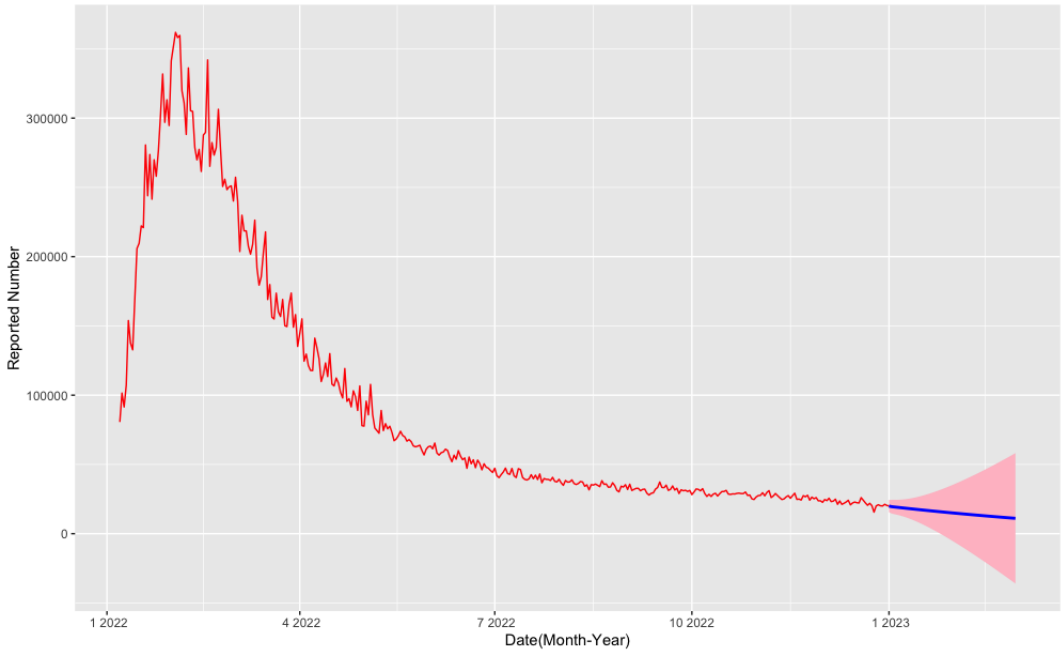

Figure 8: Prediction of Future Reported Numbers

4.2 Proportion of Hard Mode Players

To explore the pattern of the proportion of hard mode players, we utilize a data-driven time-series model to capture the potential time trends of the skill level and preference of players. Taking the effect of words' attributes into consideration, we adopt ARIMA with Explanatory Variables (ARIMAX) to fit and describe the data. By doing the Augmented Dickey-Fuller Test and checking ACF and PACF, we determine the ARIMAX model with order $(0, 1, 1)$, and the ultimate model is given by:

$$(1 - B)hp_t = (1 - \theta B)\varepsilon_t + \beta_1 \cdot Frequency_t + \beta_2 \cdot WordleScore_t + \beta_3 \cdot NumberRepeated_t \quad (10)$$

where B is the lag operator. By the result that residual series cannot reject the null hypothesis of the Box-Ljung test, we can assert $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$ for all t. The attributes of words that we included in the model are the ones that proved to be important in process of both PCA and clustering.

We report the estimation results in Table 7. By assessing the statistical significance of variables, the word frequency and the number of repeated letters are the most important attributes that impact the proportion of reported results with hard mode. The frequency of words has a negative effect on reporting proportion, while the number of repeated letters is the opposite. That is, the rarer or more strange the word, the higher propensity to choosing a hard mode and reporting it.

It's worth noting that although words' attributes significantly affect the percentage, their effects are

Table 7: Estimation Results of ARIMAX for Predicting Hard Mode Proportion

Variable	Estimate	Standard Error
θ	-0.6429^{***}	0.0402
β_1	-0.0003^{***}	0.0001
β_2	0.0002	0.0004
β_3	0.0014^{***}	0.0004
σ^2	0.00001424	
Log-likelihood	1489.28	
AIC	−2970.57	

Notes. *** $p < 0.001$.

much smaller than the time series trend, due to the small coefficients. Therefore, adopting a difficult or rare word cannot decisively increase this percentage that day. Another interesting finding during exploring this problem is that, if we run an OLS regression of hard mode percentage on the logarithmic total reports, the estimated coefficient is 1, not only significantly negative but also has quite large R^2 ($R^2 = 0.8724$). This indicates that with the decrease in the total reporting number, the hard mode reporting decreases more slowly, which means that the players reporting hard mode are more likely to be loyal players. Therefore, if the total reporting number keeps decreasing, this percentage will keep increasing.

5 Model of Reported Results Distribution

In this section, we seek to develop a model to predict the distribution of the reported results, i.e., the associated percentages of players with different times of tries for a future date. We refer to the multi-output regression that involves predicting two or more numerical values given an input example. It is worth noting that in multi-output regression, typically the outputs are dependent upon the input and upon each other, which means that often the outputs are not independent of each other and may require a model that predicts both outputs together or each output contingent upon the other outputs.

To this end, we apply the random forest (RF) regression to predict the distribution of the reported results. When training the model, we first set our extracted word attributes $Frequency$, $NumberRepeated$, $WordleScore$, and $AnyRare$ as the input and specify the number of trees as 100 with mean squared error as the criterion. Considering the nature of our time-series data, we also take time trends into account by regarding the date of the game as another input. We use 10-fold cross-validation to assess the accuracy of the model prediction, which offers an average mean absolute error of 3.6745 with a standard error of 0.3846 for the regression model with time input and an average mean absolute error of 3.7324 with a standard error of 0.3397 for the regression model without time

input. Hence, the random forest estimator can well explain and predict the variation in the distribution of the reported results.

The random forest regression can also provide us with an illustration of the relative importance of features when fitting the model. We visualize the relative importance of the model inputs in Figure 9, where the left panel presents the variables including time, and the right panel reports the variables excluding time. It can be seen that the time trend plays an important role in the random forest regression, followed by the word frequency, Wordle Score, and the number of repeated letters. The results are consistent with our previous PCA results and lend support to our classification of difficulty using the word attributes. However, the two models are similar in terms of the average absolute error, indicating that the addition of time may not improve the goodness of fit of the model and may result in biases in the prediction.

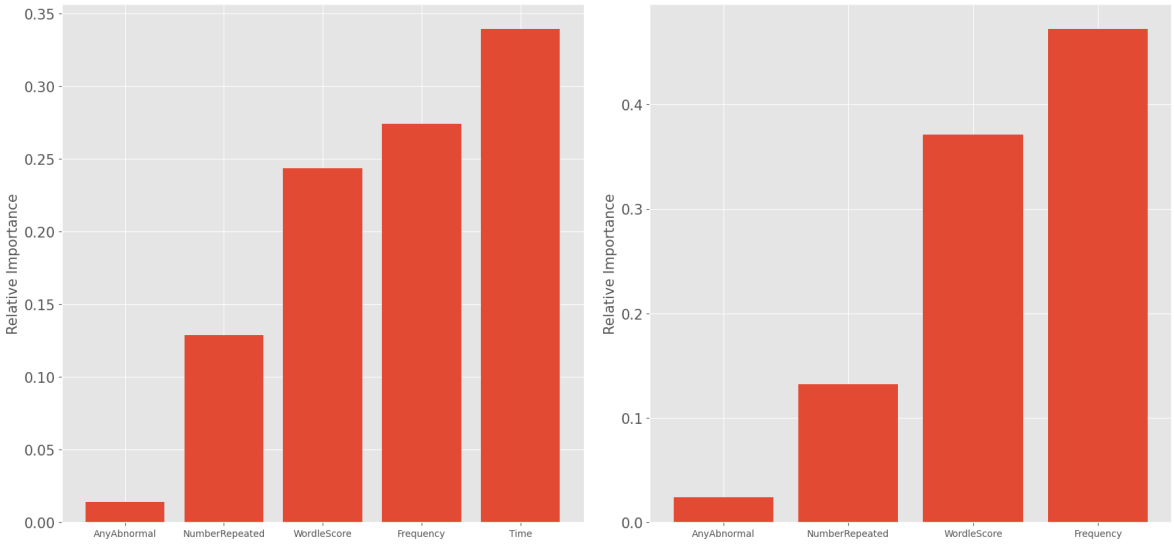

Figure 9: Relative Importance of Features

To apply our proposed model, we use our model to predict the distribution of the reported results for the word **EERIE** on March 1, 2023, reported in Figure 10. To derive the confidence intervals of the coefficients, we also re-estimate the model with 100 bootstrap samples and obtain the corresponding mean and standard deviation of the coefficient estimates. The point estimates together with 95% confidence intervals are presented in Figure 11. As shown in the figures, it is evident that excluding time as input would give rise to a higher average number of tries, which is more reasonable considering the difficulty of the word **EERIE**.

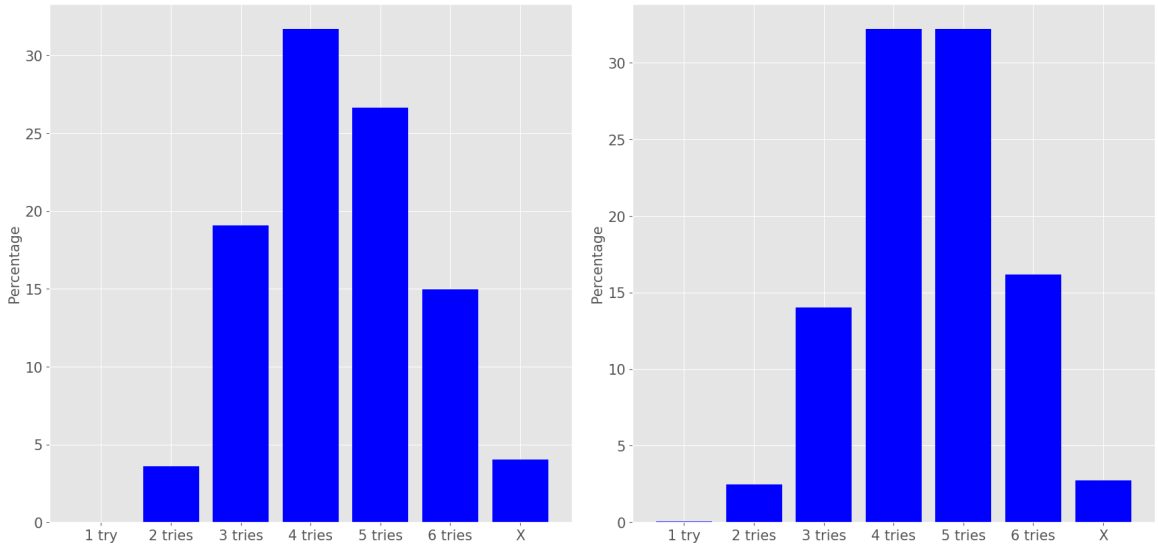

Figure 10: Relative Importance of Features Excluding and Including Time

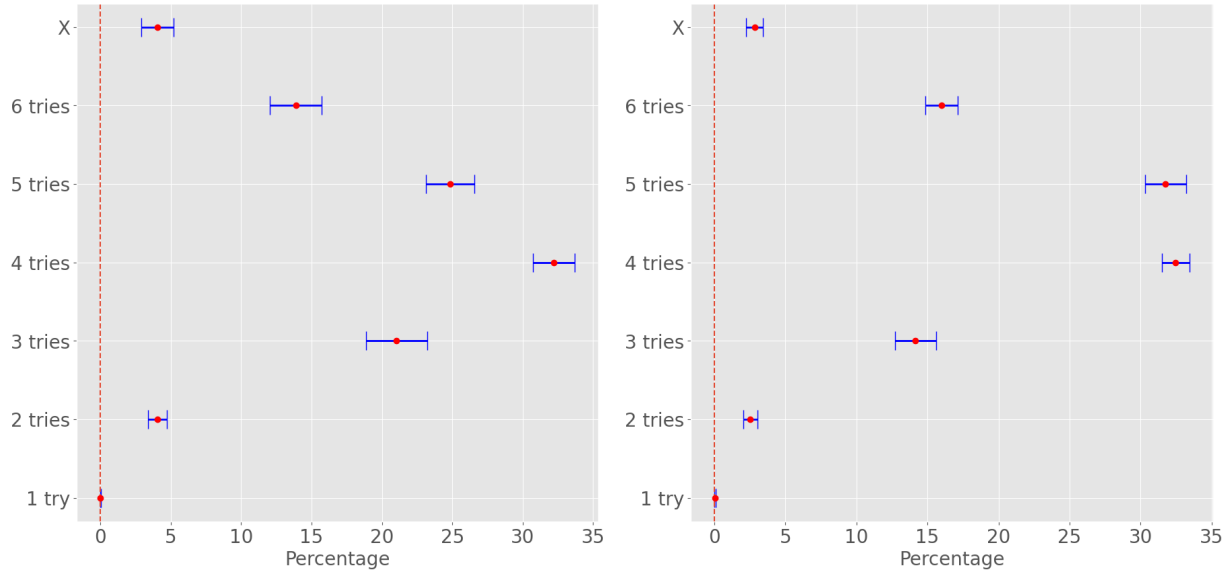

Figure 11: Confidence Intervals of the Prediction Estimates

6 Sensitivity Analysis

In the ODE model, the maximal potential market size is set to be exogenously given rather than an optimization outcome. Therefore, here we are going to fit two models with the other two given market size values, to see if the result will change a lot due to the change of the given condition.

In the former ODE model, we choose a market size of 700,000. To assess the robustness of our model, we choose 600,000 and 1,000,000, respectively. Here are two figures plotting the fitting results in Figure 12, with 600,000 in the left panel and 1,000,000 in the right panel. It can be observed that

neither of these two models fits better than the former one. Therefore, it's reasonable to set 800,000 as the maximal market size when fitting the ODE model.

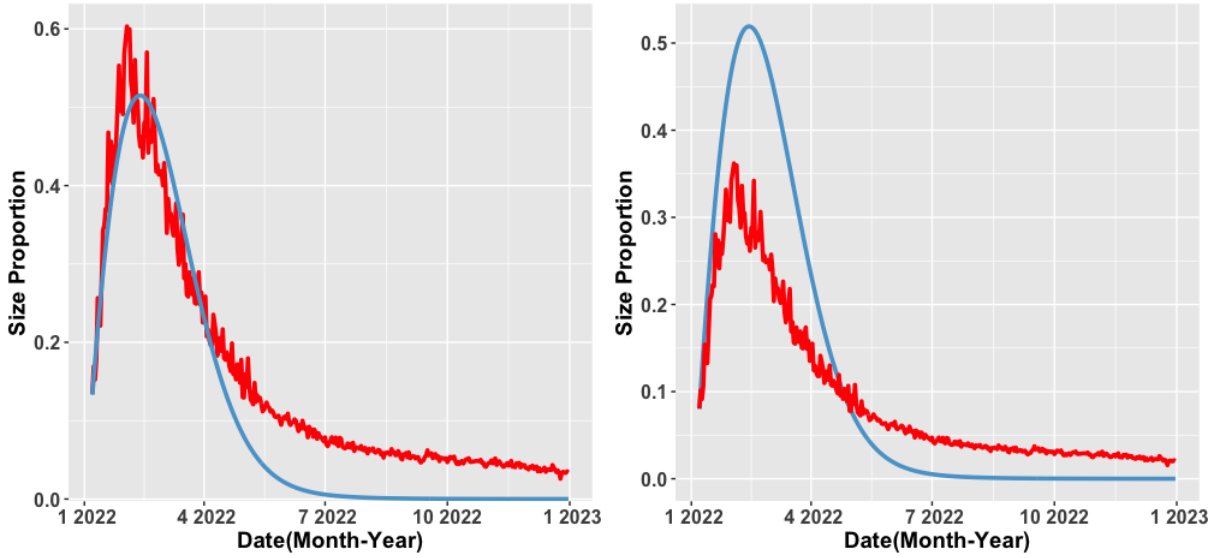

Figure 12: Different Maximal Market Sizes (Left: 600,000; Right: 1,000,000)

Meanwhile, for the ODE model itself, the market size actually only changes the equation's initial value. But when fixing other parameters, i.e., keeping p, q, μ constant, the results are insensitive to the initial value (see Figure 13). The difference between models with different market size assumptions mainly comes from the variation of true proportions when altering the maximal size.

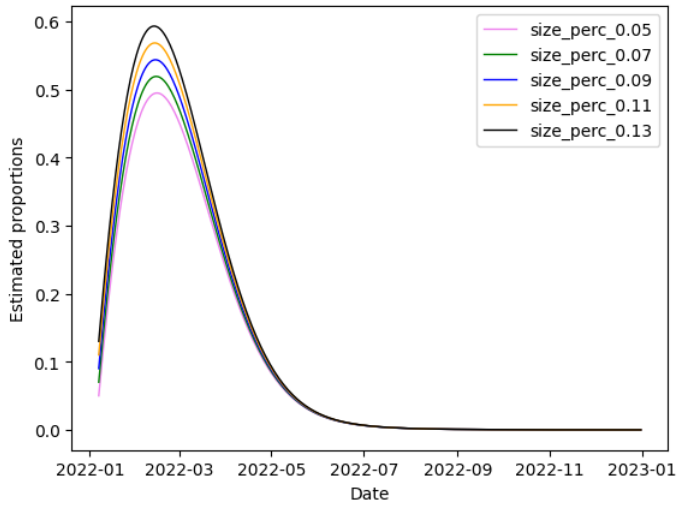

Figure 13: Robustness to Market Size

7 Model Evaluation

7.1 Strengths

1. **Novelty.** To better evaluate the words, we propose a new metric between different words based on the characteristics of Wordle. Meanwhile, we also propose an ODE theoretical model to describe the fundamental of Wordle popularity.

2. **Comprehensiveness.** In our model, we not only build our model on the data but try to get some insights and underlying patterns from the data.

3. **Generalization.** Our framework for words can be implemented to any sequence mainly of attention to positional and elemental information. Our framework for the performance of Wordle can be used to evaluate any similar internet products.

4. **Robustness.** Our models exhibit strong robustness to most of the parameters.

7.2 Weaknesses

1. **Theoretical model is not precise enough.** Our proposed theoretical model to describe the fundamental trend of reported results doesn't describe reality precisely.

2. **Some assumptions are too strict to be true.** Some of the assumptions to simplify the analysis are too hard to be satisfied, e.g., one player may enter or leave multiple times in one year.

3. **Missing other potentially relevant factors.** In this paper, all factors we adopted are given ones and characteristics of words. There may be some undiscovered information in those words.

8 Conclusion

To explore secrets behind the innovative mechanics and variation in popularity in the past year, we proposed a series of models and methods to both dig out the underlying rules of 5-letter words and their future prospects. These models successfully realize our aims, and here are our conclusions:

1. Wordle can be regarded as a game that guessing a proper sequence of letters, and the information can get from each wrong guess is positional and elemental focused, so we propose a novel metric "Wordle Score" to measure the similarity of 2 words. Further, by combining this distance and other features of words, we classify those words into 4 categories by K-means clustering, which gives a fair evaluation criterion for every word.

2. The number of reported results on Twitter can be seen as the popularity of Wordle. After looking into the literature, we employ an ODE model to theoretically construct the fundamental of popularity. And for the fluctuation term, we adopt an ARIMA model to guarantee a more precise prediction.

3. When analyzing the proportion of hard mode reports, we use an ARIMAX model to include both time series factors and word-related features. Through checking statistical significance, we find that word frequency and amount of repeated letters affect the percentage, the "harder" the word is, the higher proportion of hard mode.

4. We develop a model based on Random Forest regression to predict the distribution of the reported results, which can achieve a multi-dimensional data output simultaneously. This tree-based model realizes the prediction of a future day's percentages with a given word.

9 A Letter to the Puzzle Editor of the New York Times

Dear Puzzle Editor,

We are writing to share with you the findings of our recent study on the popular online puzzle game, Wordle, and its implications for game design. With the growing trend of players sharing their results on social media, the challenge for game developers is to leverage this data to improve the game's design, attract new users, and retain old ones. Our study aims to provide an initial exploration to address such problems by conducting a series of analyses on data collected from daily reported results from Twitter.

Our analyses focused on several attributes that may affect the difficulty of the solution word, including word frequency in the corpus, the number of repeated letters, and whether the word has less-used letters. We also developed a new measure of word difficulty by considering the distance between the solution word and the optimal starting word suggested by prior studies. Based on these attributes, we conducted Principal Component Analysis (PCA) and built a K-Means Clustering model to classify words into four difficulty levels, which showed high accuracy and generalizability.

We also developed models based on the product life cycle theory, Ordinary Differential Equations (ODE), and Auto-regressive Integrated Moving Average (ARIMA) to describe the pattern of reported results, which serves as a proxy for the popularity of the puzzle. Our analyses revealed that the number of reported results and word difficulty have significant impacts on the propensity to choose the hard mode, with players more likely to report their game outcomes and exhibit a larger proportion of hard-mode games with harder solution words.

Furthermore, we developed a model based on Random Forest (RF) Regression to predict the distribution of reported results, which reflects the performance of players given the solution word. Our proposed model can take the multi-output nature of the prediction task into account and renders an interval prediction of the percentage of each category of player tries.

Our models provide important implications for the design of online games, and we believe our analysis could be useful for the future design of the game. We would be delighted to discuss our research further with you, and we look forward to hearing from you.

Thank you for considering further communication between us.

Sincerely,

MCM Team # 2316012

References

Anderson BJ, Meyer JG (2022) Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. *arXiv preprint arXiv:2202.00557* .

Basu A, Garain A, Naskar SK (2019) Word difficulty prediction using convolutional neural networks. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 1109–1112 (IEEE).

Hiebert EH, Scott JA, Castaneda R, Spichtig A (2019) An analysis of the features of words that influence vocabulary difficulty. *Education Sciences* 9(1):8.

Kumar S, Swaminathan JM (2003) Diffusion of innovations under supply constraints. *Operations Research* 51(6):866–879.

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, volume 10, 707–710 (Soviet Union).

Li I (2022) Analyzing difficulty of wordle using linguistic characteristics to determine average success of twitter players .

Pearson K (1901) Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2(11):559–572.

Yang H, Suyong E (2018) Feature analysis on english word difficulty by gaussian mixture model. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 191–194 (IEEE).