Uncover the Hidden Secrets in Wordle Results

Since Wordle has become a popular puzzle game, it has accumulated a large amount of data. In this paper, we define a series of metrics and build several models to explore the hidden information in Wordle results.

First, after **preprocessing** the given data and analyzing the time series diagram of the number of reported results, we found that the changes can be divided into 3 stages. To forecast the number of reported results, we developed a weighted optimization model based on ARIMA and BP neural network. The prediction interval is then given using the Bootstrap method. We packaged this process as **ARIMA-BP Interval Prediction Model Based On Bootstrap**. Thus, we finally predicted the interval prediction value obtained on March 1, 2023 at 95% confidence level to be about (19504.74, 20383.26).

Then, we defined 3 qualitative and 4 quantitative attributes of words and used them to build a **Multiple Linear Regression Model** with the percentage of hard mode's players. We found that the proportion will decrease by an average of 0.618 when the initial letter changes from a vowel to a consonant while it will increase by an average of 0.017 for each one-unit increase in word internal distance.

After that, we made the percentage distribution prediction of the reported results based on **LSTM Model**. To ensure the percentage is around 100%, we first processed the component data using a **spherical coordinate transformation**. Then we use them as output variables, the 7 word attributes and number of results as input to train our LSTM model. The prediction of EERIE based on this are [2%, 11%, 25%, 24%, 19%, 14%, 5%]. We changed the model's parameters and added noise to do **sensitivity analysis**. Meanwhile, we introduced COV to measure the uncertainty of the model prediction, and found that it is around 0.4. For **error analysis**, we use MSE, RMSE and R^2 to measure the prediction accuracy, and their values are shown in Table 7.

We extracted 6 indicators: RDC, TE, SK, NFC, NON, and HL to measure the difficulty of words. We built a **GMM Clustering Model** based on these indicators and thus classifying 5 difficulty levels. We classified the word EERIE as difficulty level III.

In addition, by counting the frequency of each letter in five positions, we found S as the initial letter has the most frequency and more specific statistical results are shown in Table 9. We also used the **Association Rule Model** based on **Apriori algorithm** to mine the word combination pattern in Wordle. Ideally, we found that the letters A,S,E and F,T,L usually appear together in Wordle.

Finally, we evaluated and refined the model and reported the findings in a letter to the the Puzzle Editor of the New York Times.

**Keywords:** ARIMA-BP, LSTM, GMM, Apriori Algorithm, Word Attributes

Contents

1 Introduction

1.1 Background

Wordle is an online word puzzle game invented by Josh Wardle during the epidemic. The New York Times newspaper which is well known for the games it publishes has bought Wordle on February 2021[1]. Wordle only allows one game to be played per day, and every player in the world plays to guess the same five-letter word in six tries or less each day[2]. Players can play in regular mode or hard mode. They can share their scores via Twitter, thus attracting more people to play and share.

Figure 1: Wordle Game
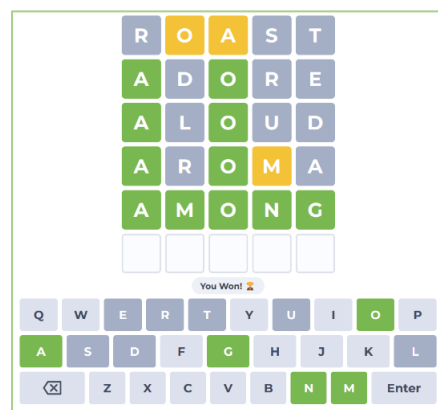
Back in October 2021, less than 5,000 visits registered to its web page while by January 2022, traffic had skyrocketed to over 45 million. Some of us also love this game, Figure 1 shows one result that we got. The green tile indicates that the secret solution word has the letter in the precise location. The yellow tile implies that the answer has the letter but not at the correct location. Grey tiles indicate the letters are not contained in the solution at all[3].

Now we have a file of daily results from January 7, 2022 to December 31, 2022. This file includes twelve key variables that are crucial in our later research. In the file, the percentages of the number of people for the seven tries may not sum to 100% due to rounding.

1.2 Problem Restatement

By analyzing the above background, we summarize the tasks that need to be addressed as follows：

- **Task 1:** Develop a model to explain the daily variation in the total number of people reporting scores on Twitter and use it to give a prediction interval for the total number of people on March 1, 2023.
- **Task 2:** Determine whether the attributes of words affect the percentage of players who choose the hard mode and explain the obtained results accordingly.
- **Task 3:** If given a future date and the specific word, build a model to predict the percentage of 1-X tries in this day. After that, the word EERIE on March 1, 2023 should be used as a specific example of the model prediction, while analyzing the uncertainties of the model and the accuracy of the prediction.
- **Task 4:** Develop a model to classify the difficulty of the words and identify the attributes of them under each category. Use this model to determine how difficult the word EERIE is? Finally, discuss the accuracy of the classification model.
- **Task 5:** List and describe some other interesting features of the dataset.
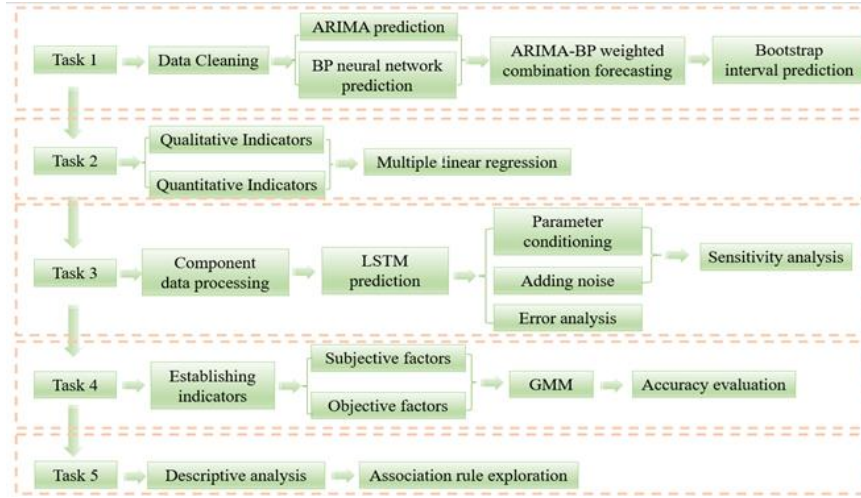
1.3 Our Work

Figure 2: Our work

2 Assumptions and Notations

Assumption 1: The number of daily online gamers in the data set is a time-dependent set of series and is independent of seasonal changes.

Reason: Therefore, we can use ARIMA as well as LSTM models to predict a series of data for March 1,2023.

Assumption 2: The scores reported by players on Twitter on a daily basis are normal and reliable.

Reason: To ensure that the model we build based on the dataset can reliably predict a range of data for March 1,2023.

Assumption 3: Assuming that Wordle's development process is consistent with the game's life cycle theory.

Reason: This assumption reduces the impact of external uncertainties on Wordle game predictions, thus making the whole process of prediction and analysis more efficient.

In this work, we use the symbols in Table 1 in the model construction. Other none-frequent-used symbols will be introduced once they are used.

Table 1: Notations

Symbols	Definition
w_A	Weights of the ARIMA model
w_B	Weights of the BP model
P_{Ai}	Predicted value of ARIMA model
P_{Bi}	Predicted value of BP model
L_i	Distance between letters
NF	Noise factor
C_i	the ratio of players at different guess counts

3 ARIMA-BP Interval Prediction Model Based On Bootstrap

3.1 Data Preprocessing

By reviewing the given data, we found that there are no missing values, but there are five outliers. One of them does not exist and two of them has less than 5 letters, we deleted these three rows of data due to the difficulty of obtaining the true values of these words. The two remaining outliers are caused by data entry errors. In order to avoid an excessive reduction in the amount of data, we changed them by combining the previous and later data as well as the semantics of them. In addition, the sum of the percentages in the original data are all in [98%, 102%], which is not much different from 100%, so they are reasonable and do not need to be processed. In summary, the preprocessing of the raw data is summarized in Table 2.

Table 2: Data Preprocessing

Outliers	Variable Type	Contest number	Preprocessing method
marxh	Word	473	Delete an entire row
clen	Word	525	Delete an entire row
tash	Word	314	Delete an entire row
2569	Number of reported results	529	Change to 22569
rprobe	Word	545	Change to probe

3.2 Point Prediction Based on a Combined ARIMA-BP Model

3.2.1 Variation Explaining

In the data preprocessing, although there are word entry errors, their corresponding numbers of reported results are not affected, so the complete reported data are analyzed in this problem. By observing and analyzing the characteristics of the changes in the number of reported results from Jan.7, 2022 to Dec 31, 2022, we found that they can be divided into 3 phases, as shown in Figure 3.

Based on the game lifecycle and player lifecycle theory[4], and combined with Wordle's game features, we explain the reasons for the changes as follows.

● Phase 1: Rapid Growth Period (January 7, 2022 - February 2, 2022)
Since the launch of Wordle's web version, it has been updated with only one puzzle a day. This artificial scarcity enhances players' desire for challenge and anticipation. The sharing function of Wordle uses emoji blocks to refer to the results of the game, which is very recognizable and easy to spread, while also avoiding spoilers, thus attracting more new players to play Wordle. In addition, the traditional popularity of crossword puzzles and the easy-to-play game mechanics have contributed to Wordle's further popularity. During this period, Wordle's overall growth in the number of results reported is 348.85%, with the number peaking at 361,908 on February 2.
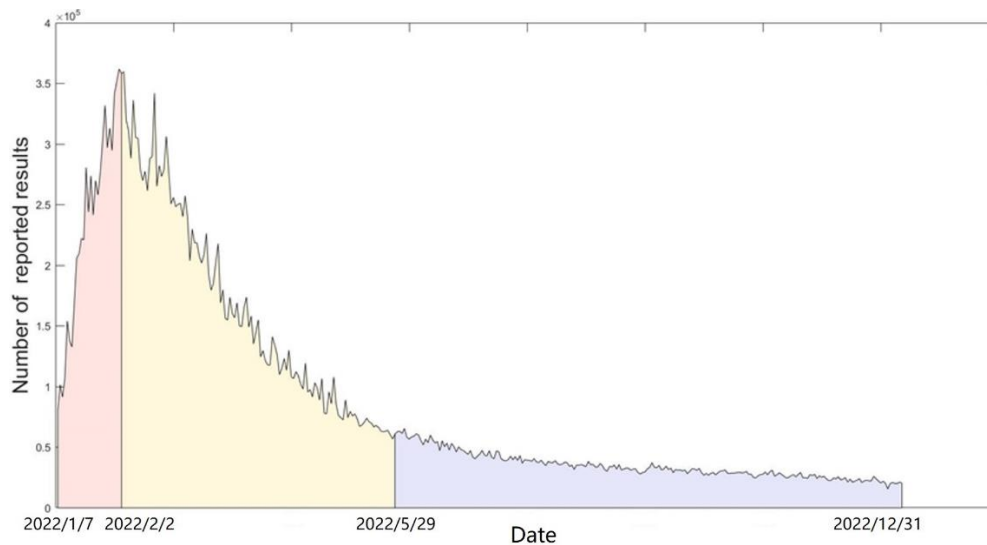
Figure 3: Changes in the number of reported results

●   Phase 2: Rapid Decline Period (February 3, 2022 - May 29, 2022)

During this period, Wordle experiences an overall 84.29% decline in the number of reported results, which is the general nature of internet fads. When the popularity of the game reaches its peak, it faces a significant loss of players due to gamers' declining interest and boredom with the game. The singularity of Wordle's game mechanic can contribute to this. Furthermore, the emergence of competing games in the market such as 'Words with Friends' and Wordle's pirated games makes Wordle lose players further.

●   Phase3: Stable Reduction Period (May 30, 2022 – December 31, 2022)

The overall decrease in the number of results reported for Wordle in this phase is 64.14%. Rapidly losing players in Phase 2 go to play Wordle generally because of its popularity among the general public and they will quickly leave when Wordle is no longer hot or the next trend emerges. The remaining players are often loyal Wordle players or players who have developed user stickiness due to the game's social circle. At this point, the number of reports will still drop but not by much.

3.2.2 Reasons for Model Selection

Predicting the number of reported results is a problem in the field of time series analysis. The traditional ARIMA model is capable of extracting deterministic information from historical data in order to predict the trend of variables over time, however, this model requires extensive testing before application, and the process of determining the order of the model is subjective to some extent [5]. In recent years, with the development of artificial intelligence, machine learning algorithms have been widely applied to the field of data classification and prediction. Among which, BP neural network model which is efficient and convenient can effectively break the limits of traditional time series prediction models. Considering these factors, we decide to combine the advantages of traditional time series models and machine learning models. It means that first of all, we use ARIMA model and BP neural network model to predict the number of reported outcomes separately, and then we use weighted average to combine the two models to obtain a more reliable ARIMA-BP combined prediction model.

3.2.3 ARIMA Forecasting

We first used the traditional ARIMA model for point prediction of the number of reported results, and this process is shown in Figure 4：
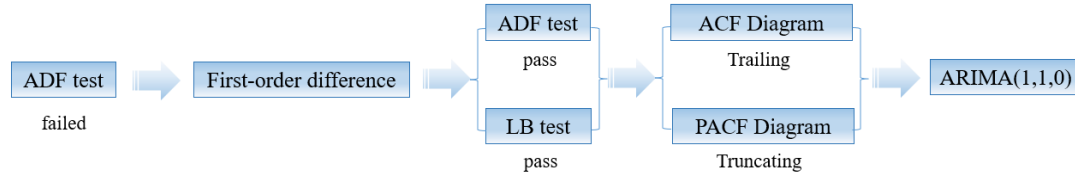

Figure 4: ARIMA model building process

First, we performed ADF smoothness test on the variable 'Number of reported results' in the preprocessed data set and found that it wasn't smooth. Therefore, we differenced the series to the first order and tested the differenced series again. After that, we observed the ACF and PACF plot of the differenced series, and found that the ACF plot showed the characteristics of trailing tails and the PACF plot showed the characteristics of 1st order truncated tails. Therefore, the following ARIMA(1,1,0) model can be developed for the original series:

$$\begin{cases} (1-\phi_1 B)(1-B)x_t = \varepsilon_t \\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s)=0, s \neq t\,, \\ E(x_s \varepsilon_t)=0\,, \forall s < t \end{cases}$$

where x_t is the tth time series value, B is the delay operator, ϕ_1 is the coefficient of the

moving self-averaging polynomial.

3.2.4 BP Neural Network Prediction

BP neural network is a multilayer feedforward neural network trained according to the error back propagation algorithm, which can iterate and repair its own weights continuously based on the relationship between input and output variables so as to finally estimate the exact functional relationship[6]. Its algorithm is explained by the following pseudo-code.：

Algorithm 1: Prediction of the Number of Reported Results
Input：Training set D
Output: test set E
Learning rate η; Data normalization (normalization formula)
Create network; Train the network
repeat for D
Forward propagation; Backward propagation
until for reaches end condition
Using the network; Inverse normalization of data
BP neural network prediction with test set E
end: The completed BP neural network trained

3.2.5 Analysis of Results

The result obtained using the ARIMA(1,1,0) model prediction is shown in Figure 5, and using the BP neural network model prediction is shown in Figure 6:
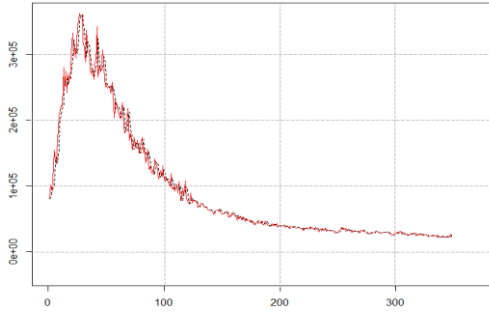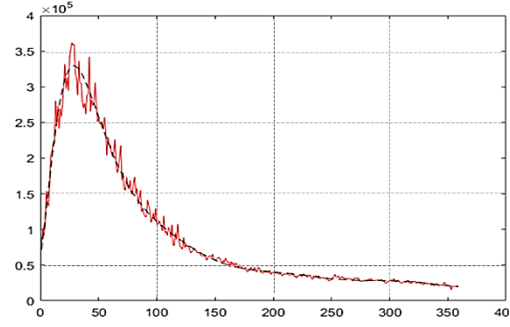

Figure 5: ARIMA prediction effect

Figure 6: BP neural network prediction effect

It can be seen that the ARIMA model has a better effect prediction, and the point prediction P_A obtained using this model for the number of results reported on March 1, 2023 is about 20598.84. BP neural network model has some bias in the early stage but has relatively good prediction effect in the later stage. The point prediction value P_B obtained using this model on March 1, 2023 is about 19496.05.

3.2.6 Weighted Portfolio Model Forecast

In order to improve the accuracy and robustness of the model prediction, we derived a new combined ARIMA-BP prediction model based on ARIMA and BP neural network model by using the prediction errors to weight the two models. The design idea of this weighted combined forecasting model is shown in Figure 7:
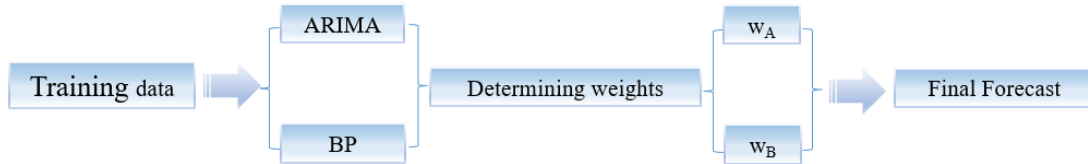
Figure 7: Combined model design

We first calculate the average prediction error for each of the two prediction models:

$$\overline{PEA^2} = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_{Ai})^2, \quad \overline{PEB^2} = \frac{1}{N}\sum_{i=1}^{N}(R_i - P_{Bi})^2.$$

After that, we use the share of the average prediction error squared of either model in the total average prediction error squared of the two models to reflect the weight accounted for by the other model, i.e.

$$w_A = \frac{\overline{PEB^2}}{\overline{PEA^2} + \overline{PEB^2}} \approx 0.406, \quad w_B = \frac{\overline{PEA^2}}{\overline{PEA^2} + \overline{PEB^2}} \approx 0.594.$$

Thus, the point prediction value of the ith series in the final ARIMA-BP combination model obtained is calculated as:

$$P_i = w_A P_{Ai} + w_B P_{Bi}. \tag{1}$$

Through the analysis of the prediction errors above, it is obvious that the overall prediction error of ARIMA model is relatively high, while the BP neural network model may be overfitted. Therefore, the combined ARIMA-BP prediction model can correct the errors of these two models, thus making the prediction results more realistic.
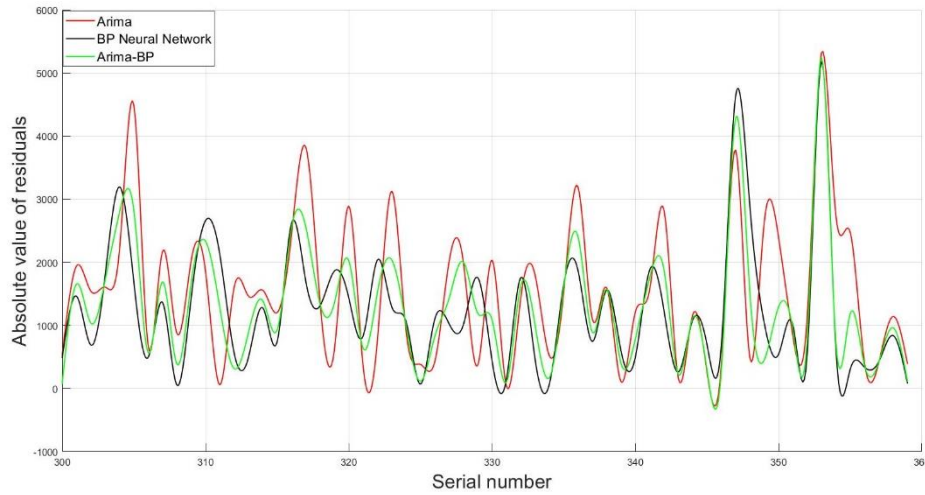
Figure 8: Prediction errors of three models

The final point forecast for the number of results reported on March 1, 2023 using this combined ARIMA-BP model is approximately 19944.

3.3 Interval Prediction Based on Bootstrap Method

The Bootstrap method is an important method used in statistics for interval prediction, which first assumes that the data set obeys an unknown distribution, and later estimates the distribution interval of the sample by repeatedly sampling the given data set [7].We took the data predicted under the weighted combination model for the next 60 periods as the sample set, after which we sampled it 1000 times with put-back to obtain the 1000 Bootstrap sample set，$Y^B=(Y_1，Y_2，\cdots\cdots，Y_{1000})$.

For each subsample, the distribution is in line with that of the sample, so we calculated the standard deviation (SD) of these 1000 subsample sets as the SD of the sample:

$$\sqrt{Var(y)}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(yi-\bar{y}\right)^2}\,.$$

According to the central limit theorem, it is known that the set of samples obtained by conducting 1000 draws approximately obeys the normal distribution. Therefore, the upper and lower limits of the confidence interval of the sample at confidence level 1-α can be calculated with the help of z-statistic：

$$y=\hat{y}\pm Z_{\alpha/2}\times\sqrt{Var(y)}\,. \tag{2}$$

Taking the confidence levels of 95%, 90% and 80% respectively, the final prediction intervals for the number of results reported on March 1, 2023 are shown in Table 3.

Table 3: Forecast interval

Confidence level	Lower limit	Upper limit
95%	19504.74	20383.26
90%	19575.33	20312.67
80%	19656.69	20231.31

From a statistical point of view, the number of people reporting Wordle scores on Twitter on March 1, 2023 will be approximately in the range of (19,504.74, 20,383.26), which will be a decrease compared to the current situation.

4 Exploring the Impact of Word Attributes on Hard Mode

Based on Multiple Linear Regression Model

Exploring the effects of certain attributes of words on the proportion of players choosing Hard Mode is a multiple-input variable, single-output variable modeling problem. We decided to develop a multiple linear regression model to derive whether word attributes affect the percentage of scores reported that were played in Hard Mode and the extent of their effects. The input variables X1-X7 are defined as follows.

4.1 Defining Qualitative Attributes of Words

First we made an overview of all the words in the dataset, and drew a word cloud map for the 356 words after preprocessing in Figure 9. We found that each word is different and its frequency is 1, so its size in the word cloud map is about the same.

Figure 9: A word cloud map for the 357 words

By searching for relevant literature, we classified the word attributes into 3 qualitative attributes and 4 quantitative attributes for a total of 7 word attributes.

- **Part of speech (X1)**

Part of speech is the most basic attribute of a word, the major classes are noun, verb and adjective[8]. By analyzing the parts of speeh of these 356 words, we divided the parts of speech of words into four categories: noun, verb, adjective and others.

- **Uncommon extent (X2)**

The extent of uncommonness is also one of the qualitative attributes of words. We used the COCA corpus to find the frequency of occurrence of each word in practical applications such as spoken speech, fiction books, magazines and newspapers, and academic journals. Using 3000 as a cut-off point, we divided the words into two categories: common and uncommon.

- **Whether the initial letter a vowel or a consonant letter (X3)**

Starting with a vowel or consonant letter is also a qualitative attribute of a word. The vowel letters are a, e, i, o, u, which may affect the rate of players who uses Hard Mode.

4.2 Defining Quantitative Attributes of Words

- **Number of vowel letters (X4)**

The number of vowel letters is one of the quantitative attributes of words. By counting that number for each word, we found that it takes values in [0, 3] (the value is an integer).

- **Maximum number of letter repetitions (X5)**

A word can be composed of different letters or may contain repeated letters, and the maximum number of letter repetitions can also be used as a quantitative property of a word. By counting that number for each word, we found that it takes values in the range [1,3] (the value is an integer).

- **Emotional polarity value (X6)**

Each word tends to imply a certain emotional attitude, and we used the SenticNet emotion dictionary to get the emotional polarity values for 356 words. In that, negative values represent negative emotions and positive values imply positive emotions.

- **Word internal distance (X7)**

We define the internal distance of a word (WID) as the sum of the distances of each adjacent letter in it (see Equation 3), and the letter distance is the distance between their positions (L) in the alphabet, e.g. A and C have a letter distance of 2 and ABBEY has an internal distance of 24.

$$WID_n = \sum_{i=1}^{4} |L_{i+1} - L_i|, \quad n = 1, 2, 3 \cdots 357 \tag{3}$$

4.3 Building a Multiple Linear Regression Model

Currently we have obtained data for 7 independent variables, but since 3 qualitative data: X1-X3, are categorical variables, we need to assign values to them before substituting them into the model. We assigned values to each category as follows.

$$X1 = \begin{cases} 1, noun \\ 2, verb \\ 3, adjective \\ 4, others \end{cases}, \quad X2 = \begin{cases} 1, commom \\ 2, uncommom \end{cases}, \quad X3 = \begin{cases} 1, vowel \\ 2, consonant \end{cases}$$

The dependent variable (Y) is the percentage of scores reported that were played in Hard Mode, which can be calculated according to equation (4).

$$Y = \frac{Number\ in\ hard\ mode}{Number\ of\ reported\ results} \times 100\% \tag{4}$$

Descriptive statistics for all variables identified are shown in Figure 10 and Table 4. As seen in (a), there are more uncommon words in the dataset than common ones, and more words with initial consonant letters than vowel letters; as seen in (b), there are the most nouns in the dataset, accounting for 53%, 61% of words with only one vowel letter,

and 29% of words with more than two maximum repetitions of letters.

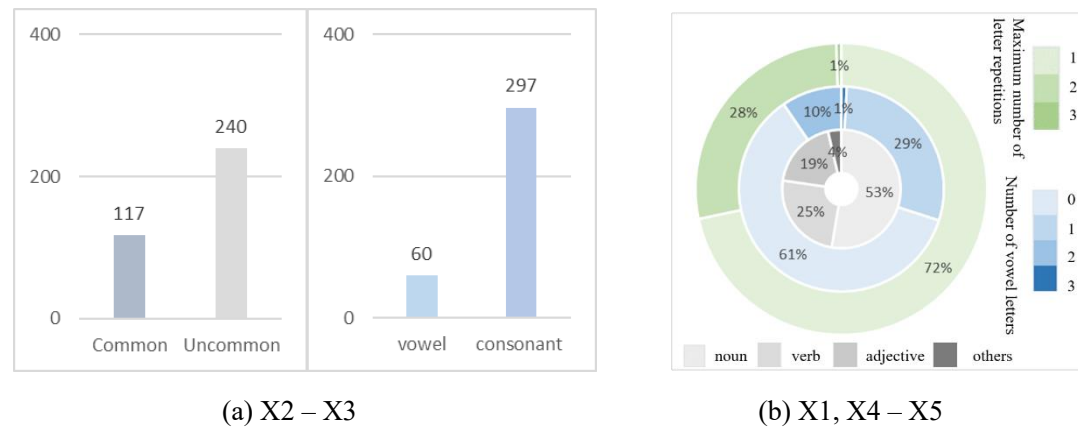(a) X2 – X3   (b) X1, X4 – X5

Figure 10: Descriptive statistics of variables

As seen in Table 4, the average value of emotional polarity value among the 356 words is 0.0355, which is positive; the average value of word internal distance is 33.3726; the average percentage of people reported to play hard mode is 7.52%

Table 4: Descriptive statistics of variables

Variable	Minimum	Maximum	Average	Standard deviation
X6	-1	0.999	0.0355	0.6448
X7	7	71	33.3726	11.9655
Y	1.17%	13.33%	7.52%	0.0223

Then, a multiple linear regression model can be developed as follows.

$$Y = \beta_0 + \sum_{i=1}^{7} \beta_i X_i + \varepsilon$$

In the model, β_0, β_i are the parameters of the model and ε is the error term.

Substituting the variable data, the final fitted regression model is:

$$\hat{Y} = 6.855 - 0.007X1 + 0.305X2 - 0.618X3 + 0.136X4$$
$$+ 0.394X5 - 0.025X6 + 0.017X7$$

4.4 Results Analysis

The results of the significance tests for the model as a whole and for the regression parameters of each independent variable are collated in Table 5.

Table 5: Model and parameter significance results

Variable	t-value (F-value)	P-value	Variable	t-value	P-value
Model	1.747	0.097	X4	0.662	0.509
X1	-0.053	0.958	X5	1.551	0.122
X2	1.210	0.227	X6	-0.138	0.890
X3	-1.798	0.073	X7	1.655	0.099

First, the P-value of the model as a whole is 0.097, which is significant at the 10% level

of significance, indicating that the model is significant. Second, among the 7 independent variables, only X3 and X7 have a P-value significant at the 10% level, which are 0.073 and 0.099 respectively, indicating that the regression coefficients of these two variables are significant. The p-values of the other variables are more than 0.1, so their regression coefficients are not significant.

Thus, the two word attributes of whether the initial letter of a word is a vowel or a consonant letter and word internal distance have a significant effect on the proportion of players who report playing the hard mode, and other five variables of part of speech, uncommon extent, number of vowel letters, maximum number of letter repetitions, and emotional polarity value do not have a significant effect on this rate. When other variables are held constant, the proportion will decrease by an average of 0.618 when the initial letter changes from a vowel to a consonant; the proportion will increase by an average of 0.017 for each one-unit increase in word internal distance.

5   Percentage Distribution Prediction Based on LSTM Model

5.1 Processing of Component Data

Currently, we know the percentage of players reporting scores on Twitter who guessed the word or were unable to solve the puzzle in 1 to 6 attempts per day, and these percentages add up to an approximate 100%.

1:	0%
2:	6%
3:	22%
4:	33%
5:	24%
6:	12%
X:	3%
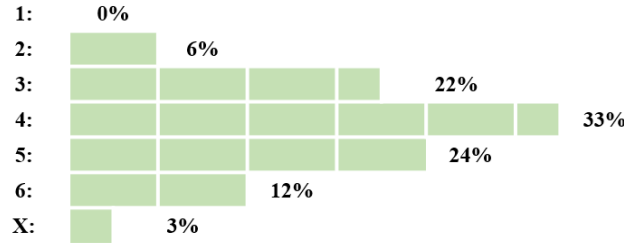
Figure 11: Percentage of successful attempts

It can be seen that these percentage data together form a set of component data $X=(x_1,x_2,\cdots\cdots,x_7)$, and that the values of the seven components in X approximately satisfy the constraint:

$$\sum_{i=1}^{7} x_i \approx 1,\ 0 \leqslant x_i < 1. \tag{5}$$

Therefore, the prediction of the percentage of future dates with respect to $(1,2,\cdots,6,X)$ can be considered as a prediction problem for the component data. In order to satisfy the restrictions of equation (5), instead of directly using the original data set as the output variable for prediction, we first performed a spherical coordinate transformation on the component data.

We open the root of each component x_i of the component data to obtain $y_i = \sqrt{x_i}$,

which gives us the new vector group $Y = (y_1, y_2, \cdots, y_7)$. Y can be viewed as a point on

the hypersphere, so we map Y into the hypersphere, i.e.

$$\begin{cases} \theta_7 = \arccos y_7 \\ \theta_6 = \arccos\left(\dfrac{y_6}{\sin\theta_7}\right) \\ \theta_5 = \arccos\left(\dfrac{y_5}{\sin\theta_7\sin\theta_6}\right) \\ \vdots \\ \theta_2 = \arccos\left(\dfrac{y_2}{\sin\theta_7\sin_6\cdots\sin\theta_3}\right) \end{cases}$$

After the above processing, we can use $\theta_2 - \theta_7$ as the output variable and use the model to predict it. After the prediction results are available, the final component data x_1-x_7 can be obtained by inverse transformation.

5.2 How to Predict

For the percentage prediction of the number of attempts by players in the specified future date, we view it as an information mapping problem between multiple inputs and multiple outputs. In this process, we consider the output variable as the percentage component data that we transformed using ball coordinates. As for the input variables, we consider two main aspects: the number of players and the properties of the words. From the previous analysis, we have found that the number of players changes over time, which may cause the percentage of attempts by players to change accordingly. On the other hand, the attributes of words will have a direct impact on the success rate of players' guesses and thus on the distribution of percentages. Therefore, in addition to the number of players per day, we include the seven words' attributes already defined earlier as input variables. Finally, for the selection of the prediction model, we chose the LSTM deep learning model, and we will describe the application of this model in detail below.

LSTM model is a long short-term memory network model that allows the analysis of input data using time series and predicts the future trend of these input data by using the core idea of cell state operation [9]. The working principle of this model can be represented in Figure 12, where yellow represents the arithmetic operations and blue represents the learned neural network.
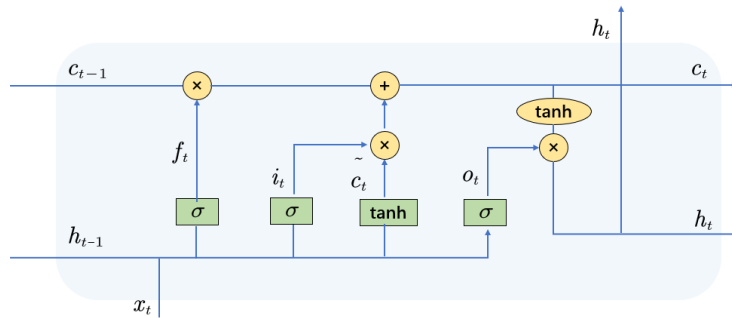

Figure 12: The working principle of LSTM

The LSTM model has the ability to remove or add information to each module by

setting up "gates", and in general it consists of four steps：

Step 1: Decide the discarded information in the cell: this operation first reads the current input x_t and the pre-neuron information h_{t-1}, after which the forgetting gate f_t determines the discarded information：

$$f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right)$$

Step 2: Determine the new information stored in the cell state: this operation uses the *sigmoid* layer as the input gate layer to determine the updated values and creates a new vector of candidate values via the *tanh* layer which are then added to the state：

$$i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right), \quad \tilde{c}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right)$$

Step 3: Update the cell state. This operation updates c_{t-1} to c_t and multiplies the old state with f_t, while discarding the information that needs to be discarded, after which the new candidate value is determined.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$$

Step 4: Determine the output: This operation uses the *sigmoid* layer to determine the portion of the cell state that is output, followed by processing the cell state through *tanh* and multiplying it with the output of the sigmoid layer by.

$$o_t = \sigma\left(W_0\left[h_{t-1}, x_t\right] + b_0\right), \quad h_t = o_t \cdot \tanh\left(c_t\right)$$

After that, we use the word EERIE as an example, and the number of players who report their scores on that day is predicted to be about 19944 using the ARIMA-BP weighted model, while the set of attribute vectors for this word is (3, 2, 4, 3, 1, 1, 26). Bringing these input attributes into our LSTM model, while inverse transforming the obtained results, we can get the final percentage distribution of the number of player attempts on March 1, 2023:

Table 6: Results

1 try	2 tries	3 tries	4 tries	5 tries	6 tries	X
2%	11%	25%	24%	19%	14%	5%

It can be seen that on March 1, 2023, for the word EERIE, about 13% of the players could guess the word in 2 trys, about 49% of the players could guess it in 3 to 4 guesses, while 5% of the players failed to guess the word in the required number of attempts.

5.3 Sensitivity Analysis

We explored the uncertainties of our model and predictions through sensitivity analysis, which was performed separately for the model using two methods: adjusting the model parameters and adding noise in the form of a normal distribution to the variables.

● **Adjusting the model parameters**

For parameter adjustment, we adjusted the batchsize and epoch parameters of the LSTM model five times, with specific adjustment values of [25, 250], [50, 500], [75, 750], [100, 1000], and [125,1500]. We considered the percentage of players with 7

attempt types per word as a prediction whole and used the coefficient of variation (COV) to measure the degree of dispersion between the 7 percentage values. The formula for COV is:

$$COV = \frac{s}{\bar{x}}$$

Based on the prediction idea of the LSTM model above, we re-predicted the data using the model after adjusting the parameters and calculated the COVs for the predicted data. If the overall distribution of COVs is similar as the model is adjusted, the uncertainty of the model is lower, and vice versa is higher. With the variation of batchsize (bs) and epoch (ep), the scatter plot of COVs of 356 words is shown in Figure 13.
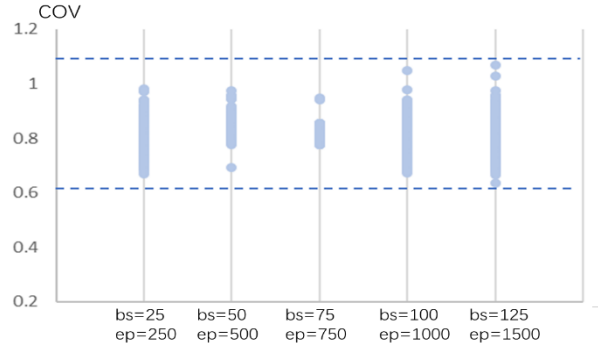

Figure 13: Sensitivity and uncertainty analysis (1)

As seen in Figure 13, the variations of the model parameters (bs, ep)have an effect on the distribution of the COVs of the predicted data. The distributions of COVs are similar under the model parameters except for bs=75, ep=750 where the distribution are more concentrated, indicating that the uncertainty of the model and prediction is low.

● **Adjusting noise**

For adding noise, we first randomly generated a normal distribution matrix $M_{356\times7}$, which was then multiplied by a noise coefficient NF and added to a matrix $T_{7\times356}$ consisting of the original data from one try to X tries to calculate the new output data set N of the model. The formula is expressed as follows. We adjusted the values of NF to 0.1, 0.15, 0.2, 0.25 and 0.3 respectively.

$$N = T_{356\times7} + NF \times M_{356\times7}$$

Similarly, we re-predicted the data using the model after adding noise and calculated the COVs for the predicted data. With the variation of NF, the scatter plot of COVs of 356 words is shown in Figure 14.
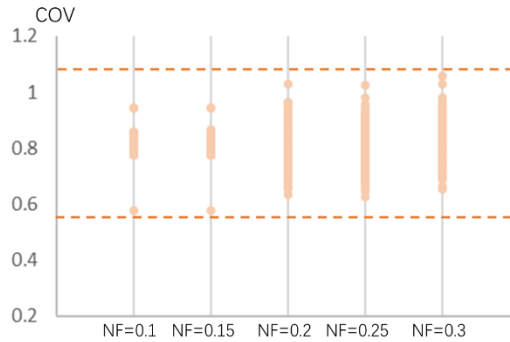

Figure 14: Sensitivity and uncertainty analysis (2)

As seen in Figure 14, the variations of the noise factor (NF) also have an effect on the distribution of the COVs of the predicted data. The distribution of COVs is more consistent across words when NF is 0.2 and above, which also indicates that the uncertainty of the model and prediction is lower at this time. We defined the overall extreme difference of the COV (Equation 6) to indicate the degree of uncertainty.

$$Uncertainty = \max(COV) - \min(COV) \tag{6}$$

From the variation of model parameters, the uncertainty of model and prediction is 0.4321; from the variation of noise coefficient, this uncertainty is 0.4794.

5.4 Error Analysis

We use error analysis to measure the confidence of the model prediction. The smaller the prediction error, the higher our confidence in the model we build. By reviewing the literature[10], we determined to use the mean absolute error (MAE), the root mean squared error (RMSE) and the coefficient of determination (R^2) to evaluate the effectiveness of the model prediction.

- **Mean absolute error (MAE)**

MAE is the weighted average of the absolute values of the differences between the predicted ($\hat{y}_n$) and true values (y_n) of the model. Mean absolute error does not consider the direction of the error, but only the average error magnitude of the predicted value. Therefore, this indicator can avoid the problem of mutual offset, and then can reflect the size of the actual prediction error more accurately. The range of MAE is [0, +∞), and the smaller the MAE is, the better the model fitting. The formula is as follows.

$$MAE = \frac{1}{N} \sum_{n=1}^{357} |y_n - \hat{y}_n|$$

- **Root mean squared error (RMSE)**

RMSE is the average root of the expected value of the square of the difference between the true value and the predicted value. It measures the deviation between the predicted and true values and is more intuitive in terms of order of magnitude. The range of the RMSE is [0, +∞), and the smaller the RMSE value, the better the model's predictive ability. The formula is as follows.

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{357} (y_n - \hat{y}_n)^2}$$

- **Coefficient of determination (R^2)**

R^2 is used to measure the goodness of fit of the model predictions to the actual values. The range of it is [0,1], a value close to 1 indicates that the model fits well; if the value is close to 0, the model is poorly fitted. The formula is as follows.

$$R^2 = 1 - \frac{\sum_{n=1}^{357} (y_n - \hat{y}_n)^2}{\sum_{n=1}^{357} (y_n - \overline{y})^2}$$

According to the formula of the above 3 indicators, we calculated the prediction error of the percentage of players making 1 try - X tries respectively, and then found the mean value of them. The calculation results are collated in Table 7.

Table 7: Error analysis

Indicators	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	X tries	Average
MAE	0.543	3.361	7.168	4.162	5.350	5.443	2.403	4.062
RMSE	0.911	4.414	8.868	5.423	6.590	6.865	4.302	5.339
R^2	0.899	0.986	0.993	0.989	0.986	0.993	0.998	0.978

As seen in Table 7, the MAE of all prediction results is within 8, with an average MAE of 4.062 and an average RMSE of 5.339, both of which have small values, indicating that the prediction error of the model is small. In addition, the overall R^2 of the model is 0.978, which indicates that the model fits extremely well and the model has good prediction effect. Therefore, we have confidence in our model to a large extent.

6 Word Difficulty Evaluation Based on GMM Model

6.1 Description of Indicators to Evaluate the Difficulty of Words

The difficulty of a word is influenced by its natural properties, and it can also be reflected by the number of attempts made by the player. Therefore, before classifying the difficulty of words, we first defined some indicators that can measure the difficulty of words in terms of both subjectivity (S) and objectivity (O).

● **Calculating the relative difficulty coefficient (S)**
The number of players of each attempt of type can be obtained from the number of results reported and the later percentage, and to reduce the effect of decreasing numbers in time series changes, we preprocessed the number of players (NP) using extreme difference normalization:

$$NP_{ij}' = \frac{NP_{ij} - NP_{\min}}{NP_{\max} - NP_{\min}}.$$

We define the score for one guess-X guesses as [1, 7], where a higher score indicates a more difficult word. The score is multiplied by the number of people to obtain the total score (TS) for each word under different guesses. Then the importance of different number of guesses to word difficulty was assigned based on AHP, and the weights W_i were obtained by the eigenvalue method as [0.0285, 0.0428, 0.0662, 0.1016, 0.1558, 0.2404, 0.3648] respectively. The consistency index C.R is 0.0176<0.1, indicating that the weights calculated based on the judgment matrix weights are valid. From this we define the relative difficulty coefficient (RDC) of words as follows.

$$RDC = W_i \times TS \tag{7}$$

● **Try expectations (S)**
We consider the ratio of players at different guess counts (C_i) as the probability of occurrence of that count (P_i), and define the try expectation (TE) which is also the average times each word is guessed as follows. Harder words have bigger TE.

$$TE = \sum_{i=1}^{7} P_i \times C_i \tag{8}$$

- **Skewness coefficient (S)**

The skewness coefficient (SK) can reflect the distribution of the data. For the number of attempts data, words are more difficult when SK is less than 0 while the data are left-skewed distributed, and vice versa. Its calculation formula is as follows.

$$SK = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

- **Word frequency coefficient (O)**

Based on the COCA corpus, we counted the word frequencies of more than 350 words in Wordle, and then we defined the word frequency coefficient of the word with the highest frequency as 1, the coefficient of the word with the lowest frequency as 0, and the coefficient of the word with the middle frequency as 0.5. In this way, all of the word frequencies can be mapped to the interval [0, 1]. The conversion equation for the word frequency coefficient in the lower interval is：

$$\underline{FC} = 0 + \frac{FRE - FRE_{\min}}{FRE_{mid} - FRE_{\min}} \times 0.5 .$$

The conversion equation for the word frequency coefficient in the upper interval is：

$$\overline{FC} = 0.5 + \frac{FRE - FRE_{mid}}{FRE_{\max} - FRE_{mid}} \times 0.5 .$$

- **Number of orthographic neighbors (O)**

Orthographic neighbors are defined as words which are the same length but differ by one letter (such as scare and stare)[11]. We collect this data of each word from "MCWord: An Orthographic Wordform Database". The more orthographic neighbors a word has, the more difficult it is for the player to pinpoint the word, and the more difficult it is.

- **Harmony level (O)**

We use the harmony level to indicate the degree of harmony between the reading and writing of a word from the perspective of the cognitive ability of Wordle players.
First, we use H to denote the ratio between the length of a word and the length of its phonetic symbols, i.e.

$$H = \frac{LW}{LT}$$

The closer the value of H is to 1, the better the harmony level of the word. Therefore, we can calculate the absolute value of the difference between H and 1. We define the harmony level of the word corresponding to the smallest absolute value as 1, and the harmony level of the word corresponding to the largest absolute value as 0. In this way, the harmony levels of all words can be also mapped to the interval [0, 1].

6.2 GMM Clustering Model Building

After specifying the indicators used to express the difficulty level of words, we chose a GMM model to cluster these indicator values to classify the difficulty level of words. Compared with the traditional K-means clustering, GMM clustering belongs to the hybrid clustering, which uses M' distance to measure the constraint relationship between different variables, and can be considered as a better clustering method.

The GMM algorithm assumes that the samples obey k mixed Gaussian distribution functions, and for each sample point, the probability that it belongs to the kth multivariate Gaussian distribution can be calculated as follows.

$$p(x^{(i)}|z^{(i)} = j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)} - \mu_j)\sum^{-1}(x^{(i)} - \mu_j)^T}$$

After that, the parameters of each multivariate Gaussian distribution function can be updated by the EM algorithm：

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \quad \sum j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^T}{\sum_{i=1}^{m} w_j^{(i)}} \, .$$

Based on the above process, the iterative loop is continued until the Gaussian parameters converge and the loop is stopped. After that, we can classify the samples into categories by the calculated Gaussian parameters.

6.3 Analysis of Clustering Results

The clustering results based on the GMM model are shown in Table 8. Since the evaluation criteria of relative difficulty coefficient (RDC) is more comprehensive among these six indicators, which is a weight score calculated by refining the proportional characteristics of players under different guesses. It is more decisive for the difficulty of words than other indicators. Therefore, we mainly divided the difficulty level by the size of RDC of each category, and the division results are as follows.

Table 8: Clustering Results

Difficulty Level	TE	SK	RDC	WFC	NON	HL
I	4.1548	0.6371	0.4158	0.351	0	0.3324
II	4.1920	0.4875	0.4345	0.3684	3.7589	0.3705
III	4.1866	0.5483	0.4656	0.3571	1.3984	0.3496
IIII	4.0888	0.1631	0.5786	0.4379	11.3750	0.4167
IIIII	4.2992	0.2261	0.5941	0.3782	6.8958	0.3524

The words in difficulty level I are the easiest compared to the words in other categories, with the lowest RDC in this category; the highest skewness coefficient (SK) and a right-skewed distribution, indicating that a higher percentage of players guessed the words in this category correctly with fewer attempts; the number of orthographic neighbors (NON) is 0, indicating that players are able to locate the words easily.

The ROC of words in difficulty level IIIII is 0.5941, the highest; the try expectation is also the highest, and the average number of times players guessed the words in this

category is 4.2992. The specific attribute values of the remaining difficulty categories are shown in Table 8.

Based on the definitions and formulas of the 6 indicators, the 6 attribute values of EERIE were obtained as shown in Figure 15. By calculating the Marxist distance between the 6 indicator values of the word and the corresponding indicators of the 5 difficulty levels classified, we found that the distance between the word and difficulty level III is the closest, so we believe that EERIE belongs to difficulty level III.
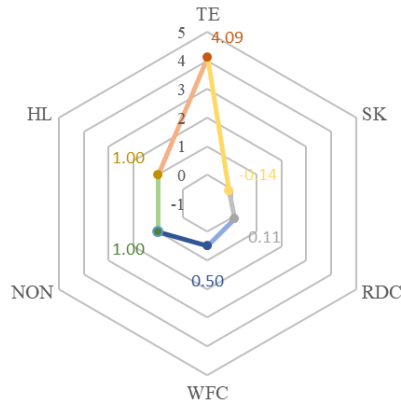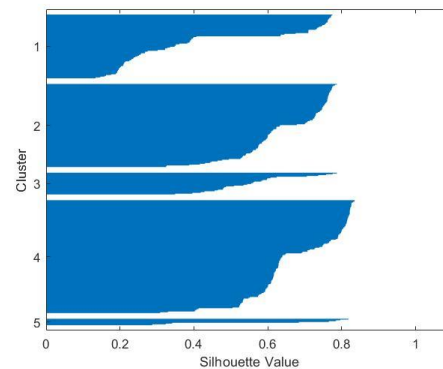

Figure 15: six attribute values of EERIE

Figure 16: Accuracy of the model

As is shown in Figure 16, the Silhouette Value can reflect the accuracy of GMM model clustering. The closer the Silhouette Value is to 1, the higher the accuracy of the model has. According to the results, our Silhouette Values of the five categories aggregated into are all in the range of 0.8 attachment, which shows that the clustering effect we obtained using the GMM model has a high degree of accuracy.

7   Other interesting features of the dataset to explore

7.1 Word Frequency Statistics for 5 Letter Positions

We counted the frequency of letters in 5 letter positions for each of the 356 words, and Table 9 demonstrates the letters with the top 8 frequencies in 5 letter positions and their frequencies. We found that 's' as the initial letter has the most frequently, 'o' is most frequent as the 2nd and 3rd letter, and 'e' is most frequent as the 4th and 5th letter.

Table 9: Letters with the top 8 frequencies in 5 letter positions

First Letter	Frequency	Second Letter	Frequency	Third Letter	Frequency	Fourth Letter	Frequency	Fifth Letter	Frequency
s	51	o	51	o	50	e	43	e	73
c	32	a	47	i	44	a	30	y	49
t	29	l	41	a	42	r	29	t	46
a	28	r	36	e	29	t	27	r	35
p	23	h	33	u	28	n	26	l	24
f	22	e	28	n	23	l	24	d	20
b	20	i	23	r	19	i	23	h	18
m	19	n	17	p	14	c	19	k	16

7.2 Exploration of the Combination Pattern Between Words

Generally speaking, the letters that make up a word are not independent of each other,

but have certain back-and-forth associations. These letters with back-and-forth associations can form a pool of resources from which players can select letters for guessing to help improve the correct rate. Therefore, from this perspective, we aim to explore the combination pattern between individual words.

To reduce the complexity of the discussion, we assume that the order of word occurrences is not considered, while only the combination pattern of the 255 words that do not appear with repeated letters is explored. We solve this problem with the help of the Association Rule Model, while using the following three definitions to determine the strength of the association between letters.

Sup: it indicates the probability that the item set X appears in the whole data N, i.e.：

$$Sup\left(X\right) = \frac{sum\left(X\right)}{N}$$

Conf: It represents the probability that one variable X changes conditional on the other variable Y also changing, i.e.

$$Conf(X \to Y) = \frac{Sup\left(X \cup Y\right)}{Sup\left(X\right)}$$

Lift: It can reflect the degree of elevation of the probability of change of one variable X on the change of another variable Y. If its absolute value is greater than 1, it means that there is indeed a correlation between these two objects.

$$Lift\left(X \to Y\right) = \frac{Sup\left(XY\right)}{Sup\left(X\right)Sup\left(Y\right)} = \frac{Conf(X \to Y)}{Sup\left(Y\right)}$$

In exploring the combination laws between words using the association rule model, we first perform letter splitting of words and convert the split words into the form of Bool values of 0 or 1. After that, we used the efficient and streamlined Apriori Algorithm for the analysis of association rules, specifying **Sup** of 0.5, and filter out the association laws and confidence values between letters when the **Lift** is greater than 1.

Table 10: Word combination pattern

Letters that appear	Letters that may appear together	Conf	Letters that appear	Letters that may appear together	Conf
a / k / s	e	1	e / n / r	i	1
l / p / t	a	1	p / r / t	e	1
b / e / n	i	1	g / h / t	i	1
c / r / t	e	1	i / m / s	t	1
d / p / t	e	1	d / f	e / t	0.6
h / i / s	e	1	f / t	l / o	0.5

Using the Association Rule Model we can get a lot of laws between letter combinations, but due to pages limitation we only show 12 possible cases, according to which players can spell specific words more efficiently to some extent.

8   Model Evaluation and Improvement

8.1 Strengths and Weaknesses

- In making the prediction of the number of people on March 1 2023, we used two models, ARIMA model as well as BP Neural Network, and performed a weighted calculation based on the sum of squared residuals of both, which compensated to some extent for the shortcomings of both algorithms.
- We first transformed the original data with the spherical coordinates when using the LSTM model before, so that the sum of the predicted values of the percentages about (1, 2, 3, 4, 5, 6, X) in the future dates is more close to 100%.
- When considering the uncertainties of the LSTM model, we added the disturbance of appropriate noise to perturb the model, and later calculated their coefficients of variation, thus making our prediction results more credible.
- When considering the characteristics of the data set, we used the Apriori algorithm, and found that the occurrence of some letters among the words given daily has some correlation by mining the features between the data.
- In our exploration, we found that the total number of players in the dataset reported daily was decreasing, but the percentage of those who chose the difficult mode was increasing when we were limited by space to give a literal description of it and did not explore it in depth.

8.2 Model Improvements

- In the subsequent study, when clustering the data using the GMM mixture model, we can take into account the importance of each indicator in depth, so that we can add certain weights to each indicator.
- We only used the Association Rule Algorithm to find correlations for some letter combinations, but in reality, the composition of words is much more complex than we thought (such as different positions of each letter), so we should also start to consider building a more comprehensive model to describe them.

9 Conclusion

In this article, we employed different models to analyze the information of Wordle results. In general, we have achieved the following conclusions：

- Compared to 2022, the number of players reporting results will approximately be in the range of (19,504.74, 20,383.26) at a 95% confidence level.
- Whether the initial letter of a word is a vowel or a consonant letter and word internal distance will affect the proportion of players who report playing the hard mode, and the proportion will decrease by an average of 0.618 when the former change from a vowel to a consonant while increasing by an average of 0.017 for each one-unit increase in word internal distance.
- On 2023/3/1, for the word EERIE, about 49% of the players could guess it in 3 to 4 guesses, and the uncertainty of this result is about 0.4 and at the same time the average prediction error is about 5.339. In addition, we found that EERIE belongs to difficulty level III, which is moderately difficult.

10 Letter to the Puzzle Editor of the New York Times

Feburary 20th, 2023

Dear Sir / Madam,

We are writing this letter to you because we want to report to you our basic findings on Wordle game data from January 31, 2022 to December 31, 2022. We hope that these results will be useful to you.

First, we found that the changes of the number of reported results on Twitter can be roughly divided into 3 phases: Rapid Growth Period (Jan.7, 2022-Feb.2, 2022), Rapid Decline Period (Feb.3, 2022-May 29, 2022) and Stable Reduction Period (May 30, 2022-Dec.31, 2022). By building a weighted ARIMA-BP model, we predict that the number of reported results on March 1, 2023 will be 19,944, and with the help of the Bootstrap method, we obtain a 95% prediction interval of (19,504.74, 20,383.26).

Second, we defined the word attributes of solution words in Wordle and divided them into 3 qualitative attributes and 4 quantitative attributes, and by building a multiple linear regression model for these 7 attribute variables and the percentage of players in Hard Mode, we found whether the initial letter of a word is a vowel or a consonant letter and word internal distance have a significant effect on the proportion.

Then, based on these 7 defined word attributes and the number of reported results, we established their relationship with the result distribution by LSTM model and predicted the distribution for the word EERIE on March 1, 2023 as [2%, 11%, 25%, 24%, 19%, 14%, 5%], which corresponded to a number of guesses ranging from one to more than seven. This prediction method can also apply to other solution words.

Again, based on the data of the number of guesses made by players for each word and the attributes of words, we defined three subjective indicators: relative difficulty coefficient, try expectations, skewness coefficient, and three objective ones: word frequency coefficient, number of orthographic neighbors, and harmony level to jointly measure the difficulty of words. Using these 6 indicators, we classify the words in Wordle into 5 difficulty levels by GMM clustering model. Taking the word EERIE as an example, we classified it into Difficulty level III according to its data for 6 indicators.

Finally, we counted the frequency of each letter in the 5 positions of the word in the dataset and found that the solution words tend to have 's' as the initial letter, the highest frequency of the 2nd and 3rd positions is the letter 'o', and the highest frequency of the 4th and last positions is 'e'. In addition, we have used the algorithm of association rules to explore the inter-letter connections in Wordle puzzles, and obtained the association rule that when 'a', 'k', and 's' appear at the same time, 'e' tends to appear as well.

Thanks for taking your time out of your busy schedule to read the letter. We hope this can help you manage and operate Wordle better.

Sincerely,
Team#2309397

References

[1] Match, S. (2022). The New York Times buys Wordle. The New York Times.

[2] Anderson, B. J., & Meyer, J. G. (2022). Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. arXiv preprint arXiv:2202.00557.

[3] Short, M. B. (2022). Winning Wordle Wisely—or How to Ruin a Fun Little Internet Game with Math. The Mathematical Intelligencer, 44(3), 227-237.

[4] Kummer, L. B. M., Nievola, J. C., & Paraiso, E. C. (2017). Digital Game Usage Lifecycle: a systematic literature review. In Brazilian Symposium on Computer Games and Digital Entertainment (SBGames) (pp. 1163-1172).

[5] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.

[6] Buscema, M. (1998). Back propagation neural networks. Substance use & misuse, 33(2), 233-270.

[7] Masarotto G. Bootstrap prediction in tervals for auto regressions [J]. International Journal of Forecasting. 1990,6(2),229-239.

[8] Kornfilt, J. (2020). Parts of speech, lexical categories, and word classes in morphology. In Oxford Research Encyclopedia of Linguistics.

[9] S. Siami-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 3285-3292.

[10] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623.

[11] Li, I. (2022). Analyzing difficulty of Wordle using linguistic characteristics to determine average success of Twitter players.