Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2306479

The Wordle Game Analysis Model

Summary

Wordle is a popular puzzle that the *New York Times* currently provides daily, and the distribution of its results has also attracted many people to explore. Using the report results from *Twitter*, our team established a related model to explore the distribution of Wordle results.

First, we established a **Gaussian mixture model (GMM)** to characterize the quantity change of the reported results. At the same time, in order to measure the volatility of the statistical data, we selected the parameter estimation results under the **95% confidence interval** as the parameter range of the Gaussian mixture model. The number of reported results for March 1, 2023 obtained by our model is **(2001,37156)**. We selected four attributes: whether the word has a prefix or suffix, the information content of the word, the probability of the word, and the probability of the letter in the word to study the impact of the word attribute on the distribution of the result percentage under the hard model.It was found that whether a word has a prefix or suffix has little effect on the percentage distribution of the result under the hard model, and the other three attributes have a larger impact on the results.

Next, we found that the distribution of the report results of each day is close to the **normal distribution**, so we use the Gaussian function to characterize the distribution of the report results of each day.At the same time, we found that the coefficient and variance of the Gaussian function have a corresponding relationship of double exponential function.So,we use the three attributes screened in the first question as independent variables,and the mean and variance of the Gaussian distribution corresponding to the report result as the dependent variable. **A particle swarm optimization algorithm based on simulated annealing(SAPSO)** is used to establish a game outcome prediction model(GOPM).But we found that the model does not include the impact of the uncertainty of the game itself on the reported results.Therefore, we use a **long-short-term memory neural network (LSTM)** to compensate for uncertainty to obtain a corrected game outcome prediction model (CGOPM).Compared with the GOPM model, the accuracy of CGOPM has increased from **52.24%** to **54.49%**.The prediction result of the word EERIE using our model is [1 try,2 tries,3 tries,4 tries,5 tries,6 tries,X]=**[6,17,28,29,16,5,1]**.

After that, we selected the mean and variance of the Gaussian function corresponding to the daily report results obtained above as the classification basis, and established a word difficulty classification model (WDCM) using **K-means**.The accuracy of the classification model is verified by calculating the silhouette coefficient and other indicators.Using **canonical correlation analysis** to study the classification results and word attributes, the results show that the word information has the greatest impact on the classification results.We combined CGOPM and WDCM to divide the difficulty of EERIE words, and the difficulty of this word is **easy**.

Finally, we summarize some interesting properties of the dataset, such as the small effect of time on the distribution of game outcome percentages, the different distribution of pronunciation of words, and the fact that repeated letters of words affect players' guesses.And we summarized our team's analysis of the results of the Wordle game and presented it to the puzzle editor of the *New York Times* in written form.

Keywords: Text Mining; GMM; Simulated Annealing; LSTM; K-Means

Contents

1 Introduction

1.1 Problem Background

A new charade game called Wordle is coming out in 2022.The rule of the game is that you can only guess words with five letters that have practical meaning. After submitting a word, the color of the square changes. If it is green, it means that the correct word contains the letter and that the letter is in the correct place; If it is yellow, it means that the correct word contains this letter but the letter is not in the correct place; If it is gray, it means that the correct word does not contain this letter .

At the same time, normal mode and hard mode are enabled, and the hard mode requirement is that the letter of the next guess must contain the green or yellow color displayed in the previous guess. The game eventually attracted thousands of people in the United States.

1.2 Restatement of the Problem

The ***New York Times*** needs you to analyze the game based on some data provided on ***Twitter*** to solve the following problems

- Develop a model to explain the variation in the number of reports, and use the model to create a prediction interval for the number of reported results on March 1, 2023. Imagery the effect of word attributes on the percentage of scores reported that were played in Hard Mode.
- Establish the percentage of attempts that the model predicts for a future day, and describes what uncertainties the model has. Explain your predictions for the word EERIE in 2023 and indicate how confident you are to believe your predictions.
- Build a model and classify words by difficulty. Identify the attributes of a given word that are associated with each classification. Use your model to illustrate how difficult EERIE is and discuss accuracy.
- List other interesting features of the dataset, and end up with a page or two of analysis for the ***New York Times***.

1.3 Literature Review

This question focuses on the conclusion related to mining data about Wordle. Wordle, as a word game, is closely related to text analysis. Text data mining refers to the extraction of high-quality information from texts, which is usually obtained by designing patterns and trends through statistical pattern learning and other means. Literature review is divided into four main parts. Stochastic model about text; Extraction of information entropy about text; Long Short Term Memory;Particle swarm optimization algorithm.

- First of all, there are many methods for stochastic models, representative of which are Bayesian random theory proposed by Bayes and Gaussian normal distribution theory proposed by Moivre. This can be done for random, non-quantifiable physical quantities. But in this problem, the use of the Gaussian model lies in the distribution of the difficulty of the problem.

- Secondly, for the extraction of text information, in paper[4], it is proposed to count the frequency of occurrence of each character, and extract character information accordingly.
- Thirdly,The paper[7] explores the search space of LSTM model structures and proposes a method called LSTM search space for automatically designing the structure of LSTM models. The method searches different LSTM model structures in order to find the optimal model structure.
- Finally, The paper[7] analyzes the convergence and stability problems of the PSO algorithm and proposes an improved PSO algorithm, called GBest PSO, which improves the convergence speed and stability of the algorithm by introducing the global optimal solution.
- The current methods and theories related to this topic will be clearly displayed in the figure below:

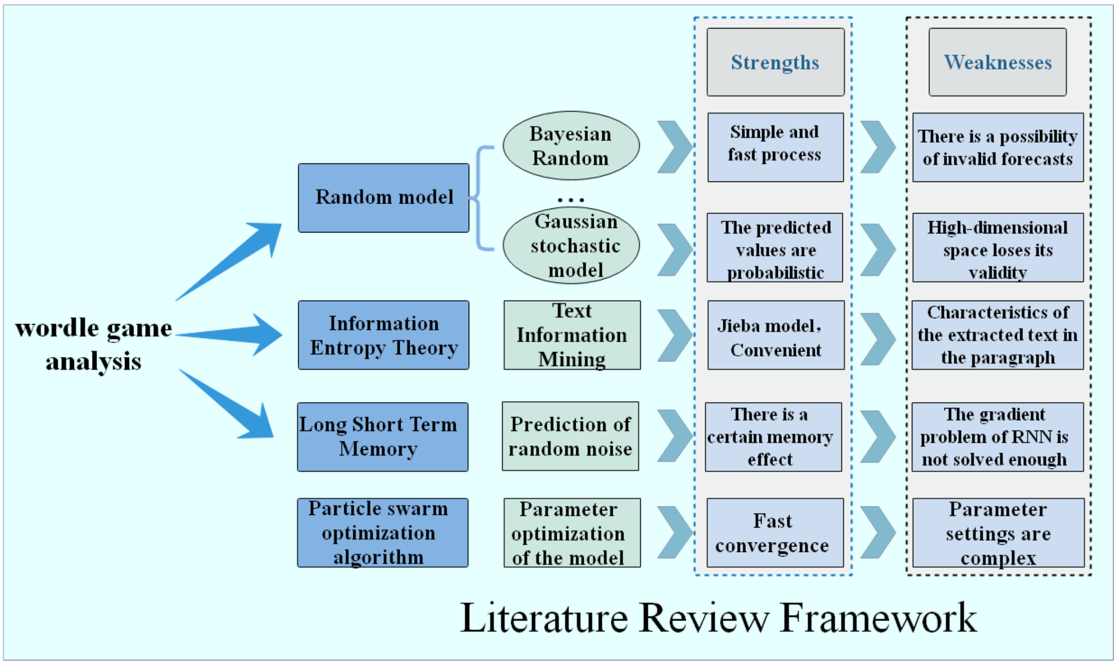

Strengths
Weaknesses
wordle game analysis
Random model
Bayesian Random
...
Gaussian stochastic model
Simple and fast process
There is a possibility of invalid forecasts
The predicted values are probabilistic
High-dimensional space loses its validity
Information Entropy Theory
Text Information Mining
Jieba model, Convenient
Characteristics of the extracted text in the paragraph
Long Short Term Memory
Prediction of random noise
There is a certain memory effect
The gradient problem of RNN is not solved enough
Particle swarm optimization algorithm
Parameter optimization of the model
Fast convergence
Parameter settings are complex
Literature Review Framework

Figure 1: Literature Review Framework

1.4 Our Work

This problem requires us to predict Wordle report data based on Twitter statistics and to classify words. Our work consists mainly of the following

1. The number of reports is predicted by building a Gaussian mixture model.
2. Four attributes were selected: the presence or absence of prefixes or suffixes, the information content of the word, the probability of occurrence of the word, and the probability of occurrence of the initial letter in the word, and their impact on the distribution of reported results was studied.

3. The results reported daily were normalized, and a game outcome prediction model was established by using a particle swarm optimization algorithm based on simulated annealing.
4. The LSTM is used to quantitatively compensate for the uncertainty in the prediction and to build a modified game outcome prediction model.
5. K-means was used to classify the difficulty of words. Through typical correlation analysis, the influence of the three attributes of the amount of information, the probability of word occurrence and the probability of the occurrence of the first letter of the word on the classification results is studied.

In order to avoid complicated description, intuitively reflect our work process, the flow chart is shown in Figure 2:

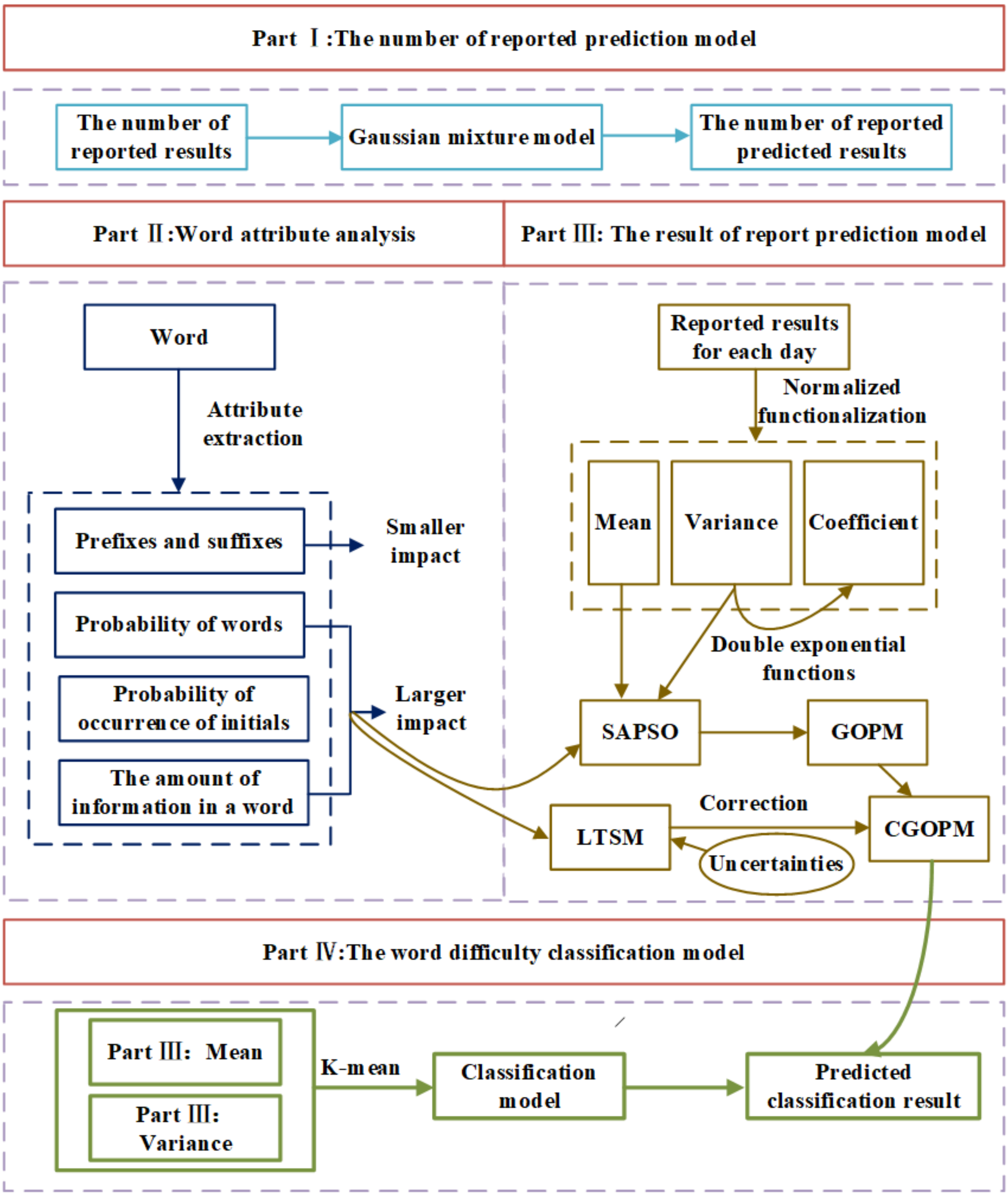

Part Ⅰ:The number of reported prediction model
The number of reported results
Gaussian mixture model
The number of reported predicted results
Part Ⅱ:Word attribute analysis
Part Ⅲ: The result of report prediction model
Word
Attribute extraction
Prefixes and suffixes
Smaller impact
Probability of words
Probability of occurrence of initials
The amount of information in a word
Larger impact
Reported results for each day
Normalized functionalization
Mean
Variance
Coefficient
Double exponential functions
SAPSO
GOPM
LTSM
Correction
CGOPM
Uncertainties
Part Ⅳ:The word difficulty classification model
Part Ⅲ: Mean
Part Ⅲ: Variance
K-mean
Classification model
Predicted classification result

Figure 2: Flow Chart of Our Work

2 Assumptions and Explanations

Considering that there are many uncertainties in the actual problem, for this analysis about wordle games, we need to make some reasonable assumptions to simplify the problem and also to make it more reasonable. Each assumption is immediately followed by a sufficient explanation

- **Assumption 1:** The impact of incidents on the number of daily reports is not considered.
  Explanation: Emergent events are not statistically significant and have no meaning for information extraction from datasets.
- **Assumption 2:** When extracting information about words and characters, frequencies are used instead of probabilities.
  *Explanation:*Because of the large sample size of the database we are using, statistically speaking, probability is approximately equal to frequency.
- **Assumption 3:** We believe that the distribution of Twitter′s daily reported numbers is the actual distribution of game results for that day.
  *Explanation:*Due to the large number of reports on twitter every day, statistically speaking, the percentage of the number of players in one attempt, two, three... is approximately equal to the actual data of the day.

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

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
u_i	The mean of the Gaussian fit for the number of game attempts
σ_i^2	The variance of the Gaussian fit for the number of game attempts
Q_{word_i}	The probability of the word
Q_{first_i}	The probability of the initial letter of theword
I_i	The amount of information for the word

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

4 Model Preparation

4.1 Data pre-processing

Inspection of the data table shows that three words are not of length 5. So there are errors in the words and corrections need to be made.As shown in the table below

Table 2: Abnormal data filtering results

Contest number	Word	Length	Amendments
545	rprobe	6	probe
525	clen	4	clean
314	tash	4	trash

5 Model 1:Report Quantity Gaussian Mixture Model

5.1 Gaussian mixture model building

Each time corresponds to the data submitted by the game platform, and the amount of data fluctuates over time. A cursory analysis of the data shows that the data shows a significant increase until January 22, 2022. After this date, the volume of data decreases significantly, showing a a sharp decline and then a gentle decline. Considering that most events in nature fit a Gaussian distribution, a Gaussian mixture model model can be built based on the trend of the data.

From January 7, 2022 to December 31, 2022, once expressed as an integer of 202-560.

$$\hat{y} = a_1 * e^{-\left(\frac{x-b_1}{c_1}\right)^2} + a_2 * e^{-\left(\frac{x-b_2}{c_2}\right)^2} + a_3 * e^{-\left(\frac{x-b_3}{c_3}\right)^2} \tag{1}$$

in the equation 1 , x represents the contest number, $x\epsilon\left(202, 560\right)$. y is the predicted value of the number of reported results.

In order to make the predicted values more objectively close to the true values, the least squares criterion can therefore be used to ensure that the sum of squares of the residuals of the total number reported is minimized.

$$\min\left(res\right) = \min\left(\hat{y} - y\right)^2 \tag{2}$$

in the above formula,res is residuals.The Lagrange multiplier method is used to ensure that the sum of squares of the residuals of the reported volume forecasts is minimized.

$$Y = res + \lambda * G\left(x, \hat{y}\right) \tag{3}$$

In equation 3, $G\left(x, \hat{y}\right) = 0$, according to the Lagrange multiplier method requires the introduction of the linkage vector λ.also, to ensure that the minimum value is reached, let the partial derivative of the constructor with respect to $x, \hat{y}$ be 0

$$\frac{\partial Y}{\partial \hat{y}} = 0 \tag{4}$$

$$\frac{\partial Y}{\partial x} = 0 \tag{5}$$

To illustrate the prediction accuracy of the regression prediction model, the confidence intervals and uncertainty of the estimation parameter are also given. In the experiments of this paper, the mean and variance of the estimation parameter are unknown. Therefore, to calculate the confidence interval of the mean estimation parameter, the following equation is defined

$$ConfidenceInterval = \left[\bar{X} - t_{\frac{\alpha_c}{2}}\left(N-1\right)\frac{S_\sigma}{\sqrt{N}}, \bar{X} + t_{\frac{\alpha_c}{2}}\left(N-1\right)\frac{S_\sigma}{\sqrt{N}}\right] \tag{6}$$

According to the above formula$\bar{X} = \sum_{i=1}^{N} X_i$ is the mean value of the number of reported results estimation parameter.α_c is confidence,$\alpha_c = 0.95$,N is is the number of samples. $N = 359$.The following formula is the variance of estimation parameter.

$$S_\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2} \tag{7}$$

5.2 The predicted result of the number of reports

According to the data given in Model 1 and the problem, the parameter values of the regression prediction function can be calculated by the least squares algorithm.

$$\begin{cases} a1 = 192300\,(114800, 269900)\,; b1 = 228\,(226.9, 229)\,; c1 = 22.24\,(18.12, 26.37) \\ a2 = 141700\,(90580, 192700)\,; b2 = 258.9\,(241, 276.8)\,; c2 = 48.09\,(32.59, 63.58) \\ a3 = 49450\,(37660, 61230)\,; b3 = 327\,(222.6, 431.4)\,; c3 = 216.2\,(123.8, 308.7) \end{cases} \tag{8}$$

The values in parentheses in the results of the above parameters represent the upper and lower bounds of their estimated results. The fitting function image and prediction results are shown in Figure 3

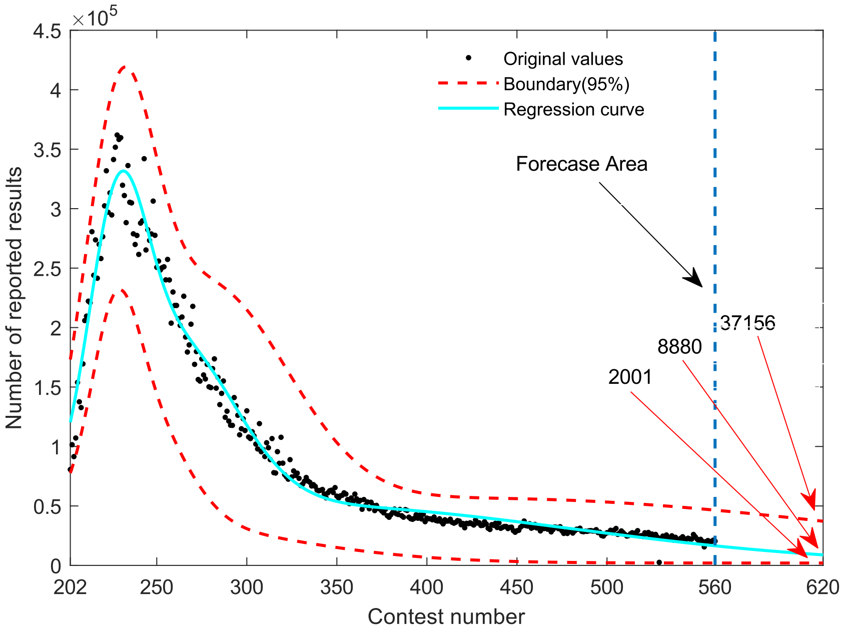


Figure 3: Reports quantity forecast results

According to the Figure 3 , the number of reported results for March 1, 2023 can be predicted to be 8880.The prediction interval is $(2001, 37156)$

The reason for this change is that when the game was first launched, people were more curious, so it was hotter and the number of submissions skyrocketed. As time goes by, users gradually lose their novelty about the game, so it will show a significant decline and finally stabilize. The people who steadily submit games to try are the loyal fans of this game.

5.3 Attribute extraction of words

5.3.1 Analysis of prefixes and suffixes

For the words themselves, the meaning of the words has greater meaning. But when it comes to guessing words, the guesser doesn't pay particular attention to the meaning of the word. Thus, the meaning of the word had little to do with the results of the final guessing game report.

In order to analyze the effect of the attributes of the words in the hard mod on the percentage of reported results, we calculated the percentage of the total number of people in the report who chose the hard mode as a percentage of the total game reports. Based on the mean value of percentage, the range with a fluctuation of 30% above or below the mean value is selected for problem analysis.

$$Interval = \left[(1-30\%)\frac{1}{n}\sum_{i=1}^{t}\frac{hard}{all}, (1+30\%)\frac{1}{n}\sum_{i=1}^{t}\frac{hard}{all} \right] \tag{9}$$

In Equation 9, $"hard"$ is number in hard mode. $"all"$ is total number of daily reports This method ensures that the reported results of the selected analysis samples are equally affected by the difficulty pattern, thus eliminating the influence of outliers. The visualization results are as follows.

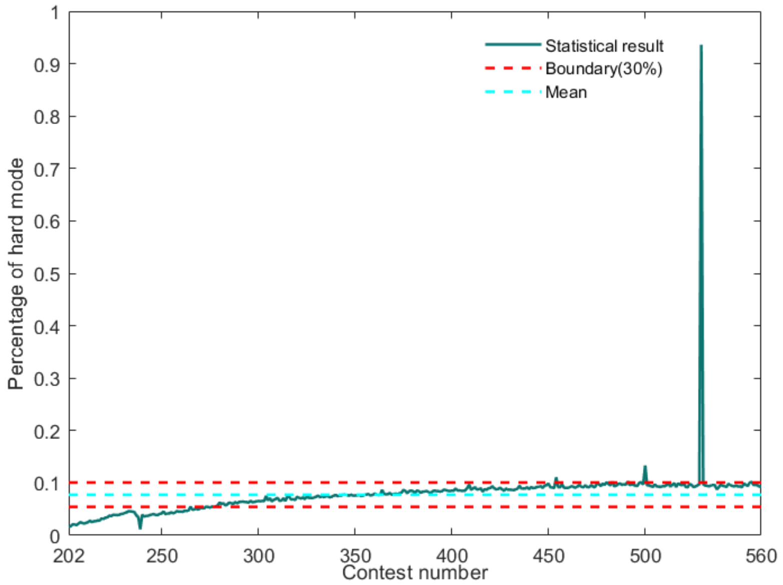


Figure 4: Schematic diagram of sample selection for analysis

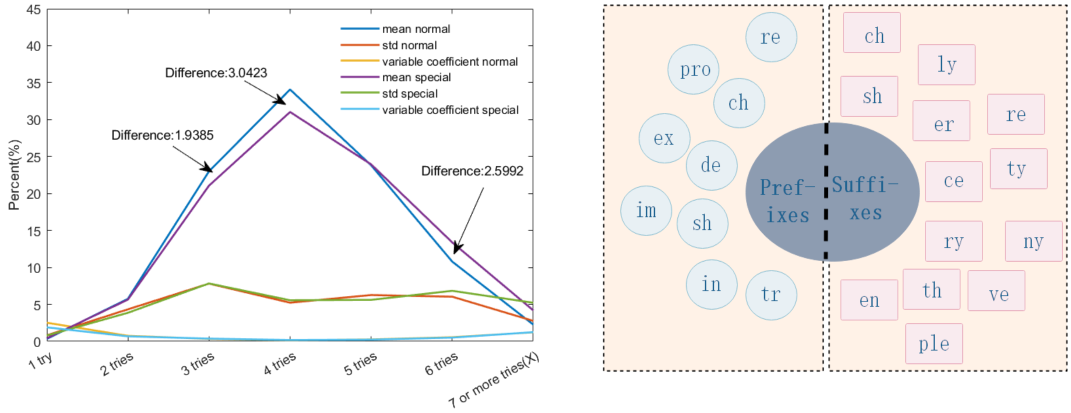
The selection of data for analysis can be clearly seen in Figure 4,According to the above chart, we can exclude the data outside the proposed range of 202-264,266-273, 454,500,529. so that the existence of similarity in the sample can be ensured.

According to the selected samples, we made statistics on the prefixes and suffixes involved in the words in the samples, and the results are shown in Figure 5(b). We statistically analyzed the characteristics of the reported result distribution for words with and without a suffix, including the mean, variance and coefficient of variation of different attempts.Finally, the curves related to the three statistics with and without prefixes are obtained separately. The statistical comparison results of three features of words with and without prefix and suffix are shown in Figure 5(a)

The coefficient of variation is calculated as follows

$$CV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}{\frac{1}{n}\sum_{i=1}^{n}x_i} \tag{10}$$

Where x_i is the sample value, which is the percentage distribution of all attempts per word per day about the report. CV is coefficient of variation.



(a) Report sample statistical analysis chart (b) Prefix and suffix diagram

Figure 5: Prefix and Suffix Analysis

According to the results in Figure 5(a) ,when the number of attempts is 3, 4, 5 and 6, for words with or without prefixes and suffixes, the variance and coefficient of variation are basically the same, and there is a certain difference in the mean value, but the difference is not large.So in general, the presence or absence of prefixes has little effect on the distribution of results in the hard mode. Therefore, we need to further analyze the influence of attributes such as the occurrence frequency of words and the amount of information on the distribution of results in difficult mode.

5.3.2 Probability of words

We can find the probability of a word appearing in the English text from the woldfram database . For this probability in normalization.

$$P'_{word_i} = \frac{P_{word_i}}{\sum_{i=1}^{359} P_{word_i}} \tag{11}$$

To reflect the normalized influence of word probabilities, the following mapping is therefore made.

$$Q_{word_i} = f\left(P'_{word_i}\right) = \frac{2}{1+e^{-200P'_{word_i}}} \tag{12}$$

5.3.3 Probability of occurrence of initials

The probability of appearing in woldfram's database when it is the first letter of a word can be obtained P_{first_j}, Then the following mapping process is applied to P_{first_j}

$$Q_{first_i} = \frac{2}{1+e^{-2P_{first_i}}} \tag{13}$$

The statistical probability of the words in question and the probability of their occurrence as initial letters can be obtained according to the above calculation.

5.3.4 The amount of information in a word

For each word containing 5 letters, the probability of each letter appearing in the English text can be obtained P_{cha_i}

$$I_k = -\sum_{k=1}^{5} \log_2 P_{cha_k} \tag{14}$$

$$All_i = I_i * Q_{word_i} * Q_{first_i} \tag{15}$$

For the already existing dataset some words have been provided, so the three features can be extracted separately according to the model of word feature extraction. Meanwhile, the composite score index All_i can be calculated according to Equation 15.Then All_i for ascending order, the top 50% and the bottom 50% were selected for data observation.

Using the number of attempts as the horizontal coordinate and the percentage of word attempts as the vertical coordinate, all words in the given data set are displayed, and then a trend is fitted to the data using a Gaussian function, and the specific expressions of the two fitted trend lines are $y_1 = 32.51 * e^{\left(\frac{x-3.3944}{1.741}\right)^2}$,$y_2 = 33.90 * e^{\left(\frac{x-4.258}{1.679}\right)^2}$.The comparison between the first half and the second half is shown in the figure below:

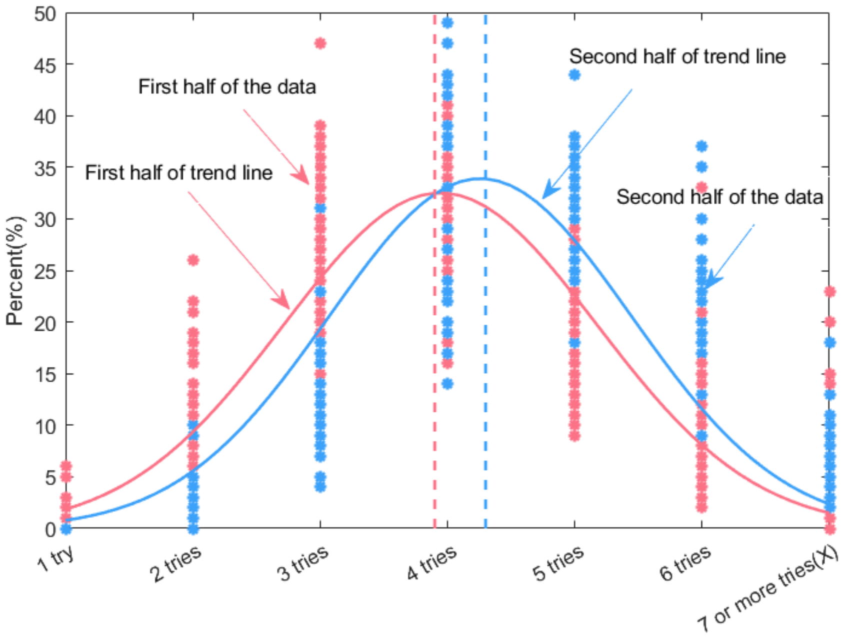


Figure 6: Word attributes affect the result

As can be seen from Figure 6, there is an obvious gap between the former part of data and the latter part of data in the report results, which also indicates that the probability of word, the probability of the first letter of a word and the information amount of a word have an obvious impact on the distribution of results in hard mode.

According to the above analysis, word attributes have significant effects on the distribution of reported results:probability of word , Probability of occurrence of initials and The amount of information in a word

6 Model 2:Game outcome prediction model

Game outcome prediction model can be simplified to GOPM.In order to predict the percentage of attempts in the report about the game on a future day, the most critical step is how to construct a bridge between the word features and the proportional distribution of features in the game report. In this paper, we first build the **GOPM** without considering the uncertainty factor, and then consider the uncertainty factor will be modified for the **GOPM**, and then build a model named **CGOPM**.

6.1 GOPM Building

6.1.1 Model Overview

In model one, feature extraction of words was performed based on words. In order to predict the percentage of a future day about the number of game attempts in the context of big data. Therefore it is necessary to model the correlation between the features of the words and the distribution of the percentage of attempts in the report. It can be observed that the reported results are characterized by an approximately normal distribution. So the relevant parameters can be solved by fitting a normal function, and then the three characteristics of the words can be used to establish a connection to each of the three parameters of the normal distribution, which can simplify the problem to a great extent. We established the **GOPM**.

The proportional distribution of the number of attempts in the game report was analyzed, and it was found that the data with the number 454 had a relatively large gap with the Gaussian normal distribution, so this data was removed directly.

6.1.2 GOPM Building

For the ith word, using the cftool toolbox in MATLAB to carry out curve fitting, we can get the parameters of Gaussian normal distribution represented by the data in the final report of each word.The Gaussian function as in equation 16. The three parameters $\tau_i, \mu_i, \sigma_i$of the normal distribution can represent the coefficient, mean and variance of the Gaussian function.

$$y = \tau * e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \tag{16}$$

For the relationship between three parameters of τ_i, μ_i, σ_i, a visual analysis of the graph is performed as follows.

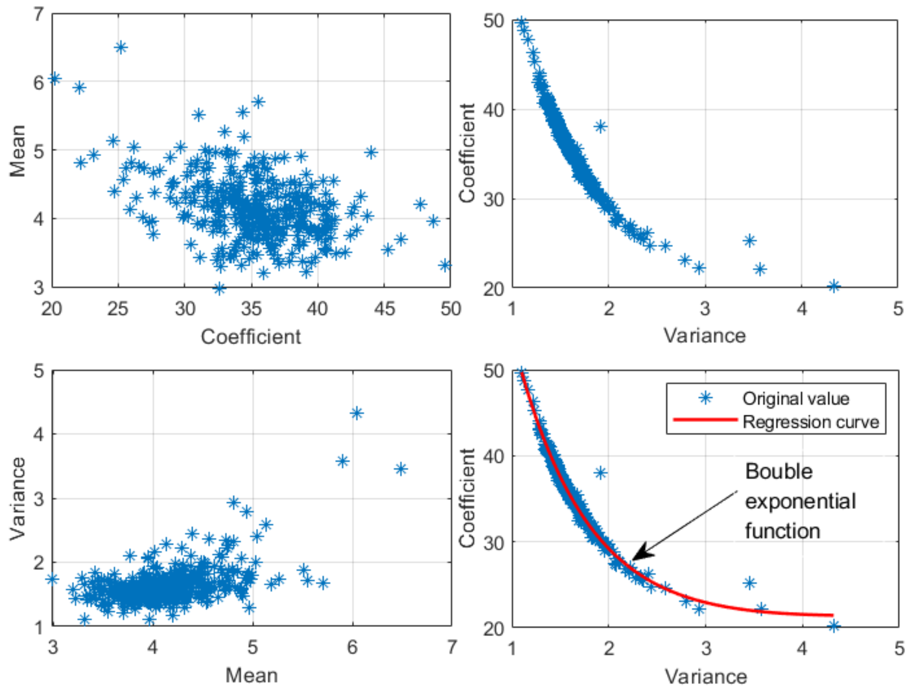


Figure 7: Schematic diagram of model coefficient relationship exploration

According to the above observation can be found that the variance σ^2 and τ there is a double exponential function relationship, you can fit the double exponential function relationship formula according to the results, and then for the formula 16 to update, update the results are as follows.

$$y = \left(129.9 * e^{\left(-1.337\sigma^2\right)} + 19.58 * e^{\left(0.0168\sigma^2\right)}\right) * e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \tag{17}$$

The three features of the words extracted in model I $Q_{word_i}, Q_{first_i}, I_i$. Since Model II needs to solve the problem of predicting the game results separately, the word probability is not normalized and function mapping is carried out directly. The first feature is corrected by the following formula

$$Q_{word_i} = \frac{2}{1 + e^{-146 P_{word_i}}} \tag{18}$$

Finally, three features $Q_{word_i}, Q_{first_i}, I_i$ are formed, and their relationships with the normal Gaussian parameters are established for these three features as follows

$$\begin{cases} \mu_i = \mu k_0 * Q_{word_i} + \mu k_1 e^{-\left(\frac{Q^2_{first_i} - \mu k_2}{\mu k_3}\right)^2} + \mu k_4 e^{-\left(\frac{I_i^2 - \mu k_5}{\mu k_6}\right)^2} + \mu k_7 \\ \sigma_i = \sigma k_0 * Q_{word_i} + \sigma k_1 e^{-\left(\frac{Q^2_{first_i} - \sigma k_2}{\sigma k_3}\right)^2} + \sigma k_4 e^{-\left(\frac{I_i^2 - \sigma k_5}{\sigma k_6}\right)^2} + \sigma k_7 \end{cases} \tag{19}$$

Then a particle swarm optimization algorithm is used, where $\tau k_0 - \tau k_7, \mu k_0 - \mu k_7, \sigma k_0 - \sigma k_7$ is the parameter value of the function, which needs to be derived. For $\tau k_0 - \tau k_7$ denotes the optimization problem in 8-dimensional space. The objective function of the optimization is

$$R^2 = \frac{\sum\left(\hat{X}-\bar{X}\right)^2}{\sum\left(X_i-\bar{X}\right)^2}, (X=\mu,\sigma) \tag{20}$$

In order to make the square of R larger,Many optimization algorithms already exist, and we use the simulated annealing particle swarm optimization algorithm. The particle swarm algorithm is a swarm intelligence algorithm designed by simulating the predatory behavior of a flock of birds. There are different food sources, large and small, in the region, and the task of the flock is to find the largest food source, in other words, the global optimal solution, and the task of the flock is to find this food source. Throughout the search process, the flock lets other birds know the location of the food source by passing information about their respective locations to each other Eventually, the whole flock can gather around the food source, i.e., the optimal solution is said to be found and the problem converges.

there exists a parameter state of $P_0 - P_7$ during the flight of the parameter particle. the flight of the particle is $V_0 - V_7$. The key optimization formula of the algorithm is shown below

$$V_i^7 = \varphi * \left[w * V_i^7 + c_1 * r_1 * \left(P_i^7 - X_i^7\right) + c_2 * r_2 * \left(X_i^7 - X_i^7\right)\right] \ (X=\tau,\mu,\sigma) \tag{21}$$

$$X_i^7 = X_i^7 + \alpha * V_i^7 \tag{22}$$

i represents the particle number, which is the number of the word, indicating specifically which word's characteristics correspond to its normal distribution function; w is the inertia factor, whose value range is non-negative; c is the acceleration constant, whose value range is non-negative constant; r is a random number in the range of 0 to 1; α is the constraint factor, which is used to control the weight of the velocity. The following relationship exists among them.

$$\varphi = \frac{2}{\left|2-C-\sqrt{C^2-4C}\right|}, C = c_1 + c_2 \tag{23}$$

In order to make the solution result of intelligent algorithm more accurate, we add the following restrictions to the solution process.We believe that the final for the proportion of 7 clock attempts in the game is $(p_1, p_2, \cdots p_7)$. When the formula 24 is satisfied, we think the solution result is reasonable.

$$95\% \leq \sum_{i=1}^{7} p_i \leq 105\% \tag{24}$$

The initialization parameters of the population are as follows

Table 3: Particle swarm optimization algorithm initialization parameters

Stock size	c_1	c_2	w	Iteration number
20	2.05	2.05	0.5	50

When the optimization result is solved, the final R_μ^2, R_σ^2 about μ_i, σ_i can be obtained separately.Normally, the closer these three R^2 are to 1, the better the effect is, which means the more confidence we have in predicting correctly.These three parameters have an equal impact on our final results,thus, the mean of R_μ^2, R_σ^2 can be interpreted as the confidence con predicted by our model.

$$con = \frac{R_\mu^2 + R_\sigma^2}{2} \tag{25}$$

6.1.3 Solution of GOPM

The solving process of the model, for the objective function R^2 will be iterated continuously to know to find the optimal solution of the objective function, and the 8-dimensional space of this optimal solution is the value we need to find. This is the parameter about the functional relationship between word characteristics and the proportion of attempts in the report. The iterative process of solving is as follows

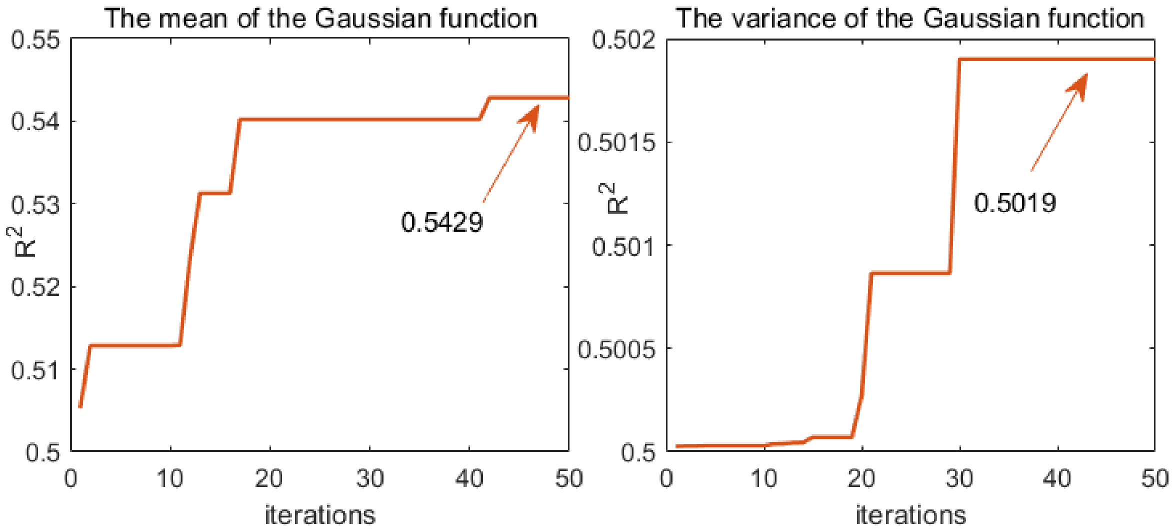


Figure 8: Particle swarm optimization algorithm iteration diagram

The results of the parameters in the iterative process are tabulated as follows

Table 4: The parameter results of the model

	Results of model parameters about μ_i							
parameters	$\mu_i k_0$	$\mu_i k_1$	$\mu_i k_2$	$\mu_i k_3$	$\mu_i k_4$	$\mu_i k_5$	$\mu_i k_6$	$\mu_i k_7$
μ_i	0.412	5.942	4.946	-2.979	-6.100	9.131	-1.576	-6.183
	Results of model parameters about μ_i							
parameters	$\sigma_i k_0$	$\sigma_i k_1$	$\sigma_i k_2$	$\sigma_i k_3$	$\sigma_i k_4$	$\sigma_i k_5$	$\sigma_i k_6$	$\sigma_i k_7$
σ_i	-1.166	-4.898	-0.444	31.472	36.702	-6.345	-27.017	-2.885

6.2 CGOPM Building

The uncertainty factor in this game is divided into two main parts, mainly the human uncertainty factor and the game uncertainty factor. For a person playing the game, there are many factors that determine how many times he can use in order to guess the correct answer. For example, the age of the person, the vocabulary of the person, the choice of words in the person's mind while playing the game. In this part of the uncertainty factor, with a relatively large amount of sample data, it can be considered that the individual differences have been eliminated, showing the characteristics of an overall normal distribution. On the other hand, the main consideration is the difficulty of the questions in the game itself, and the setting mechanism of the game itself. This aspect also affects our prediction results, which is unchangeable, which is objective. Therefore, we need to correct for the prediction errors caused by the uncertainty factors of the game itself.

6.2.1 CGOPM building

Under the influence of many uncertainties such as the setting mechanism of the game itself, some model corrections need to be made relative to the formula 19, and the corrected model is as follows

$$\begin{cases} \mu_i = \mu k_0 * Q_{word_i} + \mu k_1 e^{\left(\frac{Q^2_{first_i} - \mu k_2}{\mu k_3}\right)^2} + \mu k_4 e^{\left(\frac{I_i^2 - \mu k_5}{\mu k_6}\right)^2} + \mu k_7 + \varepsilon_\mu \\ \sigma_i = \sigma k_0 * Q_{word_i} + \sigma k_1 e^{\left(\frac{Q^2_{first_i} - \sigma k_2}{\sigma k_3}\right)^2} + \sigma k_4 e^{\left(\frac{I_i^2 - \sigma k_5}{\sigma k_6}\right)^2} + \sigma k_7 + \varepsilon_\sigma \end{cases} \tag{26}$$

In the equation 26, $\varepsilon_\mu, \varepsilon_\sigma$ is the uncertainty correction value.In the formula, a correction quantity is added. In order to correct for the correction quantity, the LSTM approach is taken to predict its correction value.

6.2.2 Bias correction based on LSTM model

The distribution of the properties of the word in question with respect to the percentage of game attempts can be obtained in the GOPM model. For any word, when its attributes are extracted, the percentage of various attempts calculated by bringing it into the model deviates from the actual percentage of game attempts. For the sample data set, the error is calculated for each word, and there is a random nature to this error. So, we use the input of the three attributes of the word and train the deviation as the output using the LSTM model so that the amount of deviation correction can be predicted.

LSTM model is a variant of recurrent neural network (RNN). It is well suited for classification, processing and prediction based on time series data. A common LSTM unit consists of a unit, an input gate, an output gate and a forgetting gate. The unit memorizes values at arbitrary time intervals, and the three gates regulate the flow of information in and out of the unit.The structure diagram of our nested two-layer LSTM model is as follows.

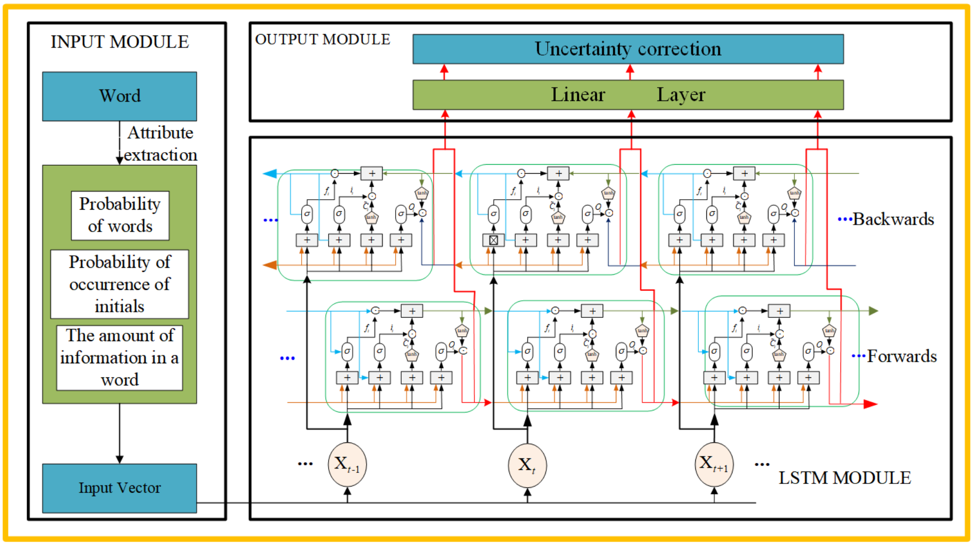


Figure 9: LSTM model schematic

1. **LSTM model data description**

   The input data of the LSTM model are the $Q_{word_i}, Q_{first_i}, I_i$ three attributes of the word; the output result of the LSTM model is the deviation of the theoretically calculated value from the actual value of the WILR model. According to the known data of the subject, the first 250 data are training data and the last 100 data are testing data. The LSTM model is trained according to the above way

2. **Input Data**

   For the input vector, we directly use the feature vector extracted from the monomer as input, and the specific formula refers to equation 12,equation 13,equation 14. For the input container v_f composed of input vectors, it is defined as follows.

Table 5: Abnormal data filtering results

Symbol	Definition
Q_{word_i}	The probability of the word
Q_{first_i}	The probability of the initial letter of the word
I_i	The amount of information for the word

3. **LSTM Model**

   The workflow of a one-layer LSTM unit is shown below

$$\begin{cases} i=\sigma\left(W_{ii}x+b_{ii}+W_{hi}h+b_{hi}\right) \\ f=\sigma\left(W_{if}x+b_{if}+W_{hf}h+b_{hf}\right) \\ g=\tanh\left(W_{ig}x+b_{ig}+W_{ho}h+b_{ho}\right) \\ o=\sigma\left(W_{io}x+b_{io}+W_{ho}h+b_{ho}\right) \\ c'=f*c+i*g \\ h'=o*\tanh\left(c'\right) \end{cases} \tag{27}$$

The notations for equations above is shown in table as follows:

Table 6: Notations for one-layer LSTM

Symbol	Definition
x	The concatenated input vector for the LSTM
h	Hidden state,containing encoded information for the sequence flow
c	Cell state,tracking dependencies between the elements in the input sequence
i	Input gate,controlling the extent to which a new value flows into the cell
f	Forget gate,controlling the extent to which a value remains in the cell
g	Gathered input value from input x and current hidden state
o	Output gate,controlling the extent to which the cell is used to compute outputs
W	The weight matrix for transitions
b	The bias for transitions
σ	The sigmoid function
$tanh$	The hyperbolic tangent function

4. Output Model

In the output module, we simply input the hidden state of the upper LSTM into the linear layer, predict the feature vector, and obtain the correction value of the uncertainty.

5. Bias correction analysis results

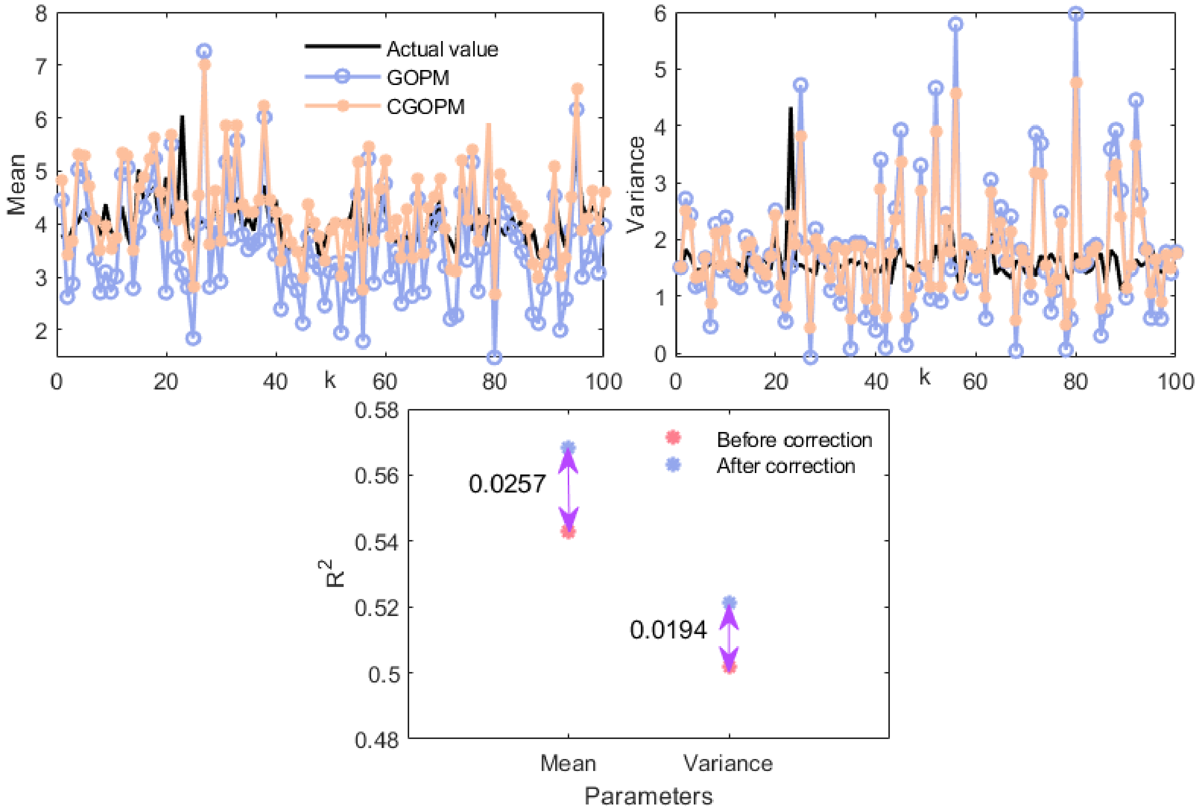


Figure 10: Bias correction analysis results

According to Figure 10, when the uncertainty factor of the game is introduced, the LSTM model is used to correct the deviation correction and find that R_2 has been improved,forecast confidence for mean increases 2.57% and variance increases forecast confidence by 1.94% ,which means that the model has received the influence of the uncertainty factor of the game, such as the uncertainty of the difficulty of the word given by the game every day.At the same time, the accuracy of our prediction model has been improved from 52.24% to 54.49%

6.2.3 EERIE prediction results

According to the above modified model, the word "EERIE" can be calculated to calculate the mean and variance, mean $\mu = 3.4568$, variance $\sigma^2 = 1.995$ of the normal distribution of its final game statistical results, and finally the prediction results of the proportion of game attempts in the report are shown in Table 7

Table 7: EERIE prediction results

Number of game attempts	1	2	3	4	5	6	X
Predicted proportion	6%	17%	28%	29%	16%	5%	1%

For the result, the predicted result can be considered reasonable according to the formula **??**

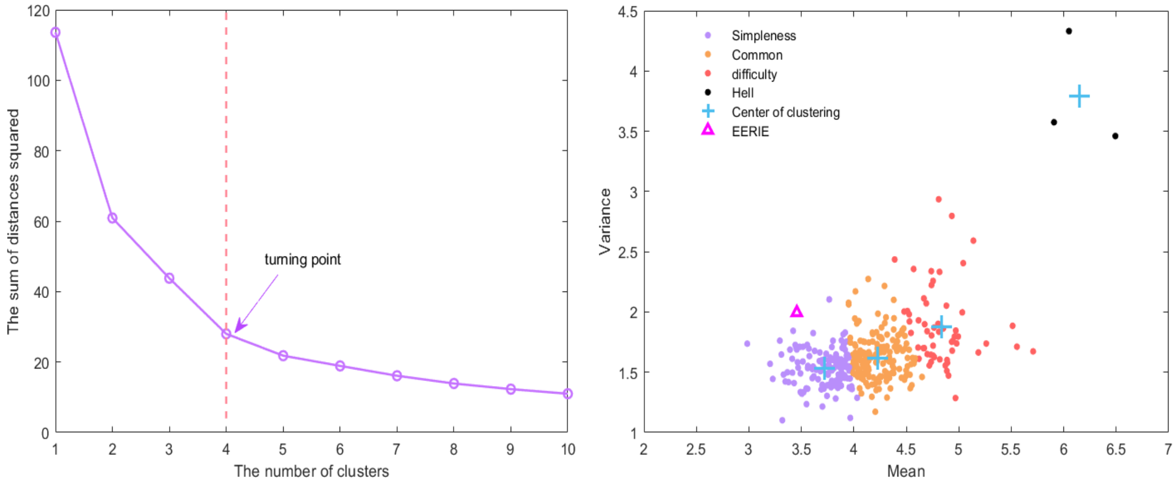
7 Model 3:Word difficulty classification model

7.1 Word K-means clustering model

In order to classify words, it is first necessary to divide the difficulty of words according to the results of the game reports known in the question. The main idea is to fit a Gaussian function based on the number of word attempts. After fitting the Gaussian function, the two parameters of mean and variance are classified by the elbow method, and the division results are shown in Figure 11(a), and the words are clustered after the classification is completed, and the visualization results of cluster analysis are shown Figure 11(b). For new words, word attributes can be established according to GOPM to establish correlations with different proportions of mean and variance in the report, and mean and variance can be classified by cluster analysis.

For Figure 11(a), the number of categories classified according to the elbow method is 4, which are simple, normal, difficult, and very difficult, and the different colors of Figure 11(b) are different categories.Combined with the prediction results of model , the difficulty of the word EERIE is simple.

In order to measure the accuracy of our classification model, we calculated the contour coefficient and CH value under K-means classification.The contour coefficient of the classification is 0.4, and the value range of the contour coefficient is -1 to 1. According to the literature, the contour coefficient of 0.4 indicates that the classification results are better. The CH value is 345, which indicates the compactness of the feature cluster, and a CH value greater than 100 indicates a higher degree of compactness. All in all, our classification results are relatively good.



(a) Word probability statistical results (b) As an acronym probability statistics result

Figure 11: Statistical results of words and letters

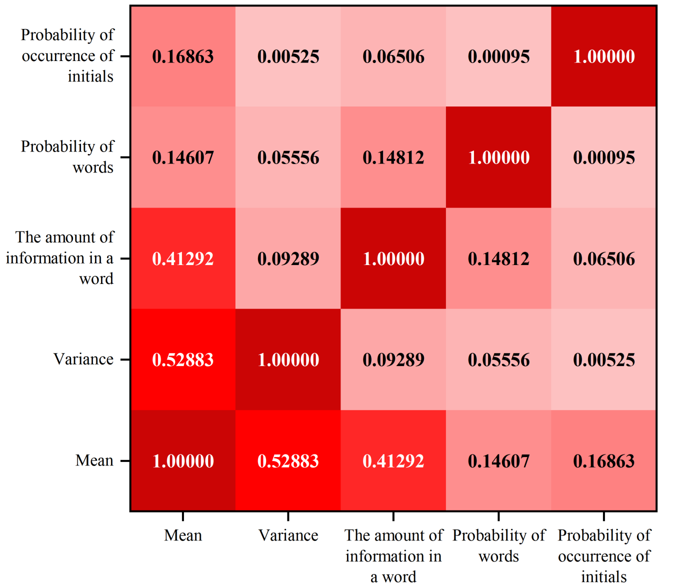
7.2 Correlation analysis of word attributes with reports

In order to explore which attributes of words the classification results are related to, we take the method of typical correlation analysis, for $X = [u_i, \sigma_i, Q_{word_i}, Q_{first_i}, I_i]$.

Covariance is calculated separately for any two of these groups.The calculation method is such as equation 28.

$$cov\left(x_2, x_1\right) = \frac{\sum_{i=1}^{M}\left(x_1^i - \overline{x_1}\right)^2}{M-1}\left(x_i = u_i, \sigma_i, Q_{word_i}, Q_{first_i}, I_i\right). \tag{28}$$

The visualization of the correlation coefficient is shown in Figure 12



	Mean	Variance	The amount of information in a word	Probability of words	Probability of occurrence of initials
Probability of occurrence of initials	0.16863	0.00525	0.06506	0.00095	1.00000
Probability of words	0.14607	0.05556	0.14812	1.00000	0.00095
The amount of information in a word	0.41292	0.09289	1.00000	0.14812	0.06506
Variance	0.52883	1.00000	0.09289	0.05556	0.00525
Mean	1.00000	0.52883	0.41292	0.14607	0.16863

Figure 12: Schematic diagram of analysis data selection

According to Figure12, ,among the three attributes of the words we selected, the information content of the words has a greater correlation with the mean value of the

Gaussian function, while the other two attributes have a lesser correlation with the mean value. At the same time, the three attributes we selected have little correlation with the variance of the Gaussian function.

It can be seen from the above classification results that the classification is mainly based on mean value, and has little relationship with variance. Therefore, it can be shown that the information content of words has a greater impact on the classification results, while the other two have a lesser impact on the classification results.

8 Interesting Features

In the above modeling process, we extracted features such as word probability, word information, word initial letter occurrence probability, word prefixes and suffixes, and daily percentage distribution from the data set.

8.1 The effect of time

In the analysis of the first question, we can see that the number of game players has a process of rapid rise and slow decline, and finally stabilizes. Therefore, we divide the time course into three stages: rising stage, falling stage and plateau stage. It can be seen from the figure below that in the first stage, the proportion of people who can guess the word four times before is greater than in the second and third stages. We guess that at the initial release stage of the game, most of the players are people who have studied or are more interested in languages, so they can guess words in a relatively small number of times.

But we also found that the data in the second stage and the third stage did not show a significant difference, indicating that there is no obvious relationship between word difficulty and time, which also shows that in this game, the choice of daily words is random.

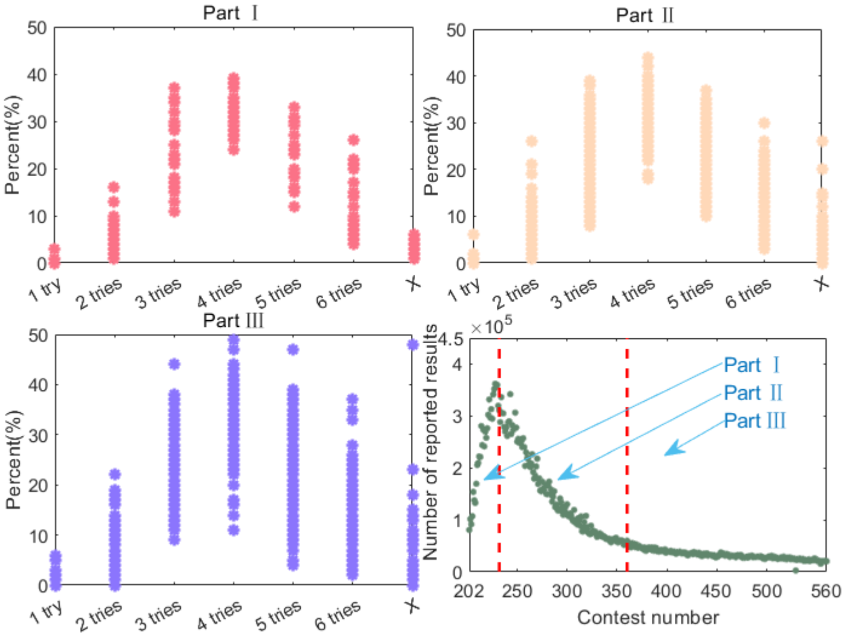


Figure 13: The effect of time

8.2 Word pronunciation features

In the English phonetic alphabet, there are 20 vowels and 28 consonants. Vowels can be divided into short vowels, long vowels and diphthongs. We find the first vowel in the phonetic symbol of each word and classify the words accordingly. The resulting distribution is shown in Figure 14, where words containing short vowels account for almost half of all words, and the proportions of diphthongs and long vowels are almost equal.

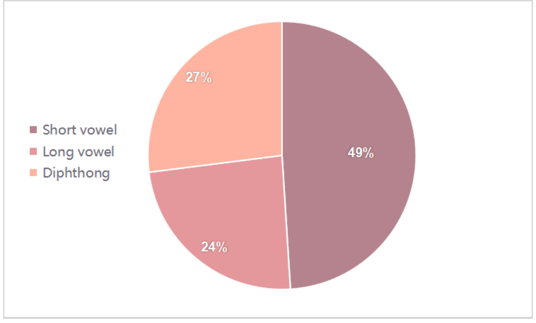


Figure 14: Distribution of words with different vowels

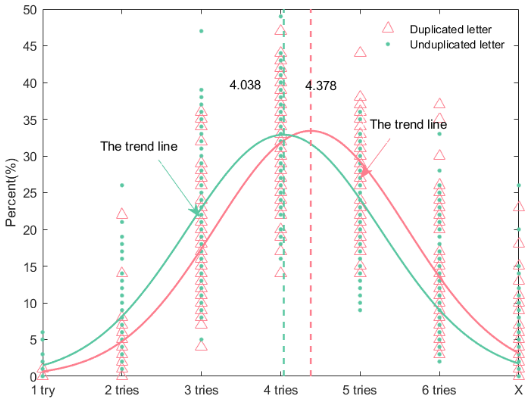


Figure 15: Distribution of game results with or without repeated letters

8.3 Repeated letters in words

Different words are constructed differently, some words have no repeated letters inside, and some words have two or more repeated letters inside. We divide all words into two categories: those that contain repeated letters and those that do not. As shown in Figure 15, comparing the statistical laws of these two types of words, we found that words with repeated letters seem to require more guesses than words without repeated letters.

9 Model Evaluation

9.1 Model Strengths

- We analyzed a large amount of data and found that the percentage of daily reports basically conforms to a normal distribution. This conclusion is realistic and reliable.
- We analyze datasets from multiple perspectives to extract information and maximize data mining.
- We used the particle swarm algorithm of simulated annealing to predict the percentage distribution of daily reports, and on this basis, used the long-short memory neural network to correct the uncertainty, which made our prediction model more accurate.
- Validation of the K-Means algorithm using Silhouette Coefficient, DBI, and CH ensures the accuracy of our word difficulty classification model.

9.2 Possible Improvements

- In the prediction model, we only extracted the word information, the probability of the word appearing in the English text, and the probability of the first letter of the word, but it is possible to extract more information from the word or the data set.
- When counting word-related information, if we use a large enough database, we can get more accurate indicators.

References

[1] Huang Xiaoyi,Xu Haoran,Chu Jizheng.(2022).Nonlinear Model Order Selection: A GMM Clustering Approach Based on a Genetic Version of EM Algorithm. Mathematical Problems in Engineering. doi:10.1155/2022/9958210.

[2] Tian Fukui,Gao Yunfeng,Yang Chuanchuan.(2022).GMM based low-complexity adaptive machine-learning equalizers for optical fiber communication. Optics Communications. doi:10.1016/J.OPTCOM.2022.128312.

[3] Rutkowski Rachel A.,Lee John D.,Coller Ryan J. ,Werner Nicole E..(2022).How can text mining support qualitative data analysis?. Proceedings of the Human Factors and Ergonomics Society Annual Meeting(1). doi:10.1177/1071181322661535.

[4] Lipovetsky Stan.(2022).Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Technometrics(1). doi:10.1080/00401706.2021.2020521.

[5] Troccoli Edric Brasileiro,Cerqueira Alexsandro Guerra,Lemos Jonh Brian , Holz Michael.(2022).K-means clustering using principal component analysis to automate label organization in multi-attribute seismic facies analysis. Journal of Applied Geophysics. doi:10.1016/J.JAPPGEO.2022.104555.

[6] Rasid Mamat Abd,Susilawati Mohamed Fatma,Afendee Mohamed Mohamad,Mohd Rawi Norkhairani , Isa Awang Mohd.(2018).Silhouette index for determining optimal k-means clustering on images in different color models. International Journal of Engineering,Technology(2.14).doi:10.14419/ijet.v7i2.14.11464.

[7] Cui Yingying,Meng Xi,Qiao Junfei.(2022).A multi-objective particle swarm optimization algorithm based on two-archive mechanism. Applied Soft Computing Journal. doi:10.1016/J.ASOC.2022.108532.

[8] Clerc, M. ,Kennedy, J. (2002). The particle swarmexplosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation, 6(1), 58-73.

Letter

To：The Puzzle Editor of the ***New York Times***
From：MCM Team #2306479
Subject：Wordle Game Results Analysis Model and
Game Development Suggestions
Date：February 21, 2023

Dear editor,

We are honored to inform you that our team has built statistical and predictive models on Wordle game outcomes and explored some of the deepest features of words, and results are described below:

First, we built a Gaussian mixture model to achieve a prediction of the number of users who will tweet the results of the game on a certain day in the future. We forecast the reported total range for March 1, 2023 to be (2001, 37156).

Then, we did some research on the English words themselves to extract three features of the letters that have an impact on the outcome of the game. Using the particle swarm optimization algorithm based on simulated annealing, combined with the long-short memory neural network for uncertainty compensation, the influence model of the above three attributes on the game result is established. We substitute a word '"eerie" that has never appeared in the game into our model, and we can predict that the result distribution is [1 try, 2 tries, 3 tries, 4 tries, 5 tries, 6 tries, X]=[0.06,0.17,0.28,0.29,0.16,0.05,0.01].

Finally, we use the K-Means algorithm to establish a classification model for word difficulty, and study the impact of word attributes on classification results through canonical correlation analysis. Also substitute the word "eerie" into the model, and the result is that "eerie" belongs to the "easy" class.

Based on our results, our team would like to give you and The New York Times some suggestions:

- In order to allow more players to reap the joy of guessing words, I hope that each player can have more guessing opportunities.
- To add to the challenge, players can choose the length of the guessed word themselves.
- In order to attract more people to participate, a PK mode has been developed, allowing users to directly share games on social software and compete with friends.

Thank you for your reading!
Your sincerely
MCM Team #2306479