Prediction model of Wordle report based on GA-BP neural network

Shiying Wang[1, *], Rui Wang[2, #], Peilin Liu[1, #]

[1]School of Information and Communication Engineering, Communication University of China, Beijing, 100024, China

[2]School of Computer and Cyber Sciences, Communication University of China, Beijing, 100024, China

* Corresponding Author Email: cucices_wsy@163.com

#These authors contributed equally.

Abstract. Natural Language Processing (NLP) plays a vital role in artificial intelligence, enabling machines to understand and generate human language. This paper proposes a novel approach to improve NLP tasks by integrating a Genetic Algorithm-based BP neural network model. The model combines the strengths of genetic algorithms and neural networks. Genetic algorithms offer effective search capabilities, while the BP neural network provides a flexible learning framework. By integrating these techniques, the proposed model aims to overcome limitations of traditional NLP models. It effectively handles the complexity of natural language data, provides efficient training by avoiding local optima, and exhibits enhanced generalization capabilities for unseen data. Extensive experiments on benchmark NLP datasets validate the effectiveness of the proposed model. Results demonstrate its superiority over state-of-the-art NLP models in terms of accuracy, efficiency, and robustness. This paper presents a novel approach to enhance NLP tasks by integrating a genetic algorithm-based BP neural network model. The model shows promising results in various NLP applications and offers advantages over traditional approaches. This research contributes to the advancement of NLP techniques, facilitating more accurate and efficient language processing systems.

Keywords: Exponential fit, GA, BP Neural Network, Normality test.

1. Introduction

Radhakrishna Rao said, “In the ultimate analysis, all knowledge in history; in the abstract sense, all science is mathematics; in a rational world all judgment is statistically.” – Statistics and Truth.

With the development of the Internet and mobile networks, more and more people play trivia games over the Internet, including word-guessing puzzles. Wordle is an alphabet puzzle game of guessing words that has become popular on the Internet, courtesy of The New York Times. And the presence of social media such as Twitter has made this game more and more popular, this also makes it a need to solve the problem or deal with the data behind it through a mathematical modeling approach[1].

Wordle offers a popular puzzle each day, which each player playing the game can try to solve by guessing a five-letter word six times or less without any tips. Each time a player enters a word, it must be a real word; unreal, made-up words cannot be used. When the player submits an entry, if one of the five letters is correct and in the right place it will turn green, if the letter is correct but in the wrong place it will show yellow, and if it shows gray, it means the answer to the puzzle has nothing to do with it. More difficult than the regular mode used for direct play is Hard Mode, in which the player's subsequent guesses of the words must use the letters that were guessed correctly before, which are the green and yellow letters.

By reading other relevant literature, it can be seen that most of the other studies only use BP neural network, which has certain limitations. In this paper, on the basis of BP neural network, genetic algorithm is used to improve the accuracy of the model. The purpose of this study is to build a model to predict the distribution of the reported results at a future date, i.e. the associated percentages of (1,

2, 3, 4, 5, 6, X). And analyze the uncertainty of the created model and give an example of our prediction for the word EERIE on March 1, 2023.

2. Materials and Methods

2.1. Notations

The related symbols of this study are shown in Table 1.

Table 1. The related symbols of this study

Symbol	Description	Unit
AIC	A measure of the goodness of fit of a statistical model, the smaller the value the better.	-
τ	Date	Days
NPR	Number of predicted reported results	PCS
R^2	The degree of fit of the time series, the closer to 1, the better.	- -
σ	The Standard Deviation	-
PNH	$\frac{Number\ in\ Hard\ Mode}{Munber\ of\ reported\ results}$	-
FQ	The frequency of word occurrence.	-
SA	Whether the word contains the same letter within the word	

Other specific notations, if necessary, will be mentioned and illustrated while we're building models.

2.2. Data Acquisition and Pre-processing

This article obtained a data report about Wordle on the Github website. Some of the data reports are shown in Figure 1:

					Percent in						
Date	Contest number	Word	Number of reported results	Number in hard mode	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
2022/12/31	560	manly	20380	1899	0	2	17	37	29	12	2
2022/12/30	559	molar	21204	1973	0	4	21	38	26	9	1
2022/12/29	558	havoc	20001	1919	0	2	16	38	30	12	2
2022/12/28	557	impel	20160	1937	0	3	21	40	25	9	1
2022/12/27	556	condo	20879	2012	0	2	17	35	29	14	3
2022/12/26	555	judge	20011	2043	0	2	8	16	26	33	14
2022/12/25	554	extra	15554	1562	1	5	20	35	28	10	1
2022/12/24	553	poise	20281	1911	2	11	34	32	15	6	1
2022/12/23	552	aorta	21937	2112	0	7	26	35	20	10	3
2022/12/22	551	excel	20490	2034	0	1	13	34	34	15	2
2022/12/21	550	lunar	22180	2036	0	5	32	40	17	4	0
2022/12/20	549	third	24137	2261	1	10	47	32	9	2	0
2022/12/19	548	slate	26010	2422	6	14	33	27	13	5	1
2022/12/18	547	taper	22166	2108	0	8	28	30	20	11	3
2022/12/17	546	chord	22336	2088	0	7	39	38	13	3	0
2022/12/16	545	rprobe	22853	2160	0	6	24	32	24	11	3
2022/12/15	544	rival	22176	2127	0	7	27	35	22	8	1
2022/12/14	543	usual	20824	2048	0	3	20	39	27	10	1
2022/12/13	542	spoke	24101	2224	0	6	31	38	19	5	0
2022/12/12	541	apply	22873	2150	0	5	28	38	22	7	1
2022/12/11	540	naïve	21947	2075	1	7	24	32	24	11	1
2022/12/10	539	knock	21157	2041	0	3	18	43	27	8	1
2022/12/9	538	braid	23640	2165	0	10	36	35	14	3	0
2022/12/8	537	infer	21199	1863	0	3	19	33	26	14	3
2022/12/7	536	joust	24899	2388	0	6	29	34	21	8	2
2022/12/6	535	amber	23509	2261	0	6	22	33	24	12	3
2022/12/5	534	woken	23153	2200	0	2	10	25	36	23	4

Figure 1. Partial Data Report

Before data analysis, the availability of data must be guaranteed.

$$X' = \frac{x-min}{max-min} \qquad X'' = X' \cdot (max - min) + min \qquad (1)$$

Dirty value processing. The data of the number of reported results contain some error numbers that are significantly lower than other numbers. For the accuracy of subsequent modeling, the dirty value is replaced by the number which is calculated by interpolation fitting.

Data standardization. In regression prediction and neural network training, standardization plays an important role in giving eigenvalues equal weight. We processed the data by normalization using Equation (1), where X and x mean arbitrary data, max is the maximum value of data, and min is the minimum value of data.

2.3. Introduction of the method

Backpropagation (BP) neural network is a widely used artificial neural network algorithm for supervised learning. It is particularly effective in solving problems related to pattern recognition, classification, and regression. The basic principle of a BP neural network involves the propagation of error signals backward through the network to adjust the weights of the connections between neurons. The network consists of multiple layers, including an input layer, one or more hidden layers, and an output layer. The network structure is shown in Figure 2[2].

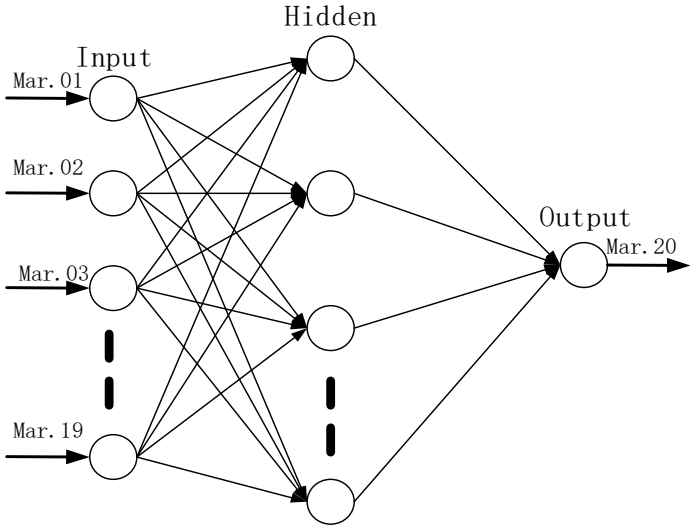

Figure 2. Neural network structure

The weight update is performed using a gradient descent optimization algorithm, which adjusts the weights in the direction that reduces the error. The gradient descent algorithm calculates the gradient of the error function with respect to the weights, indicating the direction of steepest descent. The weight update is performed iteratively for each training example in the dataset, and the process is repeated for multiple epochs until the network converges to a satisfactory level of performance.

In the backward propagation step, the error is distributed among the neurons in the previous layer based on their contribution to the error. This distribution is achieved by calculating the partial derivative of the error with respect to the output of each neuron in the previous layer.

The partial derivatives are computed using the chain rule of calculus, which allows the error to be propagated backward through the network. The weight update is then performed based on these partial derivatives, adjusting the weights to minimize the error.

The genetic algorithm (GA) is a computational method inspired by the process of natural selection and evolution. It is commonly used to solve optimization problems. The basic principle of a genetic algorithm involves creating a population of potential solutions represented as individuals, and then iteratively applying genetic operators such as selection, crossover, and mutation to evolve the population towards better solutions[3].

In a typical GA, the population starts with a set of randomly generated individuals. Each individual represents a potential solution to the problem at hand. The fitness of each individual is evaluated based on its performance in solving the problem. Individuals with higher fitness values have a higher chance of being selected for reproduction. During the selection process, individuals are chosen from the population to form a mating pool. The selection is usually based on the fitness values, where individuals with higher fitness have a higher probability of being selected. This process simulates the survival of the fittest in nature. Crossover is the genetic operator that emulates the combination of genetic material from two parents to produce offspring. It involves exchanging genetic information between selected individuals to create new individuals with a combination of their characteristics. Mutation is another genetic operator that introduces small random changes to the genetic material of individuals. This helps to maintain diversity in the population and enables exploration of new areas in the search space.

The advantages of combining GA with BP neural network include:

(1) Improved global search capability: The GA explores a large search space efficiently, enabling the identification of global optima that may be missed by traditional optimization methods.

(2) Avoidance of local optima: The GA's ability to maintain diversity in the population helps to avoid getting stuck in local optima, promoting exploration of different regions of the search space.

(3) Automatic parameter tuning: The GA can be used to optimize not only the weights and biases of the neural network but also other hyperparameters, such as learning rate or network architecture, leading to improved performance.

(4) Robustness: The hybrid model is less sensitive to the initial conditions and noise in the data, as the GA helps to find a good solution even in the presence of uncertainties[4].

2.4. Model-evaluation Index

In order to ensure the accuracy of the NPM models used in this paper, the following indicators are used for verification [5]:

$$\mathrm{SSE} = \sum(\mathrm{y} - \hat{\mathrm{y}})^2 \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{m}(y^{(i)} - \bar{y}^{(i)})^2} \tag{3}$$

$$\text{Adjusted } \mathrm{R}^2 = 1 - [(1 - \mathrm{R}^2) * (\mathrm{n} - 1)/(\mathrm{n} - \mathrm{k} - 1)] \tag{4}$$

$$RMSE = \sqrt{\frac{1}{\mathrm{m}} \sum_{\mathrm{i=1}}^{\mathrm{m}} (y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2} \tag{5}$$

3. Neural Network Percentage Prediction Model (NPM) Establishment and Solution

3.1. Features of words and model inputs

To get accurate prediction results by using a neural network model, extracting the features of data as input of the network is of great significance. This paper choose seven features as the input: FQ, SA, RT, IL, CIE, WM, and the value of PNH. To ensure that the input features of the model are not duplicated and strongly correlated, this paper tested the correlation between the feature values and each other. The result is shown in Figure 3, the P-value between each feature is close to 0, so it can assume that there is no strong correlation between the features.

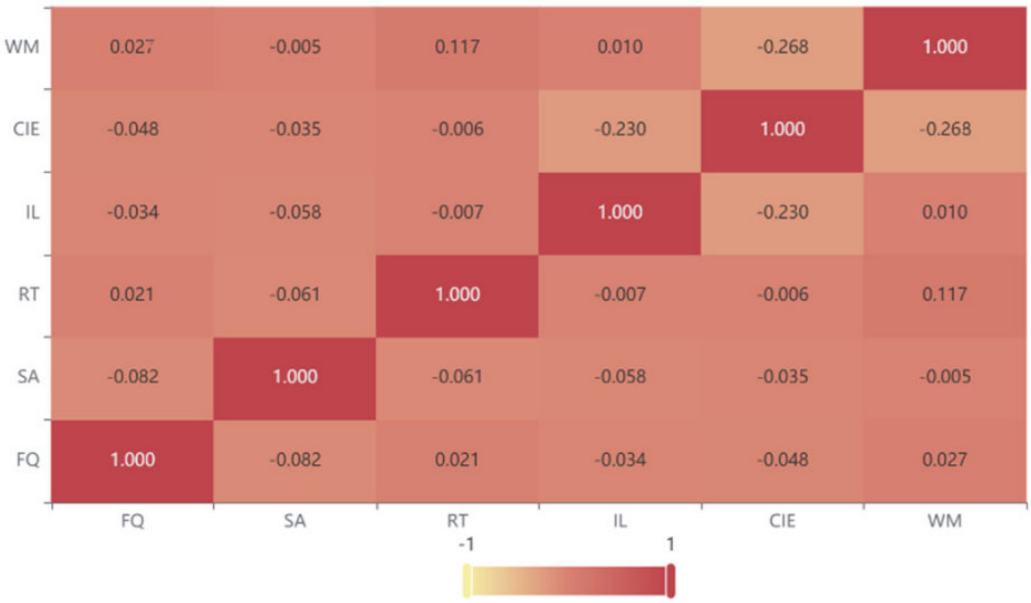

Figure 3. Correlation Heat Map

3.2. Establishment of the model

Step 1: Mapping the date as PNH

Based on the analysis of the game mechanics of the Hard Mode, a player must use the correct letters they have found in subsequent guesses, the Hard Mode gives players a smaller chance of excluding letters. Therefore, the rise of the percentage of Hard mode will affect the percentage of (1, 2, 3, 4, 5, 6, X). Thus, it is important to consider the PNH in our model. By examining the relationship between time and the PNH, this paper found that it conforms to an exponential distribution, with the PNH rising as time progresses. This paper fitted the curve using exponential fitting and obtained the curve fit as in Figure 4 and the deviation values as in Figure 5, after which this paper also obtained the results in Table 1. It can be seen that the RMSE and SSE are extremely small and errors are small, too.$R^2$is close to 1, indicating a strong correlation. Therefore, the percentage of the times of attempts can be predicted for a given date in the future[6].

$$PNH(\tau) = \mathrm{a} \cdot e^{b\tau} + c \cdot e^{d\tau} \tag{6}$$

Coefficients(with 95% confidence bounds):

$$\mathrm{a} = -17.84(-2.033e+05, 2.033e+05) \tag{7}$$

$$b = -0.002661(-0.2072, 0.2019) \tag{8}$$

$$c = 17.86(-2.033e+05, 2.033e+05) \tag{9}$$

$$d = -0.002626(-0.2057, 0.2005) \tag{10}$$

As time progresses, the number of veteran users of Wordle is increasing. Veteran users will prefer to play in Hard Mode. On the grounds of the findings, this paper believes that by mapping the date as PNH, the predictive models can take the effect of time factors into account(Table 2).

Table 2. Exponential Fit Results

SSE	0.005632
R^2	0.9681
Adjust R^2	0.9679
RMSE	0.003994

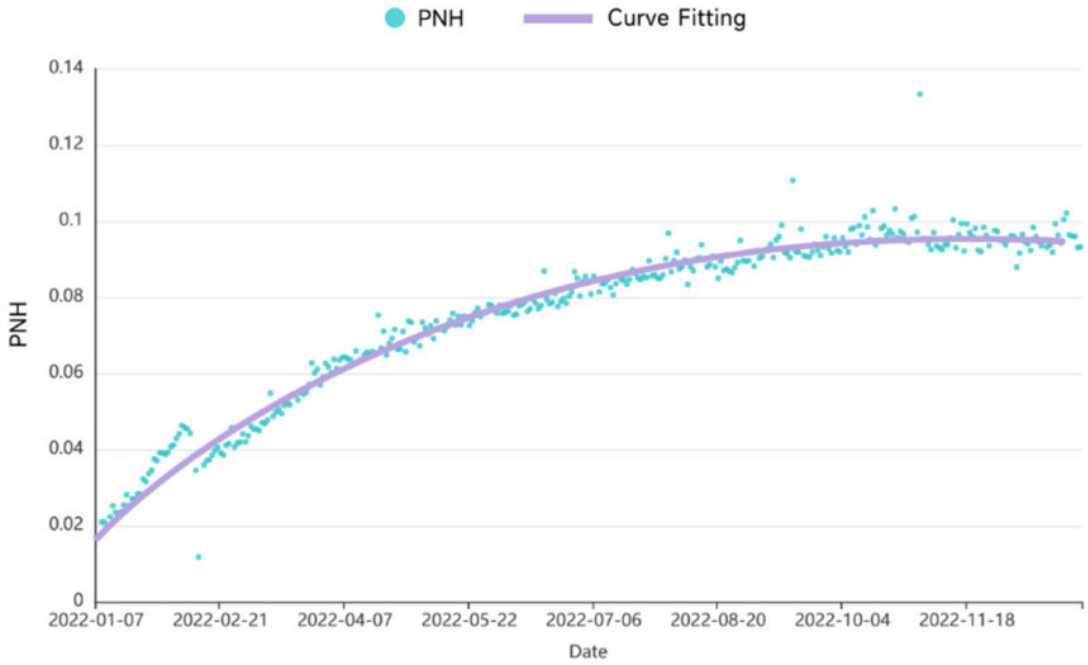

Figure 4. Exponential Fitted Curve Chart.

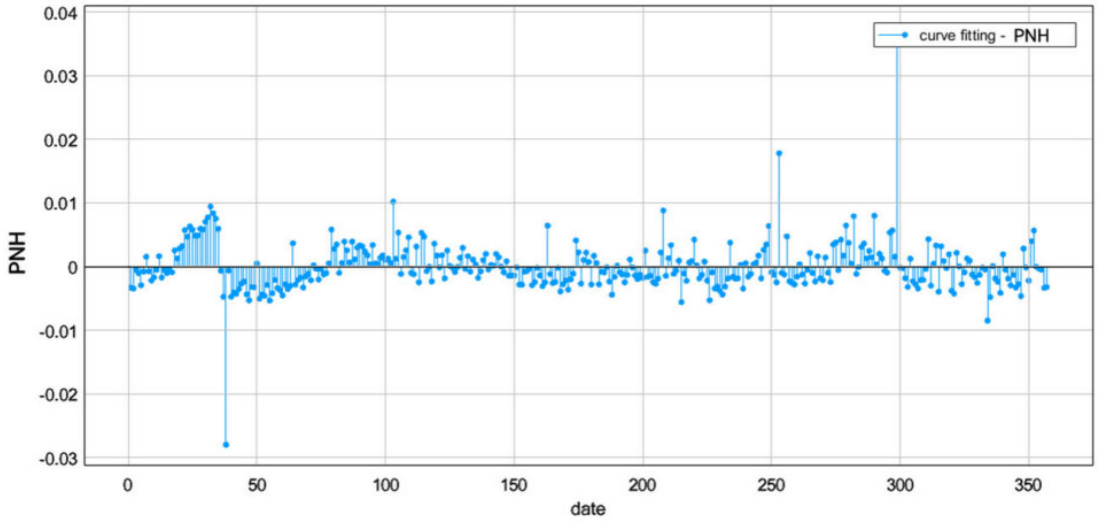

Figure 5. Deviation Value Chart.

Step 2: Establish BP neural network optimized by GA

This paper use the BP neural network optimized by GA to learn the nonlinear relationship between the extracted features and the percentage of (1, 2, 3, 4, 5, 6, X).

Firstly, this paper determined the number of neurons in the input layer, the six attributes of words and the value of PNH as the input data, the number of neurons in the output layer the percentage of (1, 2, 3, 4, 5, 6, X), the number of weights from the input layer to the hidden layer, the number of weights from the hidden layer to the output layer, and the number of variables to be optimized.

Secondly, this paper set the variables of the genetic algorithm to optimize the neural network. These variables were the number of individuals, the maximum number of genetic generations, the crossover probability, the variation probability, the initial population, and so on[7]. All value of mentioned parameters is shown in Table 3.

Table 3. Numeric Table

Hidden Num	Input Num	Output Num	Number of Weights from the Input Layer to the Hidden Layer	Number of Weights from the Hidden Layer to the Output Layer	Number of Variables to Be Optimized
25	7	7	200	175	390
Number of Individuals	**Maximum Number of Genetic Generations**	**Number of Binary Bits for Variables**	**Generation Gap**	**Crossover Probability**	**Mutation Probability**
40	100	10	0.95	0.7	0.01

Thirdly, this paper performed 100 iterations using the genetic algorithm to obtain the optimal solution and its ordinal number for each generation.

Step 3: Optimization

For the data of SA and RT, this paper changed 0 to 0.1 and 1 to 0.9 to make the Sigmoid activation function yield better values.

After the end of the iterations, this paper obtained the evolutionary graph (Figure 6), the optimal initial weights.

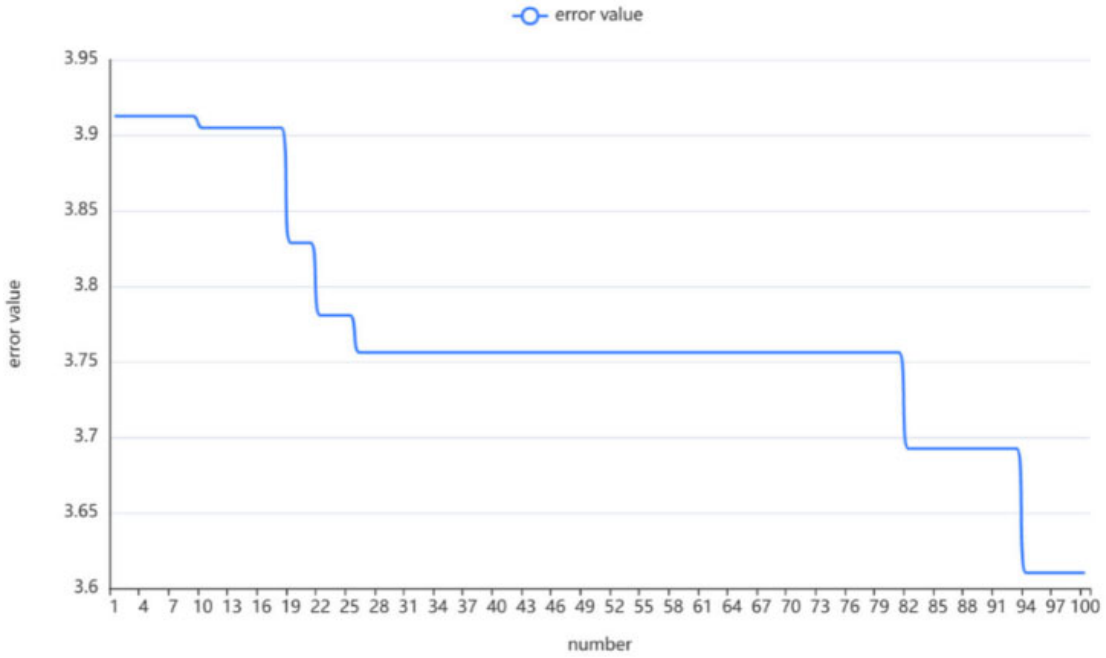

Figure 6. The Evolutionary Graph.

Step 4: Result Thresholding

The predicted value derived from the model may be a negative number. This is a fact that can't exist, since the percentage of (1, 2, 3, 4, 5, 6, X) cannot be negative. For this consideration, this paper threshold the output values by defaulting any negative number to 0.

3.3. Examples

To more visually demonstrate the predictive power of our model, this paper made predictions for the percentage of (1, 2, 3, 4, 5, 6, X) of EERIE on March 1, 2023, and LEMON on February 3, 2023.The result of those two words are shown in Figure 7.

Example 1: EERIE on March 1, 2023(Table 4).

Table 4. The predicted value of EERIE

IL	WM	CIE	FQ	SA	RT	PNH
15	133.97	1.668763033	9.86E-07	0.9	0.1	0.0923

Example 2: LEMON on February 3, 2023(Table 5).

Table 5. The predicted value of LEMON

IL	WM	CIE	FQ	SA	RT	PNH
11	108.75	2.08530507	4.95E-06	0.1	0.1	0.0958

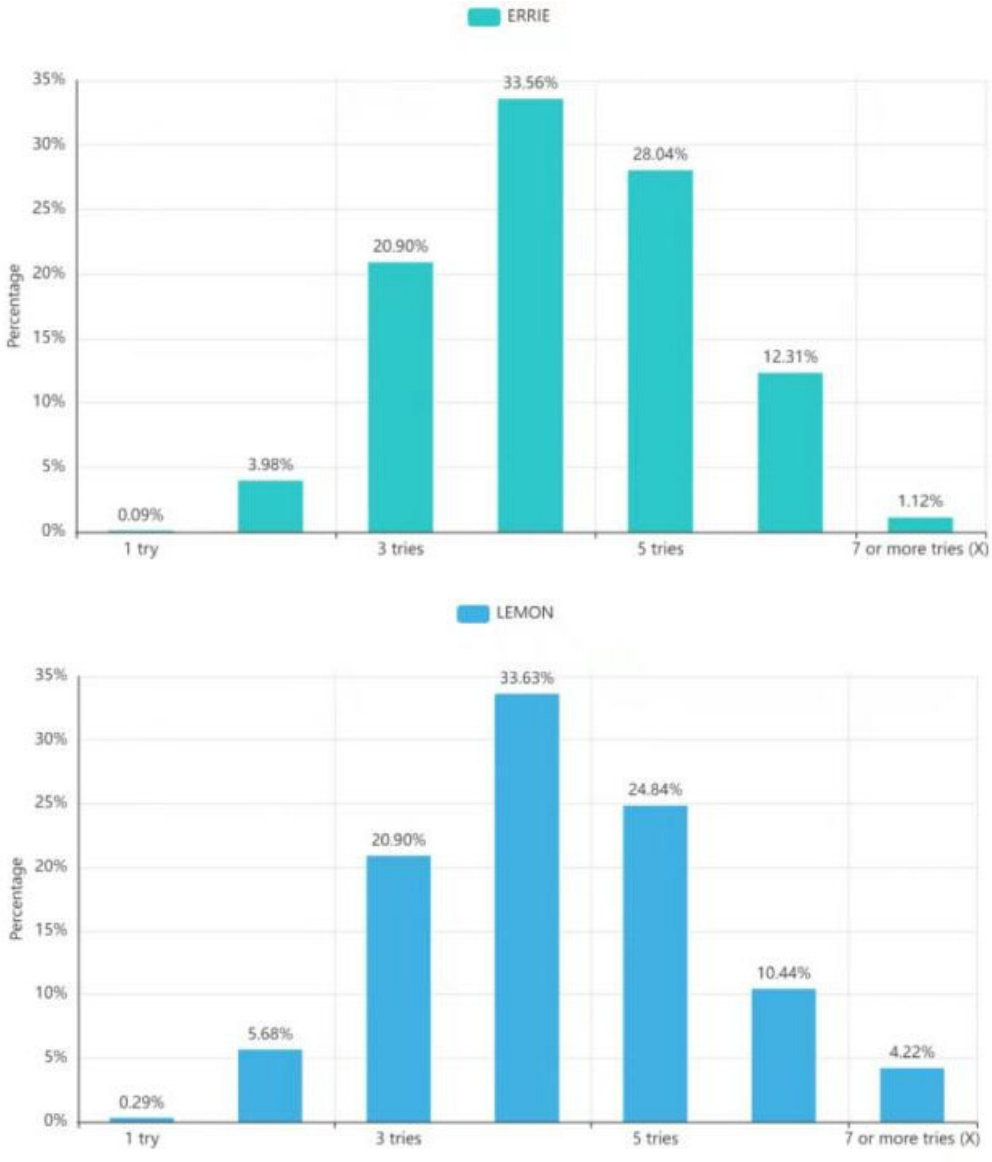

Figure 7. The predicted value of EERIE and LEMON

3.4. Accuracy and uncertainty

(1)Accuracy

The model evaluation results are shown in Table 6. As we can see, in the training set, the value of Accuracy Rate, Recall Rate, Precision Rate, and F2 are all close to 0.86, which means the model performs well in fitting the data. Accuracy Rate in the test set and the valid set is above 0.785, the value of Recall Rate is above 0.758, the value of Precision rate is above 0.754, and the F1 value is above 0.75. As a result, the model performs well in predicting.

Table 6. Model Evaluation Results

	Accuracy Rate	Recall Rate	Precision Rate	F2
Trained Set	0.859	0.859	0.869	0.864
Cross Validation Sets	0.803	0.803	0.771	0.787
Test Set	0.785	0.758	0.754	0.756

Note: F2 is the reconciled average of the Precision Rate and Recall Rate.

(2)Uncertainty

Uncertainties associated with our model and predictions are discussed.

As the number of participants in the game rises and players gradually gain experience by participating in the game, players may acquire game skills that improve the accuracy of their guesses and that uncertainty is independent of the difficulty of the words themselves. In addition, if a word has become a buzzword at a certain time or players use improper means, it may also affect the analysis of the model.

4. Model Evaluation and Discussion

In the following sections, we describe the advantages and disadvantages of our model.

4.1. Strengths

We have done enough visualization of the algorithms and results in this paper for easy understanding.

The predictive model uses an exponential fit curve and takes into account the effect of the time factor. It maps the effect of the time factor onto the percentage. Based on the rules of the Hard Mode game, the Hard Mode increases the number of attempts by the player, so the difference in the percentage of attempts is affected by the percentage distribution of the attempts. This model incorporates the rules of the game[8].

The use of genetic algorithms to optimize traditional neural networks can better avoid getting trapped in locally optimal solutions.

The neural network is used to capture the relationship between the various types of features learned for the output, and the optimal training principle is used to repeat the calculation and keep debugging the structure of the neural network until a relatively stable result is obtained. This can eliminate many human subjective factors and ensure the authenticity and objectivity of the classification and prediction results[9].

4.2. Weaknesses

The data set is too small to support the training procedure of complex neural networks. While it may work very well in the final result, there are still a lot of randomnesses. The neural networks would produce greater accuracy in identification if there was more data available. More data over time will only strengthen the model[10].

The model is unable to avoid human interference, such as variations in the number of attempts due to empirical skill or cheating.

There is a lack of data sets for verbal transmission records based on metrics such as word frequency and frequency of letter occurrences as criteria for word familiarity.

The "prediction model" does not take into account the psychology of the player when playing the game and may fixate on entering certain words to get more information.

5. Conclusions

The Wordle data report provides the basis for the establishment of data analysis and prediction model. However, there are some limitations in relying solely on BP neural network, such as easy to fall into local optimal solution, insensitive to data volume and data quality, etc. In this paper, BP neural network model based on genetic algorithm is used to construct neural network percentage prediction model. Firstly, this paper performs data cleaning and preprocessing on massive data. Secondly, this paper analyzes the number of attempts of future words according to the characteristics of words and the number of attempts of users. In addition, this paper also evaluates the model and discusses its advantages and disadvantages. The experimental results show that the neural network percentage prediction model has good predictability and has certain practical application value for natural language processing.

References

[1] Anonymous . New York Times buys Wordle and keeps it free 'initially'[J]. Computer Act!ve,2022(625).

[2] Yue T ,Qingbo K ,Xinping M , et al. Evaluation on power information data asset management system based on BP neural network[J]. International Journal of Thermofluids,2023,20.

[3] Engineering; Findings from School of Electrical Engineering and Computer Science Reveals New Findings on Engineering [A Genetic Algorithm (Ga) Approach To the Portfolio Design Based On Market Movements and Asset Valuations][J].Journal of Engineering,2020,

[4] Liu ,Guojin ,Miao , et al. Life Prediction of Residual Current Circuit Breaker with Overcurrent Protection Based on BP Neural Network Optimized by Genetic Algorithm[J]. Journal of Electrical Engineering & Technology,2022,17(3).

[5] Jun Z ,Yujie G . Correlation coefficient-based measure for checking symmetry or asymmetry of a continuous variable with additive distortion[J]. Communications in Statistics - Simulation and Computation,2019,51(5).

[6] Jun Z ,Zhuoer X ,Zhenghong W . Absolute logarithmic calibration for correlation coefficient with multiplicative distortion[J]. Communications in Statistics - Simulation and Computation,2023,52(2).

[7] M M ,M Q ,K B , et al.Comparison of a Genetic Algorithm Variable Selection and Interval Partial Least Squares for quantitative analysis of lactate in PBS.[J].Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference,2019,2019

[8] Chang C T ,Wu H M ,Lin Y . Sum of Exponential Model for Fitting Data †[J]. Engineering Proceedings,2023,38(1).

[9] Jialin L ,Tao P ,Shen Z , et al. Improved PID Controller Based on BP Neural Network[J]. Journal of Physics: Conference Series,2023,2479(1).

[10] Zhuo Z ,Fayu S ,Qingling L , et al. Establishment of the Predicting Models of the Dyeing Effect in Supercritical Carbon Dioxide Based on the Generalized Regression Neural Network and Back Propagation Neural Network[J]. Processes,2020,8(12).