Prediction and Classification Models of Wordle Data Set Based On ARMA, XGBoost

Xiangcheng Meng*, Chuanzhen Wang, Weichen Zang

Tongji University, Shanghai, 200092, China

* Corresponding Author Email: a3302478857@outlook.com

Abstract. This study intends to use machine learning method to solve the problem of predicting the number of word guessing and the difficulty of word guessing in natural language. Firstly, this study smoothed the data according to the COF method. According to the images of the sample autocorrelation function and the sample partial autocorrelation function, the ARMA model is constructed and the coefficient of the AMRA model is calculated, and the number of word contest people on March 1, 2023 is predicted to drop to [18002,20622]. Second, based on the seven attributes of the word, the XGBoost classification model is used to predict the value of the first attempt, and then the XGBoost prediction model is used to predict the results of the other attempts. Eerie distribution of different attempts to {0,8.890, 17.650, 28.191, 32.451, 11.349, 1.483}. The average accuracy of the model on the test set is 79.4%.

Keywords: Natural language, Prediction, COF, ARMA, XGBoost.

1. Introduction

In the field of Natural Language Processing (NLP), Word Prediction is an important task which aims to predict unknown words based on contextual information. With the rapid development of the Internet and digital technology, huge amounts of text data are emerging, and how to efficiently utilize these data has become one of the hotspots of people's attention. And word guessing, as a basic task of NLP, has a wide range of applications in many application areas, such as speech recognition, machine translation and question and answer systems.

In previous studies, the word guessing task mainly relies on traditional rule-based methods. These methods perform word guessing by constructing various rules and models, such as n-gram models and conditional random fields. However, due to the complexity and diversity of languages, these traditional methods have certain limitations when dealing with large-scale datasets, which make it difficult to meet the needs of practical applications.

In recent years, with the rapid development of Machine Learning (ML) technology, machine learning-based methods have gradually become the mainstream means to solve natural language processing problems. Machine Learning is able to learn patterns and laws from data, and predict and reason about unknown situations by building models. This data-driven approach shows great potential in word guessing tasks.

However, there is relatively little research on the number of people and the difficulty of guessing in word guessing tasks. The issue of number of people focuses on how many specific word samples are needed and the number of people involved in them, while the research on guessing difficulty focuses on the difficulty of word guessing in different contexts. The number of people involved in word guessing and the guessing difficulty directly affect the design of the algorithmic model and the evaluation of its effectiveness. Therefore, an in-depth study of the number of people and the difficulty of word guessing in word guessing tasks is of great theoretical significance and practical application value.

This thesis aims to study the number of people and guessing difficulty in word guessing using machine learning in natural language processing. By analyzing the existing research results and methods, we explore the way to solve the word guessing task using machine learning technology and propose a word guessing model based on machine learning. The performance and effect of the model are verified through experiments to further understand the relationship between the number of people

and the guessing difficulty and the design of the algorithm in the word guessing task, and some outlooks on the future research direction are proposed.

In the research of this thesis, we expect to provide scholars and researchers in the field of natural language processing with a new way of thinking and methodology to promote the development of word guessing tasks and provide more effective and intelligent solutions for practical applications. Through in-depth research on the number and difficulty of word guessing, we expect to further improve the accuracy and efficiency of the word guessing task and promote the wide application of NLP technology in various fields.

2. Modeling Preparation

2.1. Basic Assumptions

Assumption 1. Wordle's operators and business model will not change significantly.

Reason 1. Looking at the number of word guessing reports over time, we found that the number of reported results increased rapidly between January 17 and February 3, and then decreased significantly after February 3. This was most likely caused by the acquisition of the word guessing by the New York Times in early February, which caused the player community to question whether the word guessing could remain free [1]. After ten months of operation, Wordle's operation has been relatively stable. Its operators and operating models are less likely to change significantly. In order to ensure the stability of our forecasts and improve the accuracy of our forecasts, we assume that Wordle's operators and operating models will not change significantly(Fig.1).

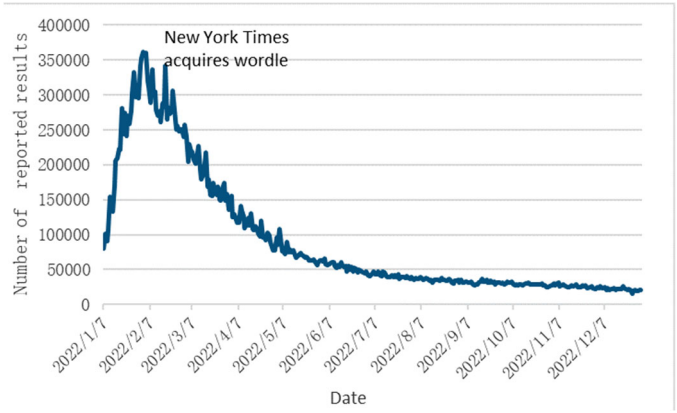


Fig 1. Number of reported results in relation to time

Assumption 2: The player's problem-solving level remains fixed.

Reason 2: Since the launch of Wordle, the number of people sharing has gradually declined and leveled off after reaching the peak in February 2022, and the player base is basically fixed. In a word guessing, the ratio of players who play for fun to those who play for problem solving remains roughly constant. It's safe to assume that Wordle players don't get much better or worse at playing. This assumption is also the premise for building a model based on past data to predict future word responses [2].

3. Modeling and Problem Solving for Word Guessing Number of Persons Reporting Quantities

3.1. Prediction of the number of reported results based on ARMA

3.1.1 Problem analysis

The first Problem to solve is to construct a model to account for the change in the number of reported results and determine a forecast range for the number of reported results on March 1, 2023. We first preprocessed the data and then built the ARMA model to solve the task.

3.1.2 Data preprocessing

According to Wordle's rule, we eliminate the data given for words that are not 5 letters or meaningless. Then we do several cubic spline interpolation [3] fill in the time series data

We have assumed that Wordle's business model will not change significantly in the future, and its business model will be significantly different from what it was before the New York Times acquisition. The use of pre-acquisition data will significantly increase the error of the model, which is not conducive to the prediction of data. Therefore, there is good reason to use only post-acquisition data (after February 3, 2022) for the construction of our model.

The remaining data is a decreasing time series, and both the top data and the bottom data differ greatly from the average value. Therefore, the traditional three sigma rule and boxplot method are not suitable for eliminating outliers in these data. We finally choose the COF（Connections-based Outlier Factor）[4] method to delete the abnormal data.

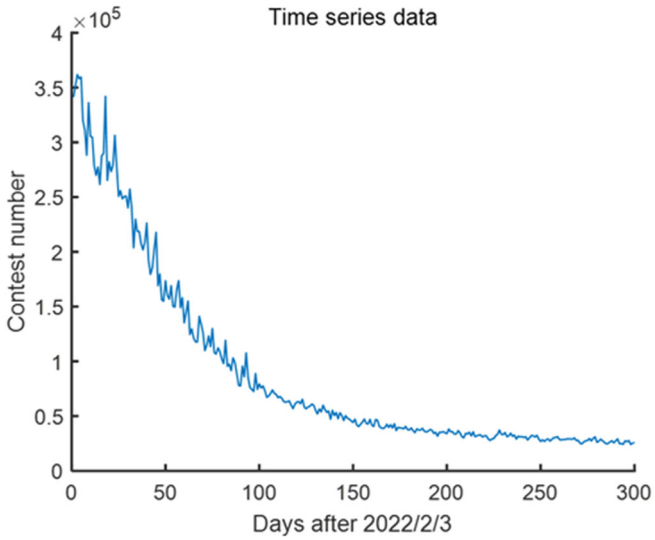


Fig 2. Data of this study

3.1.3 The establishment of ARMA model

Autoregressive moving average model(ARMA) [5] is an important method to study time series.

Correlation analysis between word attributes and the proportion of difficult patterns. This method is a typical method to study rational spectrum of stationary stochastic processes and is applicable to a large class of practical problems.

Basic principle of model: The data sequence formed by the prediction index over time is regarded as a random sequence. The dependency of this group of random variables reflects the continuity of the original data in time. On the one hand, it is affected by the influencing factors; on the other hand, it has its own change rule. Suppose that the influencing factors are $\delta_1, \delta_2 \ldots \delta_k$, by regression analysis:

$$\Gamma t = \beta 1\delta 1 + \beta 2\delta 2 + \cdots + \beta p\delta p + \Theta t \quad (1)$$

where Γ_t is the observed value of the prediction object, and Θ_t is the error. Γ_t, as a forecasting object, is affected by its own changes, and its rule can be reflected by the following formula:

$$\Gamma t = \beta 1\Gamma t - 1 + \beta 2\Gamma t - 2 + \cdots + \beta p\Gamma t - p + \Theta t \quad (2)$$

The error term has a dependency relationship in different periods, which is expressed by the following formula:

$$\Theta t = \varepsilon t + \alpha 1\varepsilon t - 1 + \alpha 2\varepsilon t - 2 + \cdots + \alpha q\varepsilon t - q \quad (3)$$

Thus, the ARMA model expression is obtaine

$$\Gamma t = \beta 0 + \beta 1\Gamma t - 1 + \beta 2\Gamma t - 2 + \cdots + \beta p\Gamma t - p + \varepsilon t + \alpha 1\varepsilon t - 1 + \alpha 2\varepsilon t - 2 + \cdots + \alpha q\varepsilon t - q \quad (4)$$

Next, we follow the process to model the AMRM of this problem. Graph sample autocorrelated function and sample partial autocorrelated function(Fig. 3 and Fig. 4).

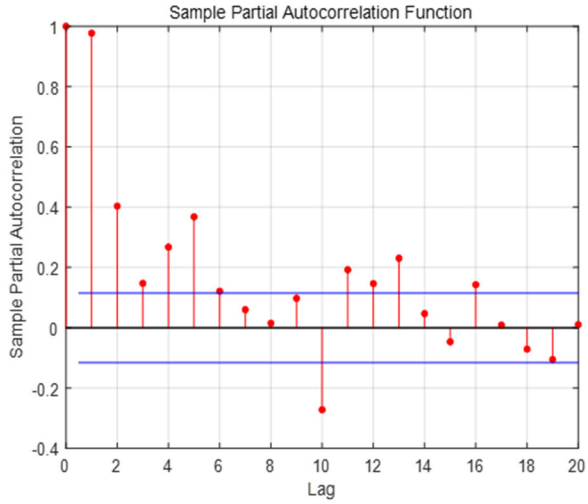


Fig 3. Sample Autocorrelated Function

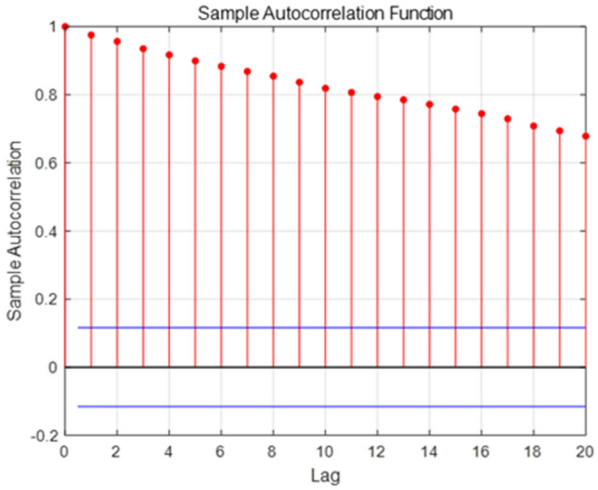


Fig 4. Sample Partial Autocorrelated Function

According to the image, it can be seen that it is reasonable for us to choose ARMA model for prediction.

And then we use the AIC criterion to find the best order of p and q. We plot the order heat map as followed(Fig.5).

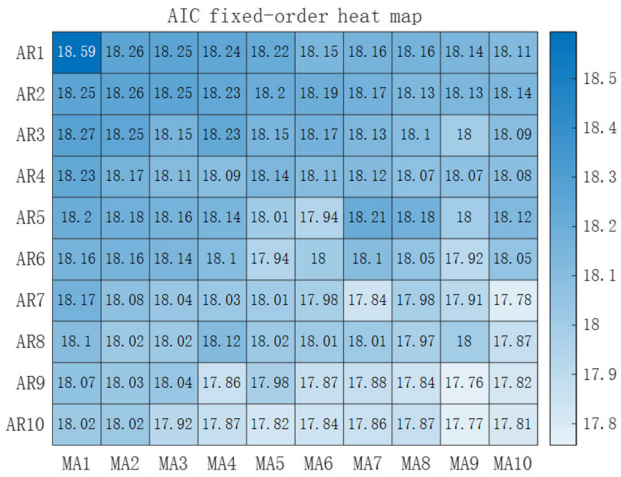


Fig 5. AIC Fixed-order Heat Map

We can easily see from the Figure 5 that AIC is the minimum when the order of AR and MA are both 9. So let's make p and q equal 9. After calculation, the model is constructed as follows:

$$\Phi(\Theta) = 1 + 0.762\Theta - 1 + 0.3251\Theta - 2 + 0.2959\Theta - 3 + 0.2145\Theta - 4 - 0.2762\Theta - 5 - 0.4326\Theta - 6 - 0.3566\Theta - 7 - 0.588\Theta - 8 - 0.7281\Theta - 9 \quad (5)$$

$$\psi(\Theta) = 1 + 1.235\Theta - 1 + 0.999\Theta - 2 + 0.9216\Theta - 3 + 0.8069\Theta - 4 + 0.6208\Theta - 5 + 0.3166\Theta - 6 + 0.08554\Theta - 7 - 0.3868\Theta - 8 - 0.6585\Theta - 9 \quad (6)$$

$$\Phi(\Theta)\Gamma(t) = \psi(\Theta)et \quad (7)$$

3.1.4 Phenomenon interpretation, error measurement and data prediction

Verify the model using a test set. Below is an image of what the model predicts and the error of the test set(Fig. 6):

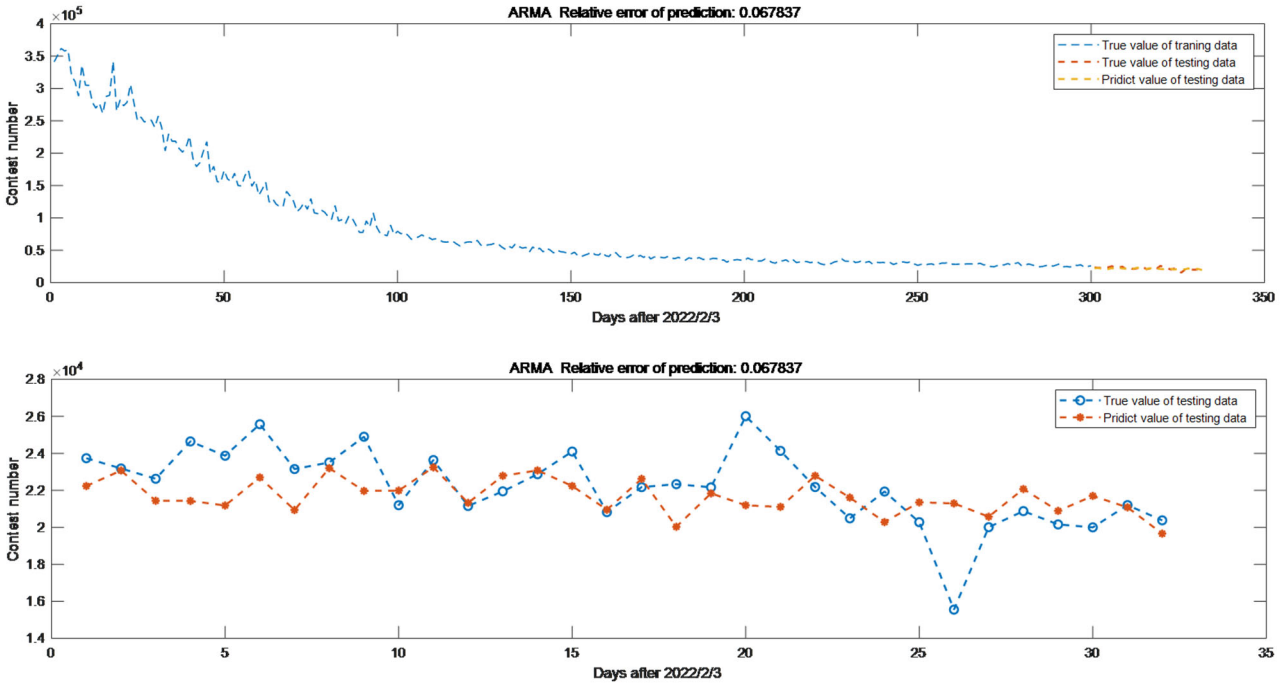


Fig 6. Prediction and Error

Based on the model, we can see an overall decline in the number of reported results since February 2022. The rate of decline slowed. The number of reported results will be around 20,000 for a long time. We think this is due to the lack of variation in Wordle over time. When Wordle was popular on social media, there were a lot of people playing Wordle, but as the popularity of Wordle waned, those same people stopped sharing their reports. The few who are really interested in the word guessing play it for a long time. That is why the number of reports decreased significantly, then slowly decreased and leveled off.

The relative prediction error is 0.067837, which fully demonstrates the accuracy of our model.

The model predicts 19,312 reported reports on March 1, 2023. Considering the relative error of the model, we believe that the prediction interval for the number of reported results on March should be 18002~20622.

3.2. Relationship between part of speech and percentage of scores reported that were played in Hard Mode

After removing words of length 5 or less, there are 356 words left in the data set. The rest of the data set includes nouns, verbs, adjectives, etc. According to statistics, there are 158 nouns, 123 verbs and 61 adjectives. Classify and calculate the average difficulty of these three types of words as follows(Table I):

Table 1. Percentage of Scores Reported That Were Played in Hard Mode

Part of Speech	Number	Percentage of Scores Reported That Were Played in Hard Mode
noun	158	5.0321%
verb	123	4.673%
adjective	61	7.329%

Apparently, when the word of the day is an adjective, percentage of scores reported that were played in hard mode is the highest; when the word of the day is an verb, percentage of scores reported that were played in hard mode is the lowest. There is an obvious relationship between part of speech and percentage of scores reported that were played in Hard Mode.

4. Word Difficulty Prediction Modeling

4.1. Problem analysis

Our goal is to predict the distribution of reported results for a given word using information we already know can be used to distinguish words. We extracted seven features of words of length 5, which will be used to predict the percentage of seven reports.

4.2. Data preprocessing

According to Wordle's rule, we eliminate the data given for words that are not 5 letters or meaningless. Unlike solving problem 1, we did not remove the data before the New York Times acquired Wordle, nor did we need to interpolate.

We also normalized [5] the proportion of attempts so that the sum is 100

4.3. Word feature extraction

(1) Distributions of Words

Obviously, the frequency of occurrence of words affects the player's choice, and we looked up the word corpus [6] and counted the frequency of the 356 words in the table.

(2) One of the more obvious indicators used to distinguish between words is the number of letter repetitions in a word. If a letter occurs in a word for three times, like ‘eerie’, or if two different letters occur in a word that are repeated once each, like the form of ‘aabcc’, then we consider both cases to be letter repetitions twice. We calculated the number of letter repetitions in each word through python string processing techniques.

(3) Since each word consists of five letters, we counted the probability of each type of letter appearing in each position from left to right. We collected a dataset of words of length 5 and counted the probability of occurrence of the letter in positions 1 to 5 from left to right. Let the letter be α, then $Xi(\alpha) = Ci(\alpha)/N$, where $Xi(\alpha)$ denotes the probability of occurrence of the letter α in the ith position, $Ci(\alpha)$ denotes the number of words in which the letter in position i is α among all words, and N denotes the total number of word corpus[7]. In the end, we have seven independent variables that can represent the characteristics of a word, which are frequency of words, repetitions, frequency of letters(x1, x2, x3, x4, x5)(Fig. 7).

Word	X1	X2	X3	X4	X5
slump	0.085893872	0.044749956	0.037237373	0.030175544	0.015826509
crank	0.061227556	0.061177472	0.083640097	0.062554779	0.022963464
gorge	0.048105577	0.138406831	0.08947487	0.028948489	0.113615306
query	0.004257131	0.072546515	0.063907044	0.05574337	0.055818496
drink	0.048731625	0.061177472	0.073072396	0.062554779	0.022963464
favor	0.036285779	0.184959808	0.019532717	0.066160819	0.053865224
abbey	0.07066837	0.010893246	0.027996895	0.138582125	0.055818496
tangy	0.053689931	0.184959808	0.078256079	0.028948489	0.055818496
panic	0.054391105	0.184959808	0.078256079	0.087421431	0.022963464
solar	0.085893872	0.138406831	0.067938797	0.100117697	0.053865224
shire	0.085893872	0.035033681	0.073072396	0.05574337	0.113615306
proxy	0.054391105	0.061177472	0.066887036	0.002929908	0.055818496
point	0.054391105	0.138406831	0.073072396	0.062554779	0.051335988
robot	0.044900208	0.138406831	0.027996895	0.066160819	0.051335988
prick	0.054391105	0.061177472	0.073072396	0.042245762	0.022963464
wince	0.025442616	0.10169534	0.078256079	0.042245762	0.113615306
crimp	0.061227556	0.061177472	0.073072396	0.030175544	0.015826509
knoll	0.042696517	0.028948489	0.066887036	0.053890266	0.049382716
sugar	0.085893872	0.072546515	0.033330829	0.100117697	0.053865224

Fig 7. Part of the Frequency of Letters from Position 1 to 5

4.4. Wordle's results distribution prediction model based on XGBoost

4.4.1 Reasons for choosing the XGBoost model

Our goal was to use the seven selected characteristics of each of the word to predict the percentage of the seven reported try times. In view of the large number of prediction targets, we will use the single objective model to predict the results of 7 cases respectively, and finally normalize the results.

Due to the small number of samples we have, we give up using BP neural network model to predict the results.

We conducted Pearson test on the independent variables and found that there was no significant correlation between the independent variables, so we gave up using the ridge regression model to predict the results.

The entropy method is too simple to construct the structure of the independent variable, and it is difficult to explain the relationship between the number of letters and the predicted results.

Random forest model is good at processing high latitude data, and random forest is a black box model, the structure of internal independent variables is not as simple as linear relationship. Since our independent variables are 7-dimensional data, we tend to use the random forest model to predict the reported results.

XGBoost model [8] is an upgraded version of random forest model, which can alleviate the excessive fitting problem of random forest model to a certain extent. We finally chose the XGBoost model to predict the outcome of the word eerie.

4.4.2 The Construction of XGBoost Prediction Model

The steps to fit the model are as follows:

Step1: Determine data set D and Chart tree set C.

$$D = \{(a_i, b_i) \mid a_i \in R^m, b_i \in R\} \tag{8}$$

$$C = \{f(x) \mid f(x) = w(r(x)), r(R^m) = s, w \in R^s\} \tag{9}$$

m is the dimension of variables. In this case, m is equal to 7; ai is a 7-dimensional vector; i depends on the number of samples; r represents the mapping of M-dimension vector to the number of leaf nodes s in a tree; w is an S-dimension vector, recording the score of each node.

Step2: Determine the bi predictive value formula

$$\hat{b}_i = \sum_{j=1}^{N} f_j(a_i) \tag{10}$$

$$\hat{b}_i^t = \hat{b}_i^{t-1} + f_t(a_i) \tag{11}$$

N represents the length of set C, i.e. the number of Chart trees. XGBoost will retain the previous t-1 simulation results when performing the t simulation prediction.

Step3: Set up the objective function

$$L^t = \sum_{i=1}^{n} l(b_i, \hat{b}_i) + \Omega(f_t) \tag{12}$$

$$\Omega(f_t) = \frac{1}{2}\mu \sum_{k=1}^{T} w_k^2 + \eta T \tag{13}$$

n represents the number of samples; $l(.,.)$ represents the error between actual value and predicted value; Ω represents the punishment for the complexity of the model; f_t represents a Chart tree in C, which can reflect the complexity of Chart numbers; μ is the complexity parameter; η is a fixed constant.

Step4: Use the greedy algorithm to select the optimal features of tree structure

$$S = \frac{1}{2}\left[\frac{\left(\sum_{i\in L_j} g_i\right)^2}{\sum_{i\in L_j} h_i + \lambda} + \frac{\left(\sum_{i\in R_j} g_i\right)^2}{\sum_{i\in R_j} h_i + \lambda} - \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda}\right] - \theta \tag{14}$$

I_j represents the set of samples on node j; L_j and R_j represent the sample set to the left and right of the segmentation point respectively; g_i and h_i are the first and second order gradients of training errors; θ represents the error caused by introducing a new node; When S is less than 0, node segmentation is stopped.

After many attempts, we found that the accuracy of regression prediction results for *T1, T4, T6* and *T7* was less than 70%. In order to improve the accuracy of the model, XGBoost classification model [9] was adopted to replace the regression model for the prediction of *T1*, because *T1* has a small value range. Then we have *T1, T2,T3,T5* with high precision. Due to the strong correlation [10] between *T1* to *T7* (Figs 8 and 9), we decided to use the results of *T1, T2,T3* and *T5* to predict *T4,T6* and *T7*.

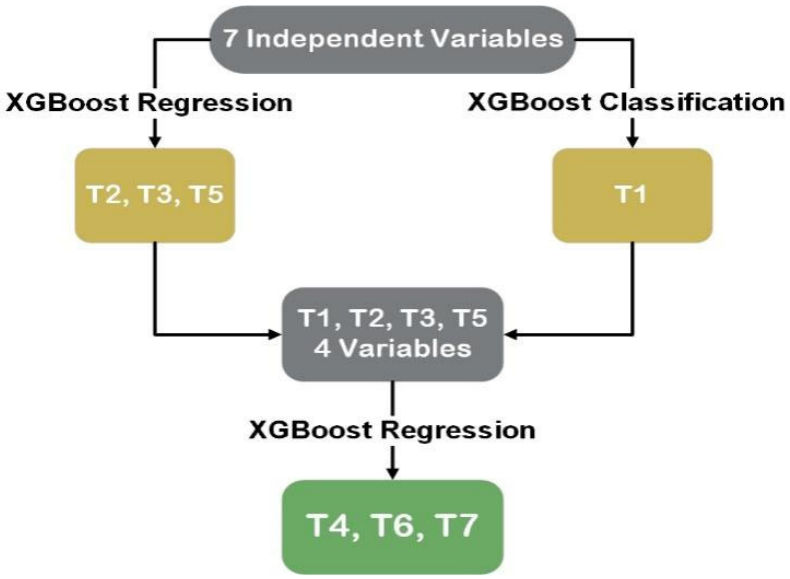


Fig 8. Flow Chart of Prediction

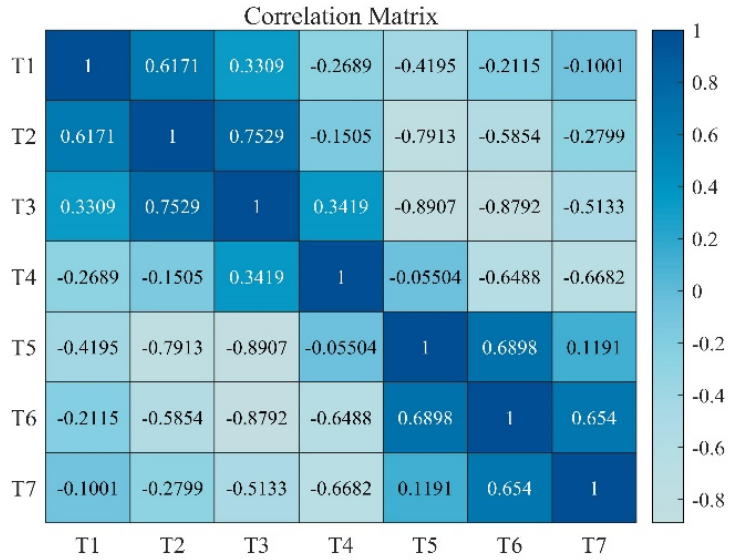


Fig 9. Pearson correlation test between different Tries

4.4.3 Prediction for EERIE

Here are the seven indicators of the word eerie:

$$Xe = \{0.035960233,0.121378309,0.08947487,$$
$$0.087421431,0.113615306,772484,2\} \quad (15)$$

From left to right are the probability of the letter e appearing in the first position of the word, the probability of the letter e appearing in the second position, the probability of the letter r appearing in the third position, the probability of the letter i appearing in the fourth position, the probability of the letter e appearing in the fifth position, the frequency of the word eerie in the word corpus and the number of letter repeats in the word.

Using this data, we predicted the reported results of the word eerie:

$$Te = \{0,8.979,17.827,28.473,32.776,11.462,1.498\} \quad (16)$$

After percentage normalization treatment:

$$T_e=\{0,8.890,17.650,28.191,32.451,11.349,1.483\} \quad (17)$$

We predict that when the word of March 1 is eerie, 0% of players will guess the word at first try, 8.890% of players will guess the word after the second try, 17.650% of players will guess the word on the third try, 28.191% of players will guess the word on the fourth try, 32.451% of players will guess the word on the fifth try, 11.349% of players will be able to guess the word on the sixth try, and 1.483% of players will try more than six times.

4.4.4 Model Accuracy

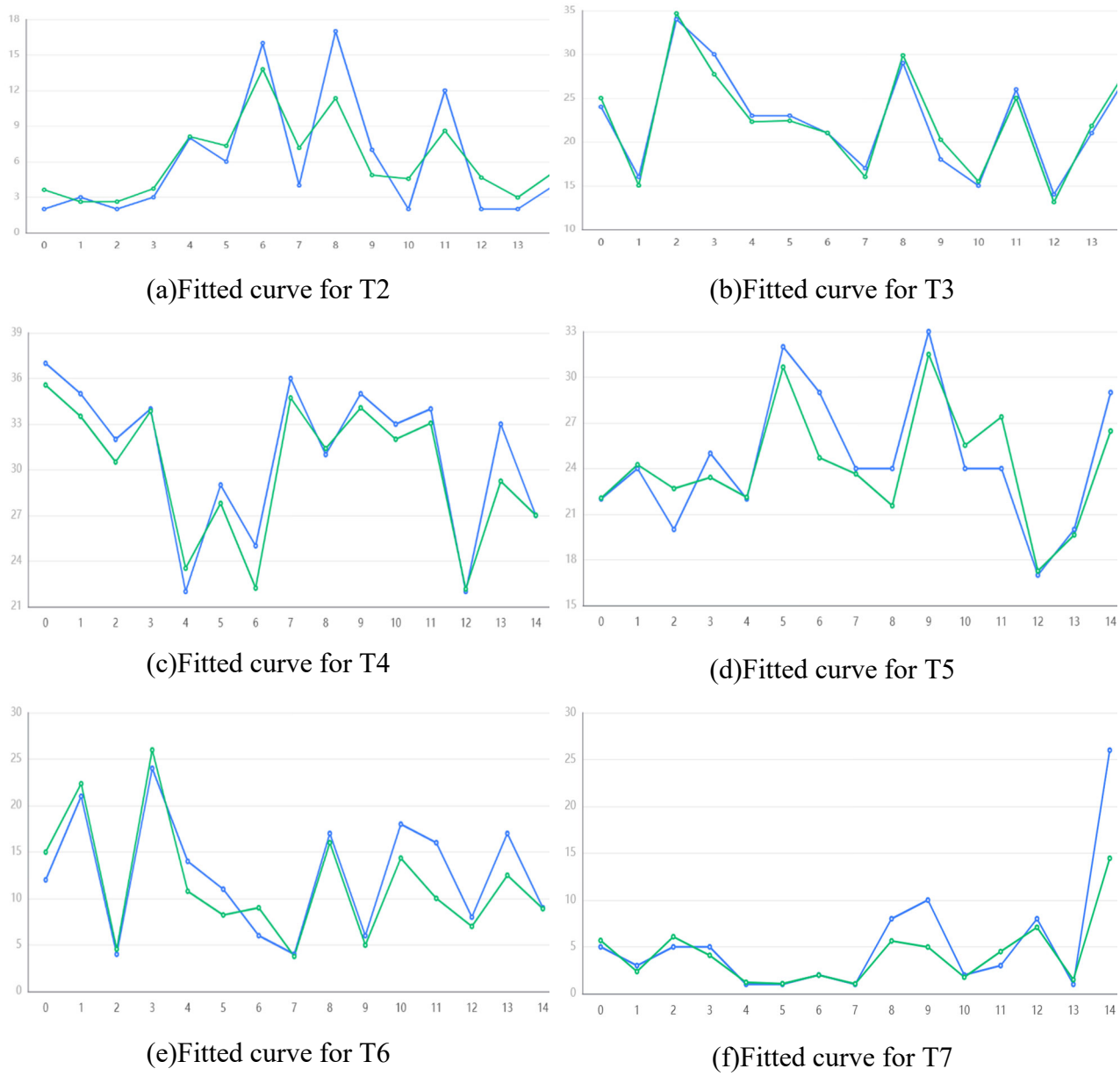
The accuracy of training set and test set of the model are shown in the table II below:

Table 2. Accuracy of Model

Indicators	T1	T2	T3	T4	T5	T6	T7
Training Set	88.4%	86.1%	89.2%	86.8%	89.0%	87.7%	88.7%
Test Set	76.4%	77.3%	86.8%	77.8%	78.2%	81.2%	77.9%

We are confident of the XGBoost model because the test set's accuracies are all over 75%.

The model's accuracy can also be reflected in these curves of test set below(Fig. 10). The test set has 15 samples in total and the sample's indexes are marked on the X label. The blue line represents real values and the green line represents the predicted values.(As T1 was predicted by XGBoost classification model instead of regression model, it does not have a curve for test set.)



(a)Fitted curve for T2

(b)Fitted curve for T3

(c)Fitted curve for T4

(d)Fitted curve for T5

(e)Fitted curve for T6

(f)Fitted curve for T7

Fig 10. Fitted curve for T2, T3, T4, T5, T6, T7

5. Conclusions

A major change in Wordle operator has been discovered through a search of information. When solving the model, the early data inconsistent with the future Wordle operation is eliminated, which significantly reduces the relative error of prediction. We use COF method to eliminate the outliers of the time series and complete the data with cubic spline interpolation, which is helpful for the construction of ARMA time series. Word frequency was obtained by searching the corpus, and words containing only five letters were screened out, and the probability of each letter appearing in each position was re-counted. Our search and processing of data material is very specific to Wordle. This greatly improves the accuracy of our XGBoost prediction model.

References

[1] https://www.nytimes.com/2022/01/03/technology/wordle-word-game-creator.html

[2] Chen T, Chen G, Chen W, et al. Application of decoupled ARMA model to modal identification of linear time-varying system based on the ICA and assumption of "short-time linearly varying"[J]. Journal of Sound and Vibration, 2021, 499: 115997.

[3] Hong S H, Wang L, Truong T K. An improved approach to the cubic-spline interpolation[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 1468-1472.

[4] Feasel K. Connectivity-Based Outlier Factor (COF)[M]//Finding Ghosts in Your Data: Anomaly Detection Techniques with Examples in Python. Berkeley, CA: Apress, 2022: 185-201.

[5] Li R, Chen W, Xu W, et al. Prediction on the Value Trends of Bitcoin and Gold-on Account of ARMA Time Series Forecasting Model[J]. Acad. J. Comput. Inf. Sci, 2022, 5: 79-84.

[6] Singh D, Singh B. Investigating the impact of data normalization on classification performance[J]. Applied Soft Computing, 2020, 97: 105524.

[7] https://www.kaggle.com/datasets/rtatman/english-word-frequency

[8] Guijun Yang, Xue Xu, Fuqiang Zhao. A user rating prediction model based on XGBoost algorithm and its application[J].Data Analysis and Knowledge Discovery,2019,3(01):118-126.

[9] Behera D K, Das M, Swetanisha S, et al. Follower link prediction using the XGBoost classification model with multiple graph features[J]. Wireless Personal Communications, 2021: 1-20.

[10] Obilor E I, Amadi E C. Test for significance of Pearson's correlation coefficient[J]. International Journal of Innovative Mathematics, Statistics & Energy Policies, 2018, 6(1): 11-23.