Data Analysis of the Wordle Game: Insights and Predictive Models Based on Twitter Data

Jiajun Jin [1,#], Buyi Geng [1,#,*], Sida Wu [2,#]

[1] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710000 China

[2] College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an, Shaanxi, 710000, China

* Corresponding Author Email: uyikatarina03@gmail.com

[#] These authors contributed equally.

Abstract. In 2022, the game "Wordle" gained immense popularity worldwide as players faced the challenge of guessing a five-letter word within six attempts, accompanied by feedback. This paper presents an extensive analysis of "Wordle" based on data mined from Twitter during the period from January 7 to December 31, 2022. The primary objective is to explore the game's dynamics and player engagement comprehensively.To achieve this, a sophisticated time-series model was developed to effectively track the fluctuation in the number of players. The model highlights an initial upward trend, reaching its peak on February 3, followed by a gradual decline and eventual stabilization, reflecting the sustained allure of the game. Additionally, this paper leveraged an XGBoost model to predict the distribution of player attempts, exhibiting remarkable accuracy, particularly for attempts ranging from three to six. This research demonstrates the powerful impact of data science in decoding intricate game dynamics and player behavior. Moreover, it emphasizes the fusion of gaming, data analytics, and social media as an exciting frontier for future research. The study's findings provide valuable insights into the gaming community's preferences and the underlying mechanisms that drive user engagement in digital gaming platforms.

Keywords: XGBoost Model, ARIMA, Prediction Model, Big Data.

1. Introduction

Wordle is a popular puzzle game in which players must guess a five-letter word in six or fewer attempts. Each guess must be a valid English word, and the player receives feedback in the form of colored tiles after each guess. A yellow tile indicates that the letter is in the word but in the wrong place and a green tile indicates that the letter is in the word and in the correct place while a gray tile indicates that the letter is not in the word at all.

The New York Times offers Wordle puzzles each day, which players can play in either normal mode or hard mode, with the latter being more difficult because players must use the correct letter in subsequent guesses. We are provided with data files for the daily Wordle results from January 7 to December 31, 2022. These data include date, contest number, solved word, number of results reported, number of players using hard mode, and the percentage of players who guessed the word in one, two, three, four, five or six attempts, or who were unable to solve the puzzle.

2. Data processing and visualization

In this study, we conducted a comprehensive analysis of Wordle gameplay data, which was sourced from Twitter between January 7 and December 31, 2022. The data included the date, contest number, the solved word, the number of reported results, the number of players using hard mode, and the percentage of players who correctly guessed the word in one to six attempts or were unable to solve the puzzle.

The first step in data processing was data cleaning, ensuring the validity of our dataset. We filtered out words that did not contain exactly five alphabetic characters, as per the game rules. An exception

in our data was the word "study," which showed significantly lower results than expected, especially in the hard mode. This data point was identified as an outlier and removed.

Next, the cleaned data were arranged in ascending chronological order. This ordering is crucial in time-series analysis, enabling the identification of trends, detection of patterns, and conduct of a more precise analysis.

To answer our research questions, we constructed a time-series model and used an XGBoost model. The former allowed us to explain and predict the fluctuating number of players over time, while the latter enabled us to predict the distribution of player attempts with high accuracy for three to six attempts. The time series plot is shown in Figure 1.

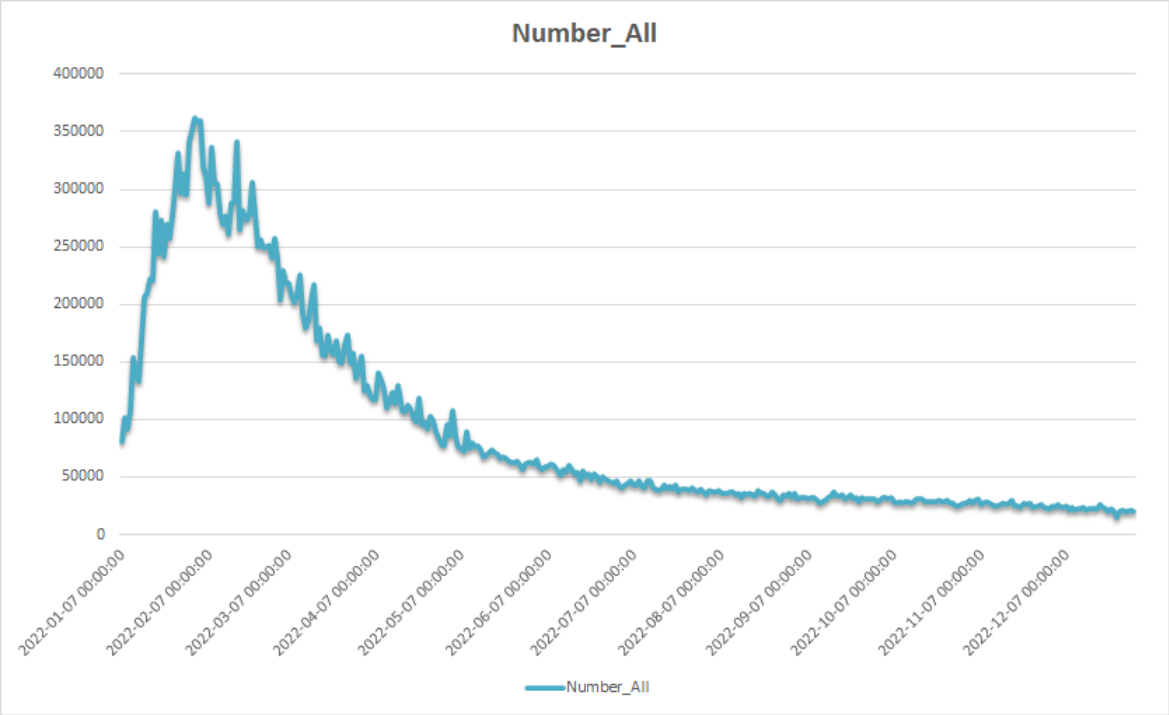

Figure 1. Time series plot of number of reported results

3. Models and Results

3.1. Model I: ARIMA

The ARIMA model is based on the autocorrelation of time series data and describes the short-term memory of the series [1-2]. Therefore has the property of short-term prediction. According to the correlation of this case, after considering a variety of time series analysis models, we choose the ARIMA model to analyze and predict.

3.1.1 Model Establishment

Develop time series models to explain the quantitative changes in reported results and use the completed model to create a prediction interval for the number of results:

Perform ADF test, view the data comparison chart before and after the difference to determine whether the data are stable; View the model test table, test the model white noise according to the P value of Q statistic ($P>0.05$), and derive the model formula combined with the time series analysis chart for comprehensive analysis to obtain the backward prediction order results. After determining that the total number of reports fit the time series predicting model, the ARIMA model is used to forecast the data for the last 60 periods.

– **Stationarity Test**: ARIMA model requires the series to satisfy the smoothness, view the ADF test results and analyze whether it can significantly reject the hypothesis that the series is not smooth according to the analyzed t-value($P<0.05$).

– **Determination of p and q**: In the ARIMA (p, d, q) model, d is the number of differences that change the time series into a stationary series, p is the lag order of the series, and q is the lag order of the stochastic series. The principle of the model is to transform the non-stationary series $\{r_t\}$ into stationary sequence $\bar{r}_t$, and then using $\bar{r}_t$ as the dependent variable. The lagged term of $\bar{r}_t$ and the random error term at and the lagged term of at are used as independent variables and regression is performed.

The ARIMA (p, d, q) model has the form:

$$\begin{gathered} \bar{r}_t = r_t - r_{t-1} \\ \bar{r}_t = \phi_0 + \sum_{i=1}^{p} \phi_i \bar{r}_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \end{gathered} \tag{1}$$

Where a_t is a white noise sequence and both p and q are non-negative integers.

Based on the historical data, we calculate the BIC values of the model at different orders by a computer programming loop. The loop calculates the BIC values of the model at different orders and finds the order p and q that minimizes the BIC. After determining the optimal order, we perform parameter estimation and then calculate a prediction interval for the number of reported results on the next day.

– **Residual test**:The ARIMA model requires the model to have pure randomness, i.e., the model residuals are white noise, to determine the validity of the model. Use the Ljung-Box statistic Q(m) to test the proximity to a white noise.

$$Q(m) = T(T+2)\sum_{l=1}^{m} \frac{\hat{\rho}_l^2}{T-l} \tag{2}$$

When the p-value of the test is greater than 0.05, it means that the residual series â passed the test at the 5% confidence level and the model is sufficient to be used to build a dynamic model. passed the test at 5% confidence level, the model is sufficient to model the linear dependence of the data. The relation schema of attributes of words is shown in Figure 2.

• To judge if any attributes of the word affect the percentage of scores reported that were played in Hard Mode:

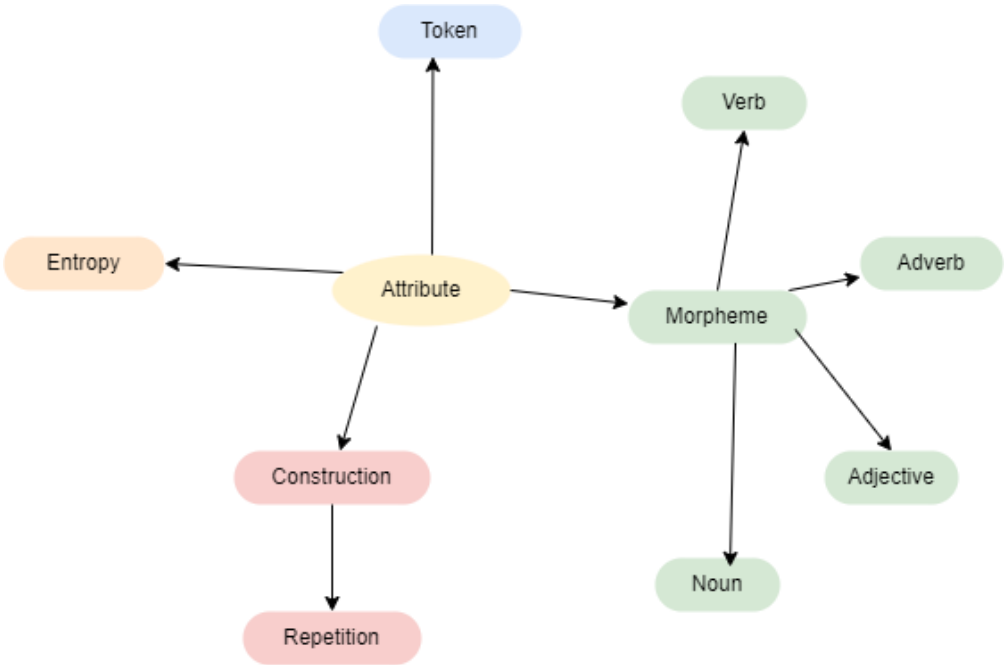

Figure 2. Attributes of words

– **Spearman Correlation Analysis**: It is a non-parametric indicator that measures the dependence of two variables. It evaluates the correlation of two statistical variables using a monotonic equation. The Spearman correlation coefficient is +1 or -1 if there are no repeated values in the data and when the two variables are perfectly monotonically correlated.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{3}$$

One way to determine whether the value of ρ for the observed data is significantly non-zero is to calculate the probability that it is greater than r as the original hypothesis and test it using a hierarchical ranking test. The advantage of this method is that it takes into account the number of data in the sample and the risk in using the sample to calculate the rank correlation coefficient.

– **Attribute 1**: Token: According to linguistics, the formal definition of a token is "a separate occurrence of a linguistic unit in speech or writing in contrast to the type or category of linguistic unit it represents" [3].

Tokenization is the process of breaking down text into smaller pieces called symbols [4]. These small pieces can be sentences, words or subwords. For example, the sentence "I won" can be tokenized as two word tokens "I" and "win".

Tokens used are from Corpus-based English-Chinese word formation comparison and translation — A case study of English derived words [5].

44 common letter combinations are selected and searched in words, as shown in Table 1, with 1 added for each occurrence, and the final result is the number of words containing the letter combinations.

Table 1. Most Frequent Tokens

ai	ay	ei	ey	ee	ea	oo	ar	ew
or	ight	gh	ph	qu	ch	sh	gi	aw
ge	ir	ur	er	tr	dr	tw	dw	oi
ul	ai	ti	ci	si	le	oa	ou	oy
au	ous	th	are	ear	ere	air	ow	

– **Attribute 2: Entropy of information:** When there are fewer known letters in a word, the harder it is for us to predict the rest of the letters, and the more letters we know, the easier it is to predict the rest of the word. This is what we mean by decreasing degrees of freedom. Using the information entropy argument, when we know nothing about a word, it has a higher entropy value, and when we know more letters, the remaining words have less degrees of freedom, and the information entropy is decreasing with it [6-7].

The lexicon of the wordle game is called up and the probability of each case is calculated for each word in the given sample, which in turn calculates the entropy, and then the expectation of the entropy of each word in all cases is calculated as the entropy of that word. At this point, We introduce the concept of information entropy to solve this problem,

$$\sum p \times \log_2 \frac{1}{p} \tag{4}$$

as well as conditional entropy :

$$\begin{aligned} \mathrm{H}(\mathrm{Y} \mid X) &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x)}{p(x,y)} \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log p(x,y) + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log p(x) \\ &= \mathrm{H}(\mathrm{X},\mathrm{Y}) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathrm{H}(\mathrm{X},\mathrm{Y}) - \mathrm{H}(\mathrm{X}) \end{aligned} \tag{5}$$

In the above figure, X and Y are two different letters in a word, and the conditional entropy is the probability that if we know that X exists in this word, Y will also appear in this word.

$$\mathrm{H}(\mathrm{Y} \mid \mathrm{X}) = \mathrm{H}(\mathrm{X} \mid \mathrm{Y}) - \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) \tag{6}$$

We can also show that H(Y|X) <= H(Y), i.e., the information entropy of the event decreases knowing whether letter Y appears after another letter X. The equal sign holds when and only when X and Y are two completely independent events. This also corresponds to our common sense in life. 44 common letter combinations are selected and searched in words, with 1 added for each occurrence, and the final result is the number of words containing the letter combinations. The wordle game's lexicon is called, and the probability of each case is calculated for each word in the given sample, which in turn calculates the entropy, and then the expectation of the entropy of each word in all cases is calculated as the entropy of the word.

3.1.2 Model Solving & Analysis

• **Prediction Model**: Time series analysis (ARIMA) is based on historical data to predict future period data: the model has a goodness of fit of 0.982 and the model performs well. Based on the variable number_all, the system automatically finds the optimal parameters based on the AIC information criterion, and the model results in an ARIMA model (1,1,0) test table and based on 0-difference data, the model equation is as follows:

$$y(t) = -169.718 - 0.355 * y(t-1) \quad (7)$$

The predicted number of reported results for March 1, 2023 is 10368.1874, and the prediction interval is [10346.92,10389.46].

• **Correlation analysis of attributes:** After conducting correlation analysis of the two attributes we selected and the percentage of scores reported that were played in Hard Mode, as shown in Figure 3, we get a heat map of correlation coefficient [8]:

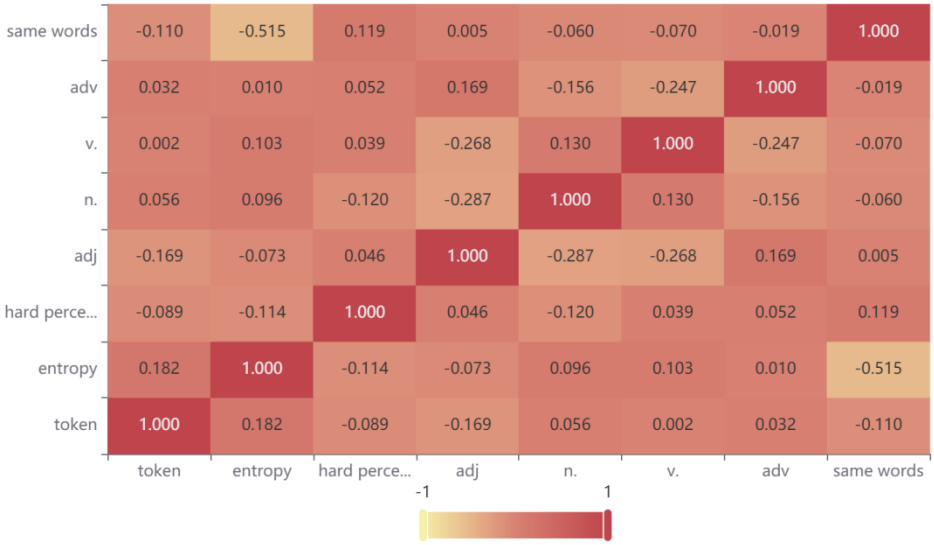

Figure 3. Correlation Analysis

After performing the correlation analysis, we can conclude that the percentage of reported scores under the total reported scores in the difficult mode is not correlated with the attributes of the words.

3.2. Model II:XGBoost

XGBoost (Extreme Gradient Boosting), an efficient gradient boosting decision tree algorithm. It improves on the original GBDT and makes the model much more effective [9]. As a forward addition model, the core of it is the integration idea - Boosting idea, which integrates multiple weak learners into one strong learner through certain methods [10]. The result of each tree is the difference between the target value and the prediction result of all the previous trees, and the final result is obtained by adding up all the results, so as to improve the effectiveness of the whole model. The derivation and establishment process of XGBoost is shown in Figure 4:

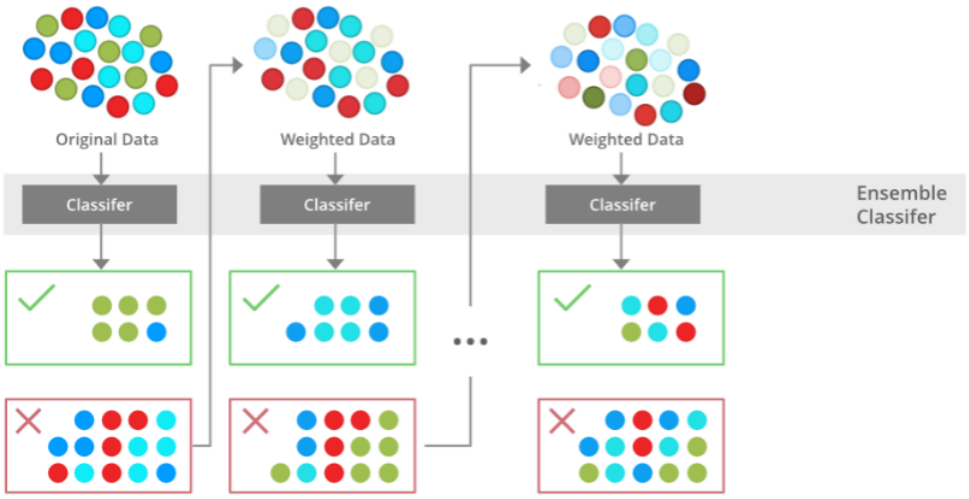

Figure 4. Workflow of XGBoost

Using regression chains to achieve multiple outputs of the model,the input of the second loop is the input and output of the first loop, and so on.

3.2.1 Model Establishment:

Following the idea of attributes in the first question, the letter combinations and information entropy of the word samples in the given data are extracted as sample features, and the percentage of three and more successful attempts are selected as sample labels. An XGBoost model was built to fit and the completed model was used to make predictions for 'eerie'.

• **Correlation Analysis:** Calculate the number of words containing letter combinations and information entropy, perform correlation analysis, and see the correlation between features and tags.

• **Predicting the percentage of correlation of 'eerie' words with the model:**The dataset was constructed and passed into the XGBoost model for training to obtain the average absolute percentage error between the predicted and true values, and the model was used to predict the percentage associated with the word 'eerie'.

3.2.2 Model Solving and Evaluation

• **Correlation Analysis:** Continue to perform Spearman correlation analysis on the characteristics and labels of the samples to obtain the correlation coefficient table as in Model I.It has a significance level of <0.05 and there is a correlation. It is shown in Figure 5.

Figure 5. Correlation coefficient heat map

The graph shows the value of the correlation coefficient in the form of a heat map, mainly through the color shades to indicate the magnitude of the value.

• **Build the dataset and pass it into the XGBoost model for training:**Multi-output regression prediction of XGBoost by RogressorChain's method. The average absolute percentage errors of the training and test sets are shown in Table 2.

Table 2. Mean absolute percentage error of the training and test sets

Number of Attempts	Training sets	Test sets
Three	0.032	0.310
Four	0.025	0.181
Five	0.040	0.197
Six	0.095	0.423
More	0.289	1.192

• **Predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date:** The number of tokens contained in eerie is calculated to be 2, and the entropy is 4.260. The percentage distribution of successful and unsuccessful guesses is 4.98, 13.81, 27.19, 26.63, and 15.40 for 3, 4, 5, and 6 attempts of eerie.

• **Uncertainties associated with your model and predictions:**

– **The Reliability of try_i:** The average absolute percentage error of the successful distribution of 3 to 6 guesses in the prediction results is small, and combined with the actual situation, the average person can guess the result in 4 guesses, so the result of about 4 attempts is the most credible.

– **Unequal probability of word occurrence:**In calculating the information entropy, we assume that each word has an equal probability of occurrence, which is obviously unrealistic. For example, words with w as the initial letter are more uncommon than words with s as the initial letter, and the wordle game requires that the answer must be a common word, so this introduces some uncertainty into our model.

4. Sensitivity Analysis And Robustness

4.1. Sensitivity Analysis of XGBoost

To analyze the sensitivity of the model, we choose different learning rates to calculate the absolute average percentage error of the prediction model. We choose each element of the set (0.01, 0.1, 0.5) in turn as the learning rate. The absolute average percentage error obtained is shown in Figure 6 and Figure 7. When the learning rate is low, the model does not converge and the absolute average percentage error of the test set is large. When the learning rate reaches 0.1, the model converges and the accuracy is high. When the learning rate continues to increase the error of the training set increases.

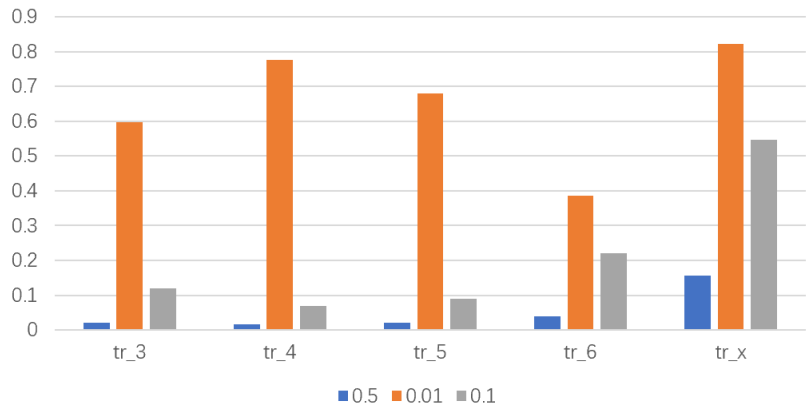

Figure 6. Absolute mean percentage error of the training set

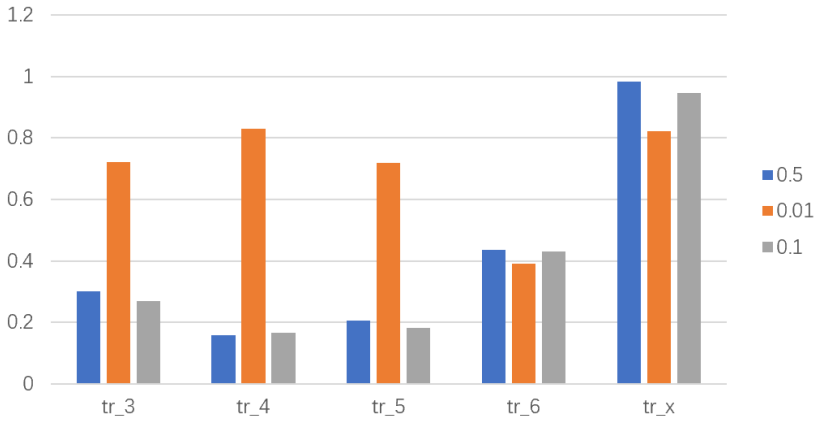

Figure 7. Absolute mean percentage error of test set

5. Evaluation of Strengths and Weaknesses

5.1. Strengths

1. Our model is an innovative model that adds the effect of tokens and information entropy to traditional analysis that only considered repetition and morpheme. This makes the model more relevant to reality.

2. The reliability is high, the model results are in high agreement with those of the actual experiments, and the research methods used are highly accurate.

5.2. Weaknesses and Further Improvements

1.The different weight of different letters is not carefully considered, They may vary in the occurance in the Wordle Lexicon.

6. Conclusions

In conclusion, our analysis of the game Wordle, mined from Twitter data between January 7 to December 31, 2022, has led to a number of insightful findings. We have successfully illustrated the game dynamics, player engagement, and attempt patterns, using a combination of time-series and XGBoost models.

We observed that the popularity of Wordle showed a significant increase at the onset of 2022, with player count peaking on February 3. Following this spike, the game's player count eventually reached a state of relative stability. This trend exhibits the temporal dynamics of the game's player base, highlighting the nature of viral phenomena in gaming and the temporal patterns of player engagement.

The application of an XGBoost model to predict the distribution of player attempts led to high-accuracy predictions for attempts ranging from three to six. This underscores the complex nature of player behavior in Wordle and suggests that machine learning techniques can be effective tools for understanding player behavior in games.

Moreover, our data processing steps effectively prepared the dataset for our analysis, providing a robust methodology for similar future studies.

The findings of this study can serve as a foundation for further research into player behavior in online gaming. They also provide game developers with insights that may be leveraged to optimize player engagement and gaming experience. Our research demonstrates the applicability of data mining and machine learning in the analysis of game dynamics, opening avenues for future interdisciplinary research.

References

[1] Kiarie J, Mwalili S, Mbogo R. Forecasting the spread of the COVID-19 pandemic in Kenya using SEIR and ARIMA models[J]. 2022, 7(2):10.

[2] Agustín Maravall.A CLASS OF DIAGNOSTICS IN THE ARIMA-MODEL-BASED DECOMPOSITION OF A TIME SERIES[J].2022.

[3] Mielke S J, Alyafeai Z, Salesky E, et al. Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP[J]. 2021.

[4] Fernandez-Herraiz C, Sara Esclapés, Prado-Dominguez A J. Tokens and Tokenization: Still a Gordian Knot for the Future of FinTech? [M]. 2020.

[5] Security S. Tokenization vs encryption: RSA touts tokens to reduce PCI DSS pain[J]. [2023-07-31].

[6] Carlos Fernández-Herraiz, Sara Esclapés-Membrives Antonio,Javier Prado-Domínguez.Tokens and Tokenization Still a Gordian Knot for the Future of Finance[J].2019.

[7] Chiarcos C, Stede R M. Linguistic Annotation || By all these lovely tokens. Merging conflicting tokenizations[J]. Language Resources & Evaluation, 2012, 46(1):53-74.

[8] Cambridge U C, Cohen P C. Applied Multiple Regression/ Correlation Analysis for The Behavioral Sciences[J]. Journal of the Royal Statistical Society Series D (The Statistician), 2003, 52(4). Chen T, He T, Benesty M. xgboost: Extreme Gradient Boosting[J].2016.

[9] Dutta R, Chen C, Renshaw D, et al.XGBoost automates the characterisation of reversibly actuating planar-flow-casted NiTi shape memory alloy foil[J].2021.

[10] Grislain N, Gonzalvez J. DP-XGBoost: Private Machine Learning at Scale[J].2021.