Quantifying the difficulty of word guessing based on lexical categorization

Shuhan Liu[a], Shuhan Xu[b], Yanqi Huang[*]

SILC Business School, Shanghai University, Shanghai, China

[a]Liushuhan0207@163.com, [b]ruby_ontheway@163.com,
*selinahuang2021@163.com

Abstract. This article focuses on the Wordle game, a word puzzle game that has become extremely popular on social media. To improve the user's gaming experience, three models have been proposed to solve the problem: a model for predicting the number of players, a model for predicting the number of attempts, and a model for classifying word difficulty. For the problem of predicting the number of players, the SIR (Susceptible Infected Recovered) model was proposed. The research findings demonstrate that the N-array tree model exhibits a certain level of effectiveness in predicting the distribution of player attempt counts. The frequency of player word guesses and the prevalence of vocabulary play a significant role in the prediction process. Finally, this paper contributes to the difficulty classification of words based on IE (Information Entropy) model, and the experimental results showed that comparing the historical data with the corresponding information entropy would obtain an absolute error of 17%, which has a high degree of confidence.

Keywords: Word Attribute Categorization; Susceptible Infected Recovered; N-array; Information Entropy;

1 Introduction

In Wordle, players guess a five-letter word within six tries with feedback. Prior studies explored the relationship between word properties and guessing difficulty but lacked quantitative analysis. This study uses Twitter data to analyze the impact of daily Wordle solutions on player guesses, forecasting player base fluctuations, predicting real-time attempts using word attributes, and categorizing word difficulty to understand how linguistic influence on guess attempts.

Recent research highlights machine learning's dependence on quality training data and its impact on model performance and fairness. Inadequate or biased data causes generalization issues, overfitting, and biases (Gao et al., 2020; Guo et al., 2017). Medical imaging studies reveal difficulties in detecting rare diseases and demographic variations due to limited representative data (Mehrabi et al., 2021). While techniques like data augmentation and knowledge distillation can aid small datasets, they can't replace comprehensive, unbiased training data (Shorten & Khoshgoftaar, 2019;

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_74

Stoean et al., 2020). Balancing model attributes like accuracy, calibration, and fairness depends on dataset quality and size (Sun et al., 2019; Wang et al., 2019). Thorough testing on diverse datasets is crucial to develop robust models and reduce the spread of erroneous correlations, measurement artifacts, and historical biases (Zhang et al., 2018). In summary, despite algorithmic advancements, machine learning still heavily relies on abundant, representative high-quality training data.

2 Model development

2.1 Susceptible-Infectious-Recovered (SIR) Model

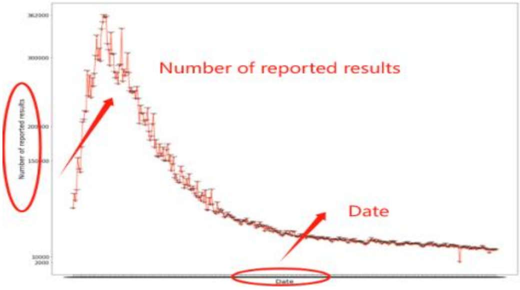

Fig. 1. Time- Number of reported results trend graph

As shown in Figure 1, weighted mean scores for 'tries' revealed fluctuating Wordle game difficulty, weakly associated with time. However, these fluctuations result from stochastic variability rather than direct temporal links, revealing no correlation with challenging mode's player count. To examine temporal-posting dynamics and predict posts, we propose applying the SIR model, which analyzes lexical attributes' interaction with achievement score proportions. Figure 2 illustrates similarities between post trends and the model's progression, driven by common sociodynamic traits. Extending the analysis, we categorize players as S (non-posting), I (playing and posting), and R (lost interest), adapting the SIR model to explore time-posting patterns in Wordle.

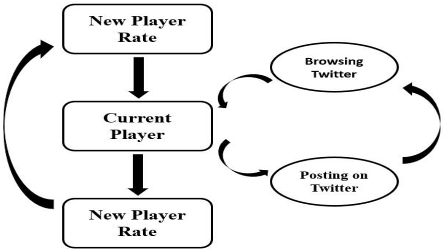

Fig. 2. Application of SIR in the context of this topic

2.2 N-array Tree model

The purpose of this model is to predict the distribution of word attempts, so we introduce a N-array tree data structure. The steps are as follows:

First of all, set the number of layers of the multi-tree to 7, and use answer as the root node of the multi-tree. Then, through bijection F, 12972 lexicon words map to the multi-tree's nodes. Utilizing backtracking, the final result becomes the root, and all possibilities unfold on the N-array tree. Rule-wise, if a letter is yellow/green in a guess, it must be in the next. Thus, bijection f must ensure similar word letters between adjacent multi-tree levels. The last step is to try to obtain random distribution by guessing words. If the player guesses the word 'W' as the correct answer in the first attempt, and W is located in the ith layer of the multi-tree. Each word in each node corresponds to its frequency of use, and the required probability P(X=m) should be the sum of all frequencies in the m layer. Then we have obtained the distribution of random variable X, m1,2,3,4,5,6,7.

2.3 Information entropy model

Using information entropy as a measure of uncertainty or information in a random variable is one of the fundamental concepts in information theory. According to information theory, information can be quantified in bits, where one bit can represent two possible states, such as 0 or 1. If the probability distribution of a random variable X is P(X=x), then the information entropy H(X) of X is defined as [4]:

$$H(X) = -\Sigma\ P(X=x)\ \log 2\ P(X=x) \tag{1}$$

The Σ represents the sum of all possible x, and log2 represents the logarithm with base 2. Higher information entropy indicates greater uncertainty in the random variable and more information conveyed. Conversely, a lower information entropy indicates less uncertainty in the random variable and less information conveyed [5]. Entropy in information theory is a fundamental concept that has been widely applied in areas such as data compression, communication encoding, and cryptography.

Step1.Compute all possible color permutations (green, yellow, gray) for each five-letter word.

Fig. 3 shows the three possible color permutations for a word when only considering i ts color in position.

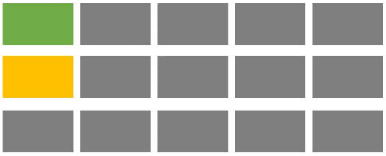

Fig. 3. Three possible permutations according to different

Step2.Calculate the probability corresponding to each possible permutation.
Example: the word 'weary' has a probability (p(🟩⬜🟨⬜⬜)=58/12972=0.0045), with all possible color arrangements calculated.

Step3. Express each probability as an entropy value.

1. Principle and feasibility expressed in terms of entropy value.
Represent observations in terms of entropy value (e.g., one bit for dividing the probability space into two parts).

2. Conversion of probability and entropy values.
In this case, the probability of the color arrangement in 324 calculated in step2 needs to be converted into an entropy value for each word.

3. Advantages of information entropy representation.
Emphasize the convenience of the information entropy representation, allowing for easy operations like addition and subtraction.

Step4. Calculate the initial entropy of each word.
The initial entropy of the word is a fixed value, defined as I.
Calculation is accomplished by adding the probability () of the color arrangement in 324 that corresponds to the word and the information Bits () provided by the arrangement. Here is the formula.

$$I = (n=1,2,3,4,5,6) \tag{2}$$

Step 5. Determine the difficulty expectation of each word:
Calculate the entropy reduction during various stages using word frequency data and the sigmoid function. Then determine the system entropy value after each phase. Last, repeat until the system entropy is zero.
This model presents a systematic approach to predicting the information entropy for the Wordle game, utilizing probability, entropy conversion, and employing sigmoid functions for smoother transitions. Step1-5 depict various stages of the model, and illustrate the three possible color permutations for a word, the probabilities for 'weary', and the Sigmoid Function Schematic, respectively.

3 Solutions

3.1 Part I: The Prediction of Reported Results

Based on the historical data, a point estimate of 19589 is obtained. The interval estimation [19115, 20064] can be calculated using the T distribution and standard deviation at the 5% significance level.

3.2 Part II: The Prediction of the Distribution of Tries

The primary solution to this issue relies on backtracking, utilizing an n-array tree structure in the process. Initially, the wordle answer becomes the tree's root, with "EERIE" as the chosen root for this case. The remaining tree adheres to the challenging mode rule. Furthermore, the projections of reported results are time-bound. As discussed in the first problem, the puzzle gained popularity, signifying increased participation of skilled players. To account for changing player numbers, a time-associated random variable 'e' is introduced. Backtracking records all potential outcomes as nodes in the n-array tree. Determining node depth requires knowledge of the player's initial guess count. Players tend to guess frequent words first. Consequently, frequently used words gain higher first-guess likelihood. Following this principle, node frequency dictates its weight. Represented as random variable 'X' for a player's score and 'Ski' for node weight in layer 'i', probability 'P(X=k)' is calculable through the equation.

3.3 Part III: Classifying Solution Words by Difficulty

In this part, we define the important concept throughout the text – the attribute of a word – by using Entropy. Entropy denotes to the uncertainty a word contains. A word with high information needs more steps to find the answer and vice versa. As shown in Figure 4, Using such tool to calculate and record all the word information available on the world's web sites to form a list. This list of information entropy provides a mean of 4.79, maximum value of 14 and minimum value of 1. If we randomly choose a word from the dictionary, it is assured that the information entropy of the word is within the interval of [1, 14]. To classify the solution word, the interval of [1, 14] is divided into 3 parts 0 to 3, 4 to 5 and 6 to 14, and entitled with easy, normal, and difficult. Under this standard, eerie with an entropy of 4 is a normal puzzle. Finally, we compare the historical data and corresponding information entropy, the absolute error of 17% is obtained, which provides a high confidence level.

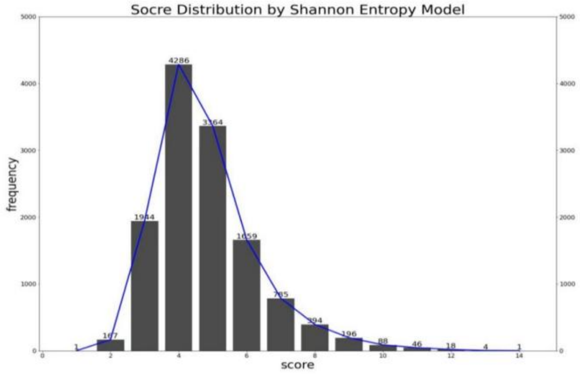

Fig. 4. Score Distribution by Shannon Model

4 Conclusion

This paper utilizes three models to analyze various aspects of the game experience. The SIR epidemiological model predicts player base changes by capturing player transitions over time. The N-array statistical model efficiently forecasts player behavior based on word attributes and historical data, offering simplicity and interpretability. The information entropy model categorizes word difficulty, providing linguistic insights. These theoretically motivated models, coupled with language pattern analysis and compartmental dynamics modeling, give a comprehensive quantitative window into the interplay between word properties and player behavior that machine learning approaches currently lack, highlighting the value of domain expertise in linguistic features categorization.

Reference

1. Gao, L., & Zhou, J. (2020). Transfer learning for EEG-based brain-computer interfaces: A review of progress made since 2016. IEEE reviews in biomedical engineering, 13, 28-37.
2. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321-1330). JMLR.org.
3. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
4. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.
5. Stoean, R., Stoean, C., Lupu, M., & Leordeanu, M. (2020). A systematic review on deep learning in medical imaging field: Fundamental, applications and challenges. Journal of imaging, 6(11), 120.
6. Sun, S., Cheng, Y., Gan, Z., & Liu, J. (2019). Patient knowledge distillation for bert model compression. arXiv preprint arXiv:1908.09355.
7. Wang, Y., Deng, Z., Pu, S., & Huang, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 563-574).
8. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).