$\mathbf{C}$ 

# Uncover the Hiddnd Secret in the Wordle Results

#### **Summary**

Since Wordle has become a popular puzzle game, it has accumulated a large amount ofdata. In this paper, we define a series of metrics and build several models to explore thehidden information in Wordle results.

First, after **preprocessing** the given data and analyzing the time series diagram of thenumber of reported results, we found that the changes can be divided into 3 stages. Toforecast the number of reported results, we developed a weighted optimization modelbased on ARIMA and BP neural network. The prediction interval is then given using the Bootstrap method. We packaged this process as **ARIMA-BP Interval PredictionModel Based On Bootstrap**. Thus, we finally predicted the interval prediction value obtained on March 1. 2023 at 95% confidence level to be about (19504.74, 20383.26)

Then, we defined 3 qualitative and 4 quantitative attributes of words and used them tobuild a **Multiple Linear Regression Model** with the percentage ofhard mode's playersWe found that the proportion will decrease by an average of 0.618 when the initial letterchanges from a vowel to a consonant while it will increase by an average of 0.017 for each one-unit increase in word internal distance.

After that, we made the percentage distribution prediction of the reported results based on **LSTM Model**. To ensure the percentage is around 100%, we first processed the component data using a **spherical coordinate transformation**. Then we use them asoutput variables, the 7 word attributes and number of results as input to train our LSTM model. The prediction of EERIE based on this are [2%,11%, 25%,24%,19%,14%.5%. We changed the model's parameters and added noise to do **sensitivity analysis**. Meanwhile, we introduced COV to measure the uncertainty of the model prediction and found that it is around 0.4. For **error analysis**, we use MSE, RMSE and R to measure the prediction accuracy, and their values are shown in Table 7.

We extracted 6 indicators: RDC, TE, SK,NFC, NON, and HL to measure the dificulty of words. We built a **GMM Clustering Model** based on these indicators and thus classifying 5 difficulty levels. We classified the word EERIE as dificulty level lll

In addition, by counting the frequency of each letter in five positions, we found S as theinitial letter has the most frequency and more specific statistical results are shown in Table 9. We also used the **Association Rule Model** based on **Apriori algorithm** tomine the word combination pattern in Wordle, Ideally, we found that the letters A.S Eand F,TL usually appear together in Wordle.

Finally, we evaluated and refined the model and reported the findings in a letter to thethe Puzzle Editor of the New York Times.

Keywords: ARIMA-BP,LSTM, GMM, Apriori Algorithm, Word Attributes

Team # 2309397 Page 2 of 11

# **Contents**

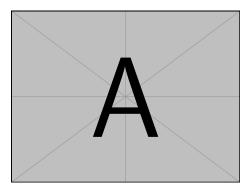
1	Intr	oduction	3
	1.1	Background	3
	1.2	Problem Restatement	3
	1.3		4
2	Assu	imptions and Notations	4
3	ARI	MA-BP Interval Prediction Model Based On Bootstrap	5
	3.1	Data Processing	5
	3.2		5
		3.2.1 Variation Explaining	5
			6
			7
4	Mod	lel 3	7
5	Test	the Model	7
	5.1	Sensitivity Analysis	7
	5.2		7
6	Con	clusion	8
	6.1	Summary of Results	8
	6.2		8
	6.3		8
M	lemor	andum	9
Re	eferen	ces 1	1

Team # 2309397 Page 3 of 11

## 1 Introduction

### 1.1 Background

Wordle is an online word puzzle game invented by Josh Wardle during the epidemicThe New York Times newspaper which is well known for the games it publishes hasbought Wordle on February 2021<sup>[1]</sup>. Wordle only allows one game to be played perday, and every player in the world plays to guess the same five-letter word in six triesor less each day Players can play in regular mode or hard mode. They can sharetheir scores via Twitter, thus attracting more people to play and share



Back in October 2021,less than 5.000 visits registered to its web page while by January 2022traffic had skyrocketed to over 45 million. Some ofus also love this game, Figure I shows one resulthat we got. The green tile indicates that the secretsolution word has the letter in the precise locationThe yellow tile implies that the answer has theletter but not at the correct location. Grey tilesindicate the letters are not contained in the solutionat all<sup>[3]</sup>.

Figure 1: Wordle Game

Now we have a file of daily results from January 7, 2022 to December 31, 2022. This file includes

twelve key variables that are crucial in our later research. In the file, the percentages of the number of people for the seven tries may not sum to 100% due to rounding

#### 1.2 Problem Restatement

By analyzing the above background, we summarize the tasks that need to be addressed as follows:

- Develop a model to explain the daily variation in the total number of peoplereporting scores on Twitter and use it to give a prediction interval for the totanumber of people on March 1.2023.
- Determine whether the attributes of words affect the percentage of players who choose the hard mode and explain the obtained results accordingly.
- If given a future date and the specific word, build a model to predict the percentage of 1-X tries in this day. After that, the word EERIE on March 1. 2023 should be used as a specific example of the model prediction, while analyzing theuncertainties of the model and the accuracy of the prediction.
- Develop a model to classify the diffculty of the words and identify theattributes of them under each category. Use this model to determine how difficult word EERIE is? Finally, discuss the accuracy of the classification model
- List and describe some other interesting features of the dataset

Team # 2309397 Page 4 of 11

#### 1.3 Our work

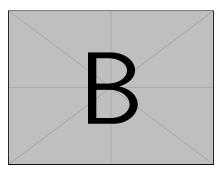


Figure 2: Our Work

# 2 Assumptions and Notations

**Assumption 1:** The number of daily online gamers in the data set is a time-dependentset of series and is independent of seasonal changes.

**Explanation:** Therefore, we can use ARIMA as well as LSTM models to predict a series ofdata for March 1.2023.

**Assumption 2:** The scores reported by players on Twitter on a daily basis are normaland reliable.

**Explanation:**To ensure that the model we build based on the dataset can reliably predict arange of data for March 1.2023.

**Assumption 3:** Assuming that Wordle's development process is consistent with thegame's life cycle theory.

**Explanation:** This assumption reduces the impact of external uncertainties on Wordle gamepredictions, thus making the whole process of prediction and analysis more efficient.

In this work, we use the symbols in Table 1 in the model construction. Other none-frequent-used symbols will be introduced once they are used.

- C 1 1	D. C
Symbol	Defination
$w_A$	Longitude within the i-th Wildfire Grid
$w_B$	Latitude within the i-th Wildfire Grid
$P_{Ai}$	The area of the i-th grid
$P_{Ai}$	the distance $d_{ki}$
$L_i$	Score for evaluating the k-th wildfire grid
NF	Noise factor
$C_i$	the ratio of players at different guess counts

Table 1: Notations

Team # 2309397 Page 5 of 11

# 3 ARIMA-BP Interval Prediction Model Based On Bootstrap

### 3.1 Data Processing

By reviewing the given data, we found that there are no missing values, but there areive outliers, One of them does not exist and two of them has less than 5 letters. wedeleted these three rows of data due to the difficulty ofobtaining the true values ofthesewords. The two remaining outliers are caused by data entry errors. In order to avoid anexcessive reduction in the amount of data, we changed them by combining the previousand later data as well as the semantics ofthem, In addition, the sum of the percentagesin the original data are all in [98%, 102%], which is not much different from 100%, sothey are reasonable and do not need to be processed. In summary, the preprocessing ofthe raw data is summarized in Table 2.

Database NamesDatabase WebsitesGoogle Scholarhttps://scholar.google.comWikipediahttps://www.wikipedia.orgwolframalphahttps://www.wolframalpha.com

Table 2: Data Processing

#### 3.2 Point Prediction Based on a Combined ARIMA-BP Model

#### 3.2.1 Variation Explaining

In the data preprocessing, although there are word entry errors, their corresponding numbers of reported results are not affected, so the complete reported data are analyzed in this problem. By observing and analyzing the characteristics of the changes in the number of reported results from Jan.7, 2022 to Dec 31, 2022, we found that they can be divided into 3 phases, as shown in Figure 3.

Based on the game lifecycle and player lifecycle theory[4], and combined with Wordle'sgame features, we explain the reasons for the changes as follows.

#### • Phase 1: Rapid Growth Period (January 7 2022-February 2 2022)

Since the launch of Wordle's web version, it has been updated with only one puzzle aday. This artificial scarcity enhances players' desire for challenge and anticipation. Thesharing function of Wordle uses emoji blocks to refer to the results of the game, whichis very recognizable and easy to spread, while also avoiding spoilers, thus attractingmore new players to play Wordle. In addition, the traditional popularity of crosswordpuzzles and the easy-to-play game mechanics have contributed to Wordle's furtheipopularity. During this period, Wordle's overall growth in the number of results reported is 348.85%, with the number peaking at 361.908 on February 2.

Team # 2309397 Page 6 of 11

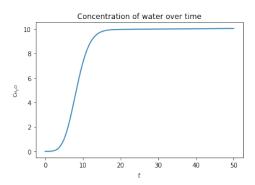


Figure 3: Changes in the number of reported results

- Parse 2: Rapid Decline Period (February 3, 2022 May 29,2022) During this period, Wordle experiences an overall 84.29% decline in the number of reported results, which is the general nature of internet fads. When the popularity of the game reaches its peak, it faces a significant loss of players due to gamers' declining interest and boredom with the game. The singularity of Wordle's game mechanic cancontribute to this. Furthermore, the emergence of competing games in the market suchas "Words with Friends' and Wordle's pirated games makes Wordle lose players further
- Parse 3: Stable Reduction Period (May 30,2022-December 31,2022) The overall decrease in the number of results reported for Wordle in this phase is 64.14%. Rapidly losing players in Phase 2 go to play Wordle generally because of itspopularity among the general public and they will quickly leave when Wordle is nolonger hot or the next trend emerges. The remaining players are often loyal Wordleplayers or players who have developed user stickiness due to the game's social circleAt this point, the number of reports will still drop but not by much.

#### 3.2.2 Reasons for Model Selection

Predicting the number of reported results is a problem in the field oftime series analysisThe traditional ARIMA model is capable of extracting deterministic information fromhistorical data in order to predict the trend of variables over time, however, this moderequires extensive testing before application, and the process of determining the orderofthe model is subjective to some extent [5]. In recent years, with the development ofartificial intelligence, machine learning algorithms have been widely applied to the fieldof data classification and prediction. Among which, BP neural network model which iseficient and convenient can effectively break the limits of traditional time seriesprediction models. Considering these factors, we decide to combine the advantages oftraditional time series models and machine learning models. It means that first of all.we use ARIMA model and BP neural network model to predict the number of reportedoutcomes separately, and then we use weighted average to combine the two models toobtain a more reliable ARIMA-BP combined prediction model.

Team # 2309397 Page 7 of 11

#### 3.2.3 ARIMA Forecasting

We first used the traditional ARIMA model for point prediction of the number of reported results, and this process is shown in 4:

First, we performed ADF smoothness test on the variable "Number ofreported resultsin the preprocessed data set and found that it wasn't smooth. Therefore, we differenced the series to the first order and tested the differenced series again. After that, we observed the ACF and PACF plot of the differenced series, and found that the ACF plotshowed the characteristics of trailing tails and the PACF plot showed the characteristics of lst order truncated tails. Therefore, the following ARIMA(1.1.0) model can be developed for the original series:

$$\begin{cases} (1 - \phi_1 B)(1 - B)x_t = \varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_{\varepsilon}^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t, \\ E(x_s \varepsilon_t) = \mathbf{0}, \forall s < t \end{cases}$$
 (1)

## 4 Model 3

## 5 Test the Model

# 5.1 Sensitivity Analysis

# 5.2 Robustness Analysis

这部分很重要,不能缺!



Figure 4: ARIMA model building process

Team # 2309397 Page 8 of 11

# 6 Conclusion

## **6.1** Summary of Results

## 6.2 Strengths

• The sensitivity analysis of the model demonstrates the effectiveness of the model under different parameter combinations and prove the robustness of the mod

• Second one ...

## **6.3** Weaknesses and Improvements

- The analysis of fish migration can be more accurate if we have more complete data;
- Some approximate analysis methods are applied to model the management of fishing companies, which may lead to a situation contrary to the actual one in extreme cases.



### Memorandum

From: Team #1887415157 of 2023 MCM

Date: February 31, 2023

Subject:

A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning.

A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning.

# Note

- 1. Text Text Text Text

- 4. Text Text Text Text

that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning. A randomly generated piece of English that has no real meaning.

This part must be a bit of an unordered list or an ordered list or something like that, otherwise it's really all text and will look super ugly.

- Solution 1. **Build more shopping centers**. Explain solution1 explain solution1 explain solution1 explain solution1 explain solution1.
- Solution 2. **Build more shopping centers**. Explain solution2 explain solution2 explain solution2.
- Solution 3. **Build more shopping centers**. Explain solution3 explain solution3 explain solution1 explain solution1 explain solution1.
- Solution 4. **Build more shopping centers**. Explain solution4 explain solution4 explain solution4 explain solution4 explain solution4 explain solution4.

Team # 2309397 Page 10 of 11

Conclusion. This part of writing conclusive things, one or two sentences to summarize it, do not write too much.

Team # 2309397 Page 11 of 11

# References

[1] Hao H, Wang Y, Xia Y, Zhao J, Shen F. Temporal convolutional attention-based network for sequence modeling. arXiv preprint arXiv:2002.12530. 2020Feb 28.

- [2] Harvey AC. 1990 Forecasting, structural time series models and the kalman filter. Cambridge, UK: Cambridge University Press.
- [3] Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (NIPS) (eds F Pereira, CJC Burges, L Bottou, KQ Weinberger), pp. 1097–1105.