Analyzing difficulty of Wordle using aggregate Twitter responses to determine public reaction*

My subtitle if needed

Ivan Li

28 April 2022

Abstract

Using the Twitter API, data was gathered recording the public scores of Wordle players over one month. The frequency of each daily score was obtained and analyzed based on level of success, geographical location, and linguistic properties of the word. It was found that the linguistic properties of words such as commonality have impacts on player success. These results may have significance in literary and linguistic analysis.

1 Introduction

(put citations later) Wordle is an online word-guessing game bearing resemblance to historically popular word puzzles such as the New York Times crossword puzzle and the board game Mastermind. Players are given a five-letter word each day and are asked to guess the word within six tries. If the word is not guessed in earlier attempts, the letters for your word will become either grey, yellow, or green: grey indicating that the letter was not found anywhere in the word, yellow indicating the presence of the letter but in the wrong position, and green indicating both letter and position are correct. The goal is to guess five green letters and therefore the correct word. The game was created by Josh Wardle, a software engineer who created the game for himself and his partner to pass the time during the 2021 COVID-19 pandemic. The meteoric rise in popularity of his creation has been attributed to a feature where players can easily copy and paste their results to share with their friends and family. Wardle's game found international spread in late December of 2021 thanks to multiple factors, including the ease in sharing results and the daily nature of the game, as well as its simplicity. Wordle's popularity can be seen to its fullest extent on Twitter, where millions of users share their results every day by pasting their scores via the website's built-in feature.

This paper aims to analyze several metrics of Wordle's popularity on Twitter. As a result of millions of users sharing their scores each day, analysis can and has been done on variables including average score- a recent study found that Sweden was the country with the lowest average guesses with 3.72 average guesses out of a maximum of 6. Here, some key findings we wish to seek include attempting to model whether certain characteristics of a word correlate with its average difficulty, and the average score of Twitter users in relation to their geographical location, such as their country of residence. These metrics were not discussed deeply in prior papers and may be of interest to future puzzle makers, linguists, and game theorists.

Using the Twitter API, tweets pertaining to Wordle were gathered and aggregated into multiple datasets used for analysis. These tweets were gathered during intervals from April 4-10 2022, and April 17-23 2022). User scores were collected by searching for common templates that are used by the website domain to share scores in their tweets. Along with the score the used reached, the time of posting, geographical location, and other data provided by Twitter were also obtained. The results were turned into graphical data and analyzed according to our key missions.

[*] Code and data are available at: https://github.com/Ivannoar/Twilight.

In the Data section, we go over the data gathered from the Twitter API and how it was cleaned and arranged for proper statistical analysis. We discuss variables used, methodology, and graphs showing distributions of data. In the Model section, we discuss our model and its implications for how we interpret our results going forward. In the Results section, graphical data is shown and used to present our story according to the key goals and findings we discover during our analysis. Lastly, we discuss what has been done, its significance, and weaknesses of the paper in our Discussion section.

2 Data

2.1 Dataset

To accomplish the goals set out in this paper, the data used consisted of a sample of tweets from the social media website Twitter. Twitter hosts and stores an immense number of messages which can be accessed via the website's API; however, access to tweet data is restricted to the past 7 days during access unless given permission by the site. The raw data collected in this paper contains 504000 tweets and 90 variables regarding various properties of the tweet and its sender, such as the message, date and time of posting, details of the sender's account, popularity metrics, and direct links to the tweet online. Tweets were searched for and collected over two timeframes: tweets from April 4-10, and from April 17-23. This was done using the programming software R (R Core Team 2020) and the tweet collecting package rtweet (Kearney 2019). Complete details as to how the data was obtained and cleaned can be found in the datasheet in the appendix.

2.2 Variables

The dataset consisted of 90 variables which needed to be trimmed down to a manageable list of relevant data. These variables included unique tweet identifiers and the exact text included, details regarding whether the tweet was original or sent as a reply or response to another, the number of likes and retweets on the initial or response posts, the device used for sending posts, and links to the posts and accounts of each observation. We seeked to only utilize the score the user had achieved in the game, their country/location, the date of the message, and the presence of 'hard mode' while playing, which adds the condition that you must use hints you are given from prior guesses. To accomplish this, variables needed to be changed and manipulated. From the raw data, which used Unix time to store chronological data, the tidyverse (Wickham et al. 2019) and lubridate (**citelubridate?**) packages were used to create new variables. From the Unix time, the date was extracted from the raw data and stored in a YYYY-MM-DD format. An indicator variable corresponding to whether the user used hard mode was created and each observation was given a classification based on the contents of their tweet. The final dataframe contained 4 variables, those variables being the date, country, score, and gamemode for each tweet instance provided.

2.3 Missing Data

There are limitations to the data gathered in this paper which will have impacts on the conclusions we make going forward. Due to regulations of Twitter and limitations of the API provided, historical data was not available for analysis, since the website only allows for mass collection of tweets from a week in the past. The paper regrets that long-term historical analysis was not possible to conduct. As well, chronological data was unavailable to be processed since the rtweet package can only collect recent tweets in bulk. The result of this was that data was unable to be gathered consistently over a long time interval due to hardware limitations and feasibility. While the data should not be negatively impacted by this, it becomes impossible to properly conduct chronological analysis, and there remains the possibility of time interval bias from a majority of the observations being taken at one timeframe.

There is also inherent bias included in the dataset as a result of the nature of tweets. Not all players of Wordle report their results online, and a smaller subset on Twitter, leaving the possibility that the conclusions drawn do not apply to the general population of all Wordle players. At best, we can confirm our conclusions from a sample of Twitter users and attempt to extrapolate to a far larger population. As Wordle is a word based off the English language, it is likely that the geographical data will also be biased in favor of English-speaking

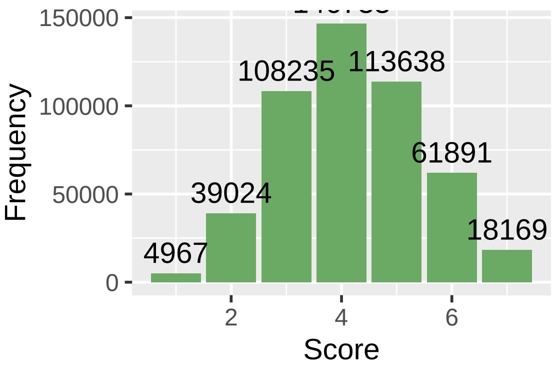
Table 1: First ten rows of a dataset showing Wordle-related tweets

Date	Country	Wordle Score	Hardmode
2022-04-04	NA	2	Off
2022-04-04	NA	6	Off
2022-04-04	NA	5	Off
2022-04-04	NA	6	Off
2022-04-04	NA	6	Off
2022-04-04	NA	4	Off
2022-04-04	NA	4	Off
2022-04-04	NA	6	Off
2022-04-04	NA	4	Off
2022-04-04	NA	5	Off

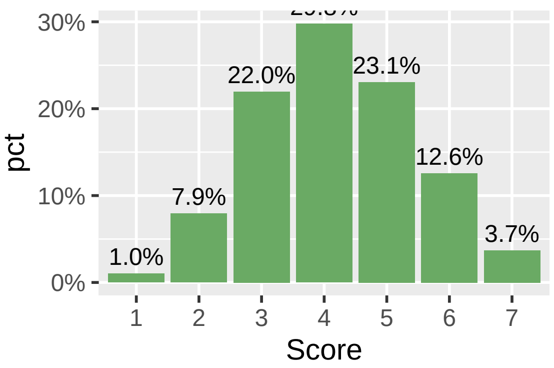
countries, as countries that are not primarily English-speaking likely have far less Wordle players, which has a magnified effect considering the nature of the dataset. Further bias exists where users who experience lower scores (higher average guesses to guess the word) or fail to complete the game may not post about these results on Twitter. All data collected is by necessity self-reported and lends itself to nonresponse bias. Users may also lie or provide false data by changing their reported score or by parodying the original Wordle message, which may affect the shape of our distribution.

2.4 Plots

Plots were created corresponding to the chosen variables in our cleaned dataset in order to familiarize with the obtained data and provide motivation for further analysis. To begin with, we wish to plot the frequency of each result/score across our entire dataset. We compare the frequency of each result and plot it as both the raw frequency and the percentage that each score appears in our dataset. Note that failures, where the user was unable to guess the word in the maximum 6 tries, is plotted as a score of 7. We observe that a majority of Twitter users guess complete the game within 4 tries, and that the distribution of guesses is sensibly unimodal. The distribution is slightly right-skewed, and the dataset shows that the number of failures is over three times as much as the number of people who guess the word on their first try, which is expected due to the low probability of guessing the exact word with no prior information compared to the chance of failure. This is



also reflected in the proportions between scores of 2 and 6.



Next, we look to compare the distributions between users who use hard mode and users who do not. It must be noted that the amount of people who do not enable hard mode dwarf the number of people who do. However, it is possible that a non-insignificant majority of people who do not report using hard mode are self-imposing hard mode conditions onto themselves, whether intentionally or unintentionally. Regardless, we plot the proportion of hard mode users to normal users and see that 94% of users in our sample do not use hardmode, which can be explained by the fact that hard mode is opt-in and users may not wish to bother with enabling it for multiple reasons. We then repeat the distribution of scores in the previous set of graphs but separate the results by the mode that users play on. From our graphs, we see that our sample of players on hard mode more commonly reach scores of 3 and 4, but also have higher failure rate than normal players. A strategy used by normal players is to disregard hints in the first 2 or 3 words by trying to 'use' as many common letters as possible to reveal as many letters in the word as possible before making a realistic guess, while hard mode players must use the hints they are given which reduces the efficiency of the strategy but means that each word they use has a relatively higher chance of being correct. However, this runs into problems when it is difficult to use the hints provided or the word shares a structure with other common words (for example, Wordle #284 was 'stove': a player could easily waste guesses on 'stone', 'store', 'stoke', etc), which results in a higher failure rate.

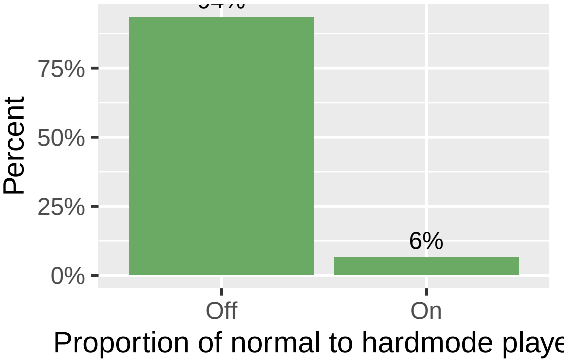
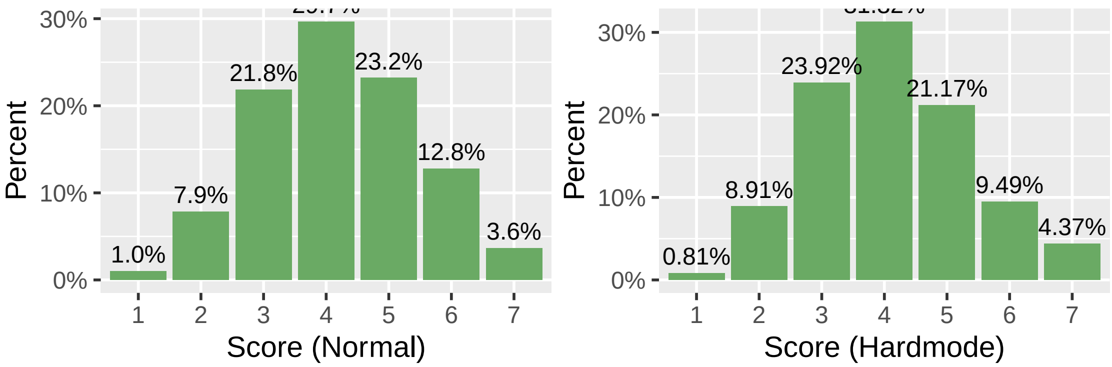


Table 2: Top 10 most common countries found in dataset

Country	Frequency
United States	4020
United Kingdom	679
Canada	589
Ireland	150
India	83
Republic of the Philippines	60
South Africa	50
Trinidad and Tobago	49
Jamaica	47
Brazil	43



We also show the frequency that countries occur in our dataset, using the geographical data provided in the tweets. We display the top 10 countries found in our dataset and observe that the Americas and the UK make up a majority of the tweets found in the data.

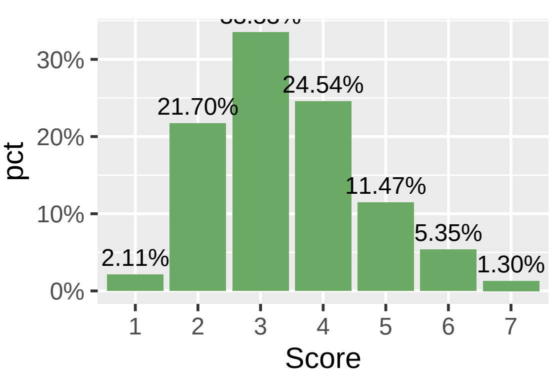
To conduct further analysis, we seek to plot data based on the average score of users on different dates and different countries. Shown is a table displaying the date of the daily word being given, the average score of Twitter users across our dataset, and the word of the day. Also plotted is the distribution of the ‘easiest’ and ‘hardest’ words determined by average score, displayed in percentages and grouped by the mode used.

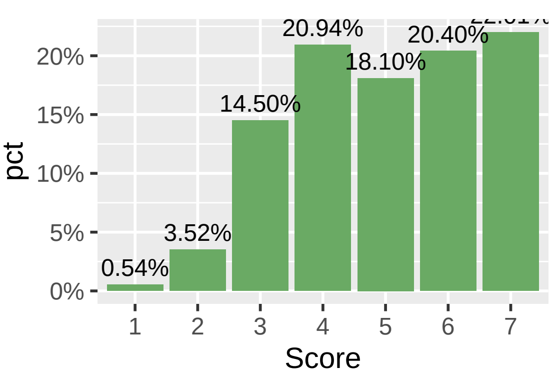
Table 3: Average score of Twitter Wordle players

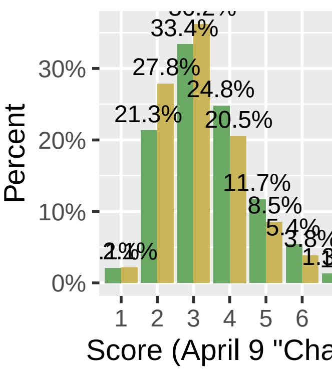
Date	Average Score	Daily Word
2022-04-04	4.4	shawl
2022-04-05	4.5	natal
2022-04-06	4.6	comma
2022-04-07	4.6	foray
2022-04-08	4.1	scare
2022-04-09	3.4	stair
2022-04-10	3.7	black
2022-04-17	4.2	ample
2022-04-18	3.9	flair
2022-04-19	5.0	foyer
2022-04-20	4.4	cargo
2022-04-21	4.6	oxide
2022-04-22	3.5	plant
2022-04-23	3.9	olive

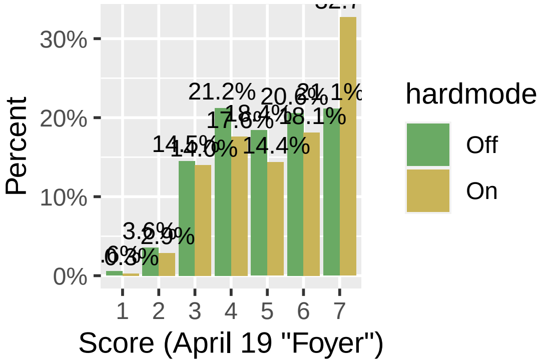
Table 4: Linguistic Properties of Daily Wordle Words

Date	Average Score	Daily Word	Word Frequency Rank	Orthographic Neighbours
2022-04-04	4.4	shawl	11938	2
2022-04-05	4.5	natal	26852	3
2022-04-06	4.6	comma	11931	1
2022-04-07	4.6	foray	10467	1
2022-04-08	4.1	scare	3837	10
2022-04-09	3.4	stair	2880	2
2022-04-10	3.7	black	253	5
2022-04-17	4.2	ample	6374	3
2022-04-18	3.9	flair	11502	1
2022-04-19	5.0	foyer	9811	2
2022-04-20	4.4	cargo	4953	1
2022-04-21	4.6	oxide	10008	0
2022-04-22	3.5	plant	623	5
2022-04-23	3.9	olive	6242	1









Finally, we conduct analysis on the words used in the game themselves. The following table shows the date the word was used, the average score users scored on it, the word of the day, the frequency of the word in texts according to the Corpus of Contemporary American English (COCA) (**citecoca?**), and the number of orthographic neighbors to the word according to MCWord (**citemcword?**). Orthographic neighbors are defined as words which are the same length but differ by one letter (for example, scare and stare).

3 Model

In this paper, we seek to use regression models to determine if there were relationships between the average score that Twitter users scored daily and properties of the given word for the day. We begin by constructing directed acyclic graphs to visualize the variables we wish to discuss and model, and clearly show the relationships we believe exist between them. Using the DAG as a visual, the paper aims to show that the frequency of the word in literature, the number of orthographic neighbours, and the game mode of the player all have significance in predicting the average score of a player.

```
## QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-r443852'
## TypeError: Attempting to change the setter of an unconfigurable property.
## TypeError: Attempting to change the setter of an unconfigurable property.
```

```
## QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-r443852'
## TypeError: Attempting to change the setter of an unconfigurable property.
## TypeError: Attempting to change the setter of an unconfigurable property.
```

The paper predicts that the linguistic properties of words such as the number of orthographic neighbors will have a different effect on the average score of a player based on their game mode; players on hard mode are hypothesized to have higher (worse) scores when the daily word shares similar structures to other possible guesses. In order to attempt to model this effect, we create multiple regression models with 2 predictors each as follows:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

In our first multiple regression model, Y_1 represents the average number of guesses it takes a normal player to complete a given Wordle puzzle. X_1 represents the frequency rank of a given word used in the game, X_2 represents the number of orthographic neighbors that the word has, β_0 represents the intercept or average score when the frequency rank and number of neighbors is 0, and β_1 and β_2 represent the regression coefficients. For every one increase in frequency rank and every one increased neighbour, Y_1 increases by β_2 and β_2 respectively.

$$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

The second multiple regression model is similar. Y_2 represents the average number of guesses it takes a player on hard mode to complete a given puzzle and the other predictors and coefficients share the same meaning. We then wish to model the probability of failure to complete the game based on the same groupings of game mode and linguistic properties. The models are as follows:

$$Y_3 = \Pr(y_i = 1) = \text{logit}^{-1}(\beta_{0H} + \beta_{1H} X_{1H} + \beta_{2H} X_{2H})$$

$$Y_4 = \Pr(y_i = 1) = \text{logit}^{-1}(\beta_{0H} + \beta_{1H} X_{1H} + \beta_{2H} X_{2H})$$

$$Y_3$$

and

$$Y_4$$

respectively represent the average probability that a normal Wordle player and a Wordle player on hard mode will fail in guessing the daily word within 6 guesses. The other predictors and coefficients share the same meaning as the prior models.

```
## Warning in `[.data.frame`(masterdata_clean_modelmain, !
## is.na(as.numeric(as.character(masterdata_clean_modelmain$text))), : NAs
## introduced by coercion
```

3.1 Features

We are interested in the linguistic properties of words and how they may translate to the difficulty of guessing it in the context of Wordle, so the variables we decide to include for our model include strictly predictors related to our goal. The frequency rank of a word directly correlates to its use in literature, speech and academia as reported by the COCA, and is used here to represent how common a word is and how likely one is to have encountered it in their lives. It is believed that more familiar words will be easier to recall and guess, and therefore have a lower score, while less common words will take longer to guess. We also look at the number of orthographical neighbors as a predictor: structurally similar words can serve to 'remind' players of a word's presence and lead them to the correct guess, but they can also confuse and discourage players who continually guess the wrong neighbor. The lack of neighbors can also assist the player, since a lack of structural similarity to other words is representative of 'uniqueness', and therefore cannot be confused for other words: for example, if one is asked to fill in '_nique', it becomes apparent that only 'unique' completes the word, and in general unique has 0 neighbors and is entirely distinct in structure. This may serve to avoid confusion in players and reduce the amount of guesses required to win the game.

3.2 Model Concerns

The most concerning element of modelling these relationships is the sample size of the amount of words in our database. Due to Twitter limitations and constraints, a very small number of words were collected which is not nearly enough to form an accurate model of what we wish to predict. The number of data points per day is suitable and reduces variance for each word, but the small number of words means that each day shows great significance in the data and does not lead to an accurate conclusion. The model would benefit from a dataset spanning a much longer timerange in order for more words to be analyzed and fit into the model.

4 Results

Two variations of two models were created and analyzed for a total of four completed models, and interpreted by analyzing the values of the predictor coefficients and their implications on the overall model and the value which they served to predict. We continue by sharing the results of each type of model and how their findings can be compared to each other.

4.1 Modelling Average Score

The first form of model was created to predict the average score of a Wordle player based on the frequency of the word and the number of orthographic neighbours the word had. Our results show that if a player is playing normally, i.e. not using hard mode, their average number of guesses will increase by 0.00004126 for each rank the word is from 1. This corresponds to the value of

$$Y_1$$

= 0.00004126 in our model equation. For example, a word around the ranking 10000 would increase the average number of guesses by 0.4126. Furthermore, the value of

$$Y_2$$

was determined to be -0.04117. This can be interpreted as: For every orthographical neighbour the word of the day has, the average number of guesses it takes a normal Wordle player will be reduced by 0.04117 guesses. Lastly, the intercept was found to be

$$\beta_0$$

= 3.954. The interpretation of this result is that a word at ranking '0' and with 0 neighbours will be guessed in an average of 3.954 tries; it represents the average number of guesses when the other two predictors are held constant.

```
##
## Call:
## lm(formula = score ~ rank + neighbournum, data = masterdata_clean_modelmain_norm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9382 -0.9382 -0.1170  0.8298  3.2996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.954e+00  4.192e-03  943.21   <2e-16 ***
## rank          4.126e-05  3.021e-07  136.56   <2e-16 ***
## neighbournum -4.117e-02  7.810e-04  -52.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.248 on 458509 degrees of freedom
## Multiple R-squared:  0.06022,    Adjusted R-squared:  0.06021
## F-statistic: 1.469e+04 on 2 and 458509 DF,  p-value: < 2.2e-16
```

The second model used the same predictor variables but used a dataset consisting of only players using Wordle's hard mode. This means that the interpretations of the variables remain the same, but they have the effect of changing the average number of guesses for a hard mode player and not a normal player. This has an effect on the values of the coefficients, which are reflected in the model. In the secondary model, the intercept has a value of 3.843, the ranking coefficient has a value of 0.00004232, and the neighbour coefficient has a value of -0.04185. In general, there is not a drastic change in the values of the coefficients between each model. The second model features a lower intercept, implying that hard mode players finish the game slightly faster, but their results are very slightly more impacted by the frequency of the word each day.

```
##
## Call:
## lm(formula = score ~ rank + neighbournum, data = masterdata_clean_modelmain_hard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8534 -0.8808 -0.0649  0.7563  3.4135
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.843e+00  1.602e-02  239.82   <2e-16 ***
## rank          4.232e-05  1.184e-06   35.74   <2e-16 ***
## neighbournum -4.185e-02  3.047e-03  -13.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.243 on 31637 degrees of freedom
## Multiple R-squared:  0.05952,    Adjusted R-squared:  0.05946
## F-statistic:  1001 on 2 and 31637 DF,  p-value: < 2.2e-16
```

4.2 Modelling failure rate

The second form of model was created to predict the failure rate of players across both game modes using the same metrics as the first model. The results of these models have similar structure due to the simplicity of the models and the small number of predictor variables, but the values and their interpretations are different and reveal separate results from the previous models. The third model, which predicts failure rate of a normal Wordle player, shows that the intercept has a value of

$$\beta_0$$

= 0.02647. Because this model is predicting a percentage, the interpretation of the intercept shows that if a word has 0 rank and 0 orthographical neighbours, a normal Wordle player will have a 2.647% chance of failure on average according to the model. The value of the rank coefficient is

$$Y_1$$

= 0.000001314. For every rank a word has, the percentage chance that a normal will fail to guess the word within 6 tries increases by 0.0001314. If a word had a rank of 10000, then the chance that a player would fail increases by 1.314 from the intercept value. The neighbour coefficient is recorded as being

$$Y_2$$

= -0.0004534, and shows that for every neighbour a word has, the chances of failure drop by 0.04534%.

```
##
## Call:
## lm(formula = failed ~ rank + neighbournum, data = masterdata_clean_modelmain_norm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06040 -0.04113 -0.03422 -0.02935  0.97547
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.647e-02  6.275e-04  42.177  < 2e-16 ***
## rank          1.314e-06  4.523e-08  29.058  < 2e-16 ***
## neighbournum -4.534e-04  1.169e-04  -3.878 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1868 on 458509 degrees of freedom
## Multiple R-squared:  0.00228,    Adjusted R-squared:  0.002276
## F-statistic:   524 on 2 and 458509 DF,  p-value: < 2.2e-16
```

The final model which predicts the failure rate of hard mode players shows significantly different results than the previous model predicting failure rates of normal players. We observe an intercept of 0.0301, a rank coefficient of 0.000001598, and a neighbour coefficient of -0.00134. A table shows the two comparisons side by side along with the difference by value and percentage.

```
##
## Call:
## lm(formula = failed ~ rank + neighbournum, data = masterdata_clean_modelmain_hard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07252 -0.04982 -0.04185 -0.03535  0.97225
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.301e-02  2.624e-03  12.580   <2e-16 ***
## rank          1.598e-06  1.939e-07   8.241   <2e-16 ***
## neighbournum -1.134e-03  4.990e-04  -2.272   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2037 on 31637 degrees of freedom
## Multiple R-squared:  0.003012,   Adjusted R-squared:  0.002949
## F-statistic: 47.79 on 2 and 31637 DF,  p-value: < 2.2e-16
```

5 Discussion

5.1 Data and Model Findings

In the time of writing of this paper, numerous reports and articles have been written about various aspects of Wordle, thanks to its meteoric rise in popularity since the beginning of 2022 as a result of the COVID-19 pandemic. This paper attempts to go a step further into analysis and cover the linguistic aspects of the popular word game and if any significant conclusions can be drawn from the results of the author's findings. The models constructed in this paper lead to both conclusions that are both reflected in the data and which are surprising at first glance but can be understood. It was the paper's hypothesis that the number of orthographic neighbours a word has would negatively impact the performance of the player; the failure rate and the number of guesses taken would both increase the similar a word was to others. This was concluded from the author's own experiences and the thought that words sharing high similarity would lead to many incorrect guesses used on the neighbours. However, the four models constructed show that while the less common a word is, the harder it is to guess, as predicted, the number of neighbours has a negative correlation with the average number of guesses as well as the failure rate of both types of players. Some explanations could include that having more neighbours means that the word's structure is easier to recall and thus provides the player with an early lead if guessed, or that the common structure of the word is correlated with words that players like to begin with as their first or second words, and therefore are easier to obtain hints on at the beginning of the game.

The data had shown that on average, players on hard mode have lower average guesses than normal players but exhibit a higher failure rate overall. This is reflected in the paper's models- while the model of the hard mode dataset showed a lower intercept value in the first type of model, there was a significant difference in the failure rates of the two types of players as reflected in the intercepts of the third and fourth model. It was expected that players on hard mode would have lower guesses on average, for multiple reasons. One reason hypothesized in the paper was that always utilizing hints given by the game results in continual progress and eventual deduction of the word. Another reason is that players on hard mode may take the game more seriously or competitively, and therefore may have lower averages on virtue of using techniques or strategies that other players are indifferent to.

5.2 Weaknesses and next steps

Weaknesses of the paper largely pertain to the dataset used and the significance and validation of the model used to predict gameplay. The dataset used in this paper was gathered from the author and is largely incomplete in observations. While Wordle has been released and its rise in popularity present for months as of writing, the dataset only covers two weeks and there is a gap in observations in between them. This results in an extremely small sample size which is not suitable for modelling and analysis of sample statistics. Prior articles referenced in the paper had analyzed Wordle games over several weeks or months, which would be a more appropriate timeframe in order to gather enough data to conduct a proper analysis on average gameplay statistics. While the dataset is lacking in sample size regarding days, the volume of tweets for each day recorded is sufficient to gauge public opinion on the recorded days accurately such that statistics such as geographical and gamemode proportions should hold. However, more data is required on both the amount of words and scores in order to make sure that the results in this paper hold.

The models created in the paper also struggle to hold as a result of the dataset and the lack of predictor variables. It is likely that the model does not hold when compared to the full population of Wordle games and words, as we have only taken a small sample of words which is prone to large amounts of variance. The models have not been tested for validation or regression coefficients which may result in an unfounded and inaccurate model. Future work may involve refinement of the models and addition of more predictor variables in order to accurately model the linguistic relationships between the chosen words and player performance.

It is the author's hope that the results from this paper serve as a start and as inspiration for continual research into word-based games and puzzle games in general. As a viral game, Wordle is a popular topic of discussion internationally and online, and already many spin-offs of the game have seen their own share of success. Future game designers and puzzle game makers may seek to use the results and impacts of Wordle to tailor their own works to provide a fun challenge for players, and research should be continued to determine how one can find a balance between difficulty and frustration as players of Wordle frequently experience. Researchers of linguistics may also find interest in Wordle as a method of gathering public knowledge and opinion of the English language. As the language continually evolves and words take on new meanings or become obsolete and forgotten, Wordle may be useful in teaching and reminding the general population of literacy much like the crossword puzzles of the past. Popular trends can not only find success among the general population and provide a source of bonding and discussion, but can also be a source of research in order to learn from these trends and utilize their implications for further purposes.

Appendix

.1 Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to conduct analysis on Twitter trends regarding the popular online game Wordle. Twitter is a convenient website to gather large amounts of player statistics as users regularily and daily share their scores in an easy to analyze format.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by the author of this paper to serve the paper's purposes.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - No monetary cost was required in the creation of the dataset; it was created free of cost using the rtweet (Kearney 2019) package and the programming software R (R Core Team 2020).
4. *Any other comments?*
   - Due to data gathering errors, there is a gap in the temporal data of the dataset. A large majority of the tweets are gathered at a general timepoint correlating to when the tweets were obtained. As such, the tweets are not evenly distributed timewise.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - Each instance of an observation represents a message known as a 'tweet'. This message is akin to a text message which may contain emoticons, pictures, and GIFs attached with the message. Tweets can also be sent as replies to other tweets, or as a quote tweet which is standalone but is connected to an initial message.
2. *How many instances are there in total (of each type, if appropriate)?*
   - In total, there are 508000 instances and 508000 gathered tweets.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset does not contain all possible instances. Billions of messages are sent and processed by Twitter daily and it is impossible to process and filter such a colossal stream of data to obtain only relevant data. The dataset is a sample of tweets relevant to Wordle and the larger set would consist of all relevant tweets made each day for the period of data gathering. The sample is not chronologically representative of the entire set because of limitations with data gathering. rtweet is limited in that it can only gather 'recent' tweets related to when the request is sent. Due to this, evenly distributed chronological coverage was not able to be gathered.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of a message sent in text and unicode characters, which may have a picture or GIF attached. There are also properties of the tweet and sender account such as chronological data in the dataset.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Each tweet has a status id which is shown in the raw data. This status id is unique and directly corresponds to the pertaining message in the Twitter databse.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally*

removed information, but might include, for example, redacted text.

- There is no information missing that was blocked by Twitter or our data collecting programs. Information such as the location of the sender of the tweet or the account information may be intentionally withheld due to privacy concerns of the user.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Relationships between invididual instances are made explicit. Part of the raw data contains information about if the instance was a reply or 'quote tweet' in response to an initial tweet. These initial tweets may or may not be in the dataset, and this can be verified by using the status id of the initial message and checking if it is in the dataset.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - A datasplit of 80/20 for the training and testing datasets will be used to create and test models for the data. These splits will consist of randomly sampled partitions of the data and will be done so to mitigate bias from the models created.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - Some sources of noise may include tweets which matched our criteria for relevance and were obtained but do not provide any usable information. There may also be 'joke' results which parody the data we wish to obtain but is itself not accurate information.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset relies on external resources, as we are looking at tweets. There are no guarantees that they will exist in the future as Twitter and its users have the right to delete and hide tweets at any moment. However, archival services exist which would allow for the messages to be preserved. As well, deleted tweets are stored in Twitter's databases, although unavailible to the public.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - The dataset may contain confidential data if the user explicitly and voluntarily mentioned this in their tweet.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - The dataset may contain offensive, insulting, and malicious text if the user chose to include such text in their original tweet.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - The dataset does not identify any sub-populations as this data is not obtainible from tweets.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It would be possible to identify individuals from the dataset if there were enough information in their account details and their message, as this data is included and gathered alongside the tweet instance.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - The data contains the geographical location of where the tweet was sent and the geographical

location of the user. There is no other sensitive information included.

16. *Any other comments?*
    - None

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data was directly observable from each tweet and user, as tweets are sent as raw text and other details are directly obtainible from the timestamp and self-provided user information.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - The data was collected using the rtweet package found in R. These procedures were validated by manually checking that the program was gathering valid tweets using the status id and referencing of tweets and accounts.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - The sampling strategy was to take recent tweets matching the search criteria as of the time of request. Recent tweets are shown in order of time posted.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - Only the author of the paper was involved in data collection.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data was collected over two timeframes- on April 10 and on April 23 2022. The timeframe roughly matches the creation timeframe of the data: The creation timeframes are roughly from April 4-10 and April 17-23 2022. These times are when daily tweets are collected.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - None
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - The data was collected from other sources, using the rtweet package.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - The individuals in question were not notified about the data collection.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - The individuals in question consented to the collection and use of their data in the Twitter Terms of Service which all users must agree to for site access. The link can be found here: https://twitter.com/en/tos
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Consenting individuals may choose to deactivate their account which would delete all messages and tweets they had sent from the platform.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a*

data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
  - None

12. *Any other comments?*
    - None

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - Cleaning of the data was done to obtain more relevance in our dataset. After the raw data was collected, observations were filtered by specific message criteria. The relevant data (Wordle scores) are contained in messages through specific phrases which are copied and pasted by users from the game's website. Scores take the form of the phrase "Wordle _____ *" /6" where the number of the daily word and the number of guesses taken by the player are displayed. Tweets were filtered according to whether this phrase was included in their tweet. Once this was done, the numerical data of the number of guesses the user took was extracted from the message text and stored as a new variable. Games in which the user did not successfully guess the word in 6 tries, and therefore failed, were counted as taking 7 guesses to complete. After this was done, the date and time that the tweets were sent was extracted from the Unix time provided in the raw data using the lubridate (**citelubridate?**) package. A binary variable was also created indicating if the user had 'hard mode' enabled in their game, which adds additional conditions to the game making it more challenging. This indication was availible in the raw text; messages from hard mode players would contain a "*/6*" in their tweet, the star indicating the use of hard mode. Once this was done, the data was aggregated into a single dataframe. Partitions were also made according to the creation timeframe of the tweets for potential future use.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - The "raw" data is contained in the repository that this paper is contained in.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - The software used to clean the data consists of R packages, which have been cited and credited.
4. *Any other comments?*
   - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - None
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - Yes. Code and data are available at: https://github.com/Ivannoar/Twilight
3. *What (other) tasks could the dataset be used for?*
   - Other tasks might consist of more analysis pertaining to Wordle, as well as using the dataset to draw conclusions based on Twitter in general.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - Tweet geographical and chronological data, as well as the status and description of the accounts used could be used to unfairly treat the individuals involved. A dataset consumer might avoid making these conclusions and taking the data at more of face value.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- The dataset should not be used for any malicious or illegal activity.

6. *Any other comments?*
    - None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The dataset will not be explicitly distributed to third parties, but will be made availible publically online.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be distributed on GitHub via this paper's repository.
3. *When will the dataset be distributed?*
    - The dataset will be distributed on April 27, 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - None
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - None
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - None
7. *Any other comments?*
    - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be hosted by the author of the paper.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The owner of the dataset can be reached via the GitHub account the repository is hosted on.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - None
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - The dataset will not be updated.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - There are no limits on the retention of the data associated with the instances.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - Older versions of the dataset may be hosted by the author of the paper on their local files. The obsolescence will be communicated from the repository the paper is hosted on.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- Others can contribute to the dataset via GitHub's built-in collaboration features, which will be verified and finalized by the author.

8. *Any other comments?*
   - None

A Additional details

References

Kearney, Michael W. 2019. "Rtweet: Collecting and Analyzing Twitter Data." *Journal of Open Source Software* 4 (42): 1829. https://doi.org/10.21105/joss.01829.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.