A preliminary demographically informed analysis of Toontown Rewritten chat message data

E. Ciereszynski

2023-07-04

Introduction

Toontown Rewritten, released in 2013 as a community reincarnation of Disney's Toontown Online, is a MMORPG in which the player creates a character called a Toon and completes in-game tasks in order to advance through various game areas. Toons are highly customizable and the player is able to select from many options for species, colours, clothing items, and weapons, which are referred to in-game, as well as from this point forward in the present document, as Gags. It is also possible to communicate through text chat, which relies on a set of whitelisted words, numerals, and symbols. Demographic analyses of Toon characteristics have been carried out previously, and natural language processing, referred to from this point forward as NLP, analyses of chat data have also been performed. The document at hand is a preliminary marriage of these two spheres; a linguistic analysis of a message corpus which takes demographic factors into account. Previous demographic work on Toontown Rewritten has been very informative while previous NLP work has been much less informative. This current work was thus motivated by a belief that combining the two areas and allowing them to inform each other would generate meaningful and novel observations. To carry out this analysis, a corpus which consisted of 4000 text chat messages and various pieces of demographic information related to the speaker of each message was assembled by hand between August 2022 and January 2023. The data was subsequently preprocessed, cleaned, explored, and analyzed with R and Python. Unfortunately, analysis yielded few interesting or notable results. The lack of conclusive results in this study demonstrates firmly that more investigation, the possibility of which will be created through the collection of much larger amounts of data, is necessary.

Background

Previous work

I have done previous Toontown Rewritten-related data analysis and NLP work but until this present analysis these two domains had remained separate in my research. In January 2022, I published my paper *An Exploration of Toontown Rewritten Demographics* which reported on the demographic characteristics of a population sample of 3000 Toons and contained extensive correlation and statistical analyses related to this information. The most notable finding of this project was the existence of co-occurring bundles of demographic features and a certain subpopulation of Toons who appear to be perpetuating various countercultural trends. In April 2022, I published a corpus analysis of 4000 chat messages I had collected in-game. The analysis reported on exclusively NLP-related metrics, namely sentiment, subjectivity, and message length, as well as the most frequently occurring tokens in the corpus. In May of the same year, I published another brief NLP-focused analysis carried out on that preexisting corpus wherein I performed latent Dirichlet allocation, which represents documents as groups of topics which output words at varying probabilities, and non-negative matrix factorization, which performs dimensionality reduction and clustering. Neither of these analyses yielded meaningful or particularly coherent results, which I tentatively hypothesize may be due to the relatively small size of the corpus, their online chat context origin, and the inherent short length of the messages due to in-game constraints. The seed for the project at hand was planted by a desire to integrate

these areas of research into a broader as well as more in-depth analysis. There is an inherent overlap: the demographic characteristics which I chose to analyze here were present in my first demographic analysis and I examined the same NLP metrics in this analysis as I did in my first corpus analysis.

. . . and how it has informed the current work

Hindsight has made it clear to me that my initial demographic analysis was far too ambitious to the point of likely producing incorrect or misleading results. Some of the metrics that I recorded and analyzed were highly mutable, such as in-game clothing and the name tag the individual was wearing, as both of these can be changed at will, and I had opted to omit a very large portion of the population, Toons who had not yet entered the final area of the game.[1] I kept these previous methodologically questionable decisions in mind in terms of my research design for the current project.

It is important to note that all previous analyses have been performed in Python while the work at hand was conducted in R (v4.1.2). There are a multitude of reasons for this, but my primary three were that I wanted to take advantage of R's data visualization capabilities in terms of plots and tables, improve my R skills, and utilize the R Markdown functionality to craft and format a report about my work.

Dataset

Collection

The dataset utilized to carry out the current project is a corpus consisting of 4000 chat messages and various demographic details of the speaker of each message collected by hand between August 2022 and January 2023. Toontown data must be collected by hand, a somewhat laborious task which has limited the scope of my research in this area.[2]

Structure

The core dataset consists of five columns: **message**, the text of an in-game chat message, and four demographic characteristics of the speaker. I will explain briefly to what each of these metrics pertains as they will be unfamiliar to those who are not acquainted with Toontown. **species** refers to the animal species of the Toon, **laff** refers to the Toon's total health points, referred to in Toontown as Laff points, **gender** refers to the Toon's gender, and **missing** refers to which of the seven types of Gags the Toon does not possess, as it is only possible to obtain six in total. This is a significant reduction in demographic metrics from my demographic study published in January 2022. The metrics **colour**, **dl**, and **leg_colour**, referring to the colour of the Toon, whether or not their legs were the same colour as the rest of their body, and the alternate leg colour if present, **organic**, the Gag track to which the Toon had opted to give extra power by growing it in their garden, **name_tag**, the font style of the name which floats above the Toon's head, and **flippy**, a Boolean expressing whether or not a Toon was wearing a certain relatively exclusive style of shirt, were omitted from the present study. My primary motivation for the omission was that all of the omitted characteristics are mutable. An individual is able to quite easily change their Toon's colour, name tag, clothing, and which Gag track they choose to grow organically. I believe that basing such a large portion of my previous analysis on mutable demographic characteristics was not methodologically sound and thus may have led to incorrect or misleading results because it inherently introduces repeated observations. I dropped exact duplicate rows from the final dataset, but this is somewhat useless when easily mutable characteristics are used as factors in

[1] In Toontown, a Toon may possess six out of seven Gag tracks, or varieties of Gags. When a Toon acquires their sixth and final Gag track, they enter the final gameplay area. I had included which Gag track a Toon was missing as a demographic metric and had therefore deemed it necessary to record demographic information exclusively for Toons who had progressed significantly in the game, omitting a very large section of the population.

[2] I suspect that the true accuracy of my previous Toontown data analysis work and the lack of conclusivity or meaningful results delivered by various Toontown projects has been heavily affected by the fact that my datasets are quite small by the standards of the field of data analysis. However, due to the sheer amount of time it can take to type out each observation by hand, there does not seem at present to be a way to easily change this characteristic of the corpora. My Toontown research stretches over a large timescale for the time being but will presumably slowly become easier and more efficacious as larger corpora are progressively constructed.

the analysis. A Toon who returns to their estate to change which name tag they were utilizing is counted as an entirely different Toon if witnessed again a short time later.

All demographic characteristics included in the present study with the exception of **laff** are immutable. It is not possible to change the species or gender of your Toon once they have been created, nor is it possible to change your mind about which Gags you have once the sixth and final track has been received. It is possible to increase your Laff by completing tasks and activities, and this is the main way by which progression through the game's storyline is measured. However, I made the choice to include it as a metric in the present study because it turned out to be a very important predictor in my previous demographic study and the issue of repeated observations is not as damning of an issue here due to the fact that I am not concerned with representing the demographic characteristics of a sample of a larger population. It also allows the analysis to represent any changes in NLP metrics as a Toon increases in Laff. It is additionally important to note that I decided to expand the scope of the population to analyze Toons at all Laff levels, as in my demographic study I only included Toons who had received their final Gag track in order to include **missing_track** as a factor. A similar metric is included here as **missing** but it now includes a level *nm* which indicates "not maxed", referring to a Toon who has less than six Gag tracks.

Methods

Preprocessing / cleaning

Prior to beginning the NLP analysis, the corpus was cleaned and preprocessed. Data cleaning consisted of finding typos which were inserted into the corpus during data collection. Preprocessing consisted of removing punctuation from chat messages, rendering all messages lowercase, and trimming any whitespace. The **stringr** package was utilized for preprocessing.

```
unique(corpus$species)
```

```
##  [1] "cat"       "duck"      "rabbit"    "deer"      "mouse"     "dog"
##  [7] "crocodile" "monkey"    "pig"       "bear"      "horse"     ""
## [13] "monk"      "deeer"     "cat`"      "moiuse"    "bean"
```

```
corpus[corpus$species == 'monk',]$species <- 'monkey'
corpus[corpus$species == 'deeer',]$species <- 'deer'
corpus[corpus$species == 'moiuse',]$species <- 'mouse'
corpus[corpus$species == 'bean',]$species <- 'bear'
corpus[corpus$species == 'cat`',]$species <- 'cat'
```

```
corpus$message <- corpus$message %>% str_replace_all("[[:punct:]]", "")
corpus$message <- corpus$message %>% tolower()
corpus$message <- corpus$message %>% trimws(which='both')
```

The columns **species**, **gender**, and **missing** were converted to factors as they encode categorical data and **laff** was converted to a numeric vector.

```
corpus$species <- as.factor(corpus$species)
corpus$gender <- as.factor(corpus$gender)
corpus$missing <- as.factor(corpus$missing)

corpus$laff <- as.numeric(corpus$laff)
```

NLP metric calculations

After cleaning and preprocessing were performed, various NLP metrics were calculated. Each message was split into a vector of its component word tokens and these vectors were assigned to the column **tokens**.

```
corpus$tokens <- corpus$message %>% strsplit('[ ]')
```

For each chat message, word count and character count were calculated and assigned to the columns **wordlen** and **charlen**.

```
corpus$wordlen <- corpus$tokens %>% lengths()
corpus$charlen <- corpus$message %>% nchar()
```

Sentiment was calculated for each message using a two-step process. First, a numeric sentiment value was calculated for each message using the **vader** package, a sentiment analysis tool which is specifically attuned to text from social media contexts. Each numeric sentiment value was appended to a vector which was subsequently assigned to the **sentiment** column. Each message was designated either positive, neutral, or negative based on its numeric sentiment value. A numeric value of zero indicates a message deemed by the algorithm to be neutral, a positive value indicates a message of positive sentiment, and a negative value indicates a message of negative sentiment. These designations were appended to a vector which was assigned to the column **sentimentjudgment** and converted to a factor.

```
messages <- corpus$message
sentiment.values=c()
for (m in messages) {
  s <- as.numeric(get_vader(m)[2])
  sentiment.values <- append(sentiment.values,s)
}
corpus$sentiment <- sentiment.values

sentiment.judgments = c()

for (val in list(sentiment.values)) {
  if (val == 0) {
    sentiment.judgments <- append(sentiment.judgments,'neutral')
  } else if (val>0) {
    sentiment.judgments <- append(sentiment.judgments,'positive')
  } else {
    sentiment.judgments <- append(sentiment.judgments,'negative')
  }
}

corpus$sentimentjudgment <- sentiment.judgments
corpus$sentimentjudgment <- as.factor(corpus$sentimentjudgment)
```

Subjectivity metrics, referring to the degree of bias or personal opinion in a text, for each message unfortunately had to be calculated externally in Python, read into R as a text file, and appended to the corpus as the **subjectivity** column. I repeatedly attempted to integrate TextBlob into my R environment in order to calculate the values in R, but I was unable to do so. I had intended to explain in this section the errors which I encountered during my attempts to integrate TextBlob, but upon returning to my previous code to troubleshoot, it appears that I am now successfully able to integrate the package into R. I remain unsure as to why it was not initially possible to do this. Using TextBlob in an R environment is the anticipated method of subjectivity analysis for future NLP research.

For this investigation, I used the **pandas** package to read the corpus into Python as a data frame and **textblob** to calculate subjectivity for each message. The calculations were subsequently written to a text file.

```
import os
import pandas as pd
import textblob
```

```
messages = corpus[0]
sentiments = []

for m in messages:
    sentiment = textblob.TextBlob(m).sentiment.subjectivity
    sentiments.append(round(sentiment,3))

with open('sentiments_nlp2.txt','w') as f:
    for s in sentiments:
        f.write(f"{s}\n")
```

The text file of subjectivity calculations was read into R and used to create a column called **subjectivity**.

```
sentiments <- read.delim('sentiments_nlp2.txt',header=F,sep='\n')
corpus$subjectivity <- sentiments
```

Results

Population details

It is very important to note here that all figures below refer to frequencies of chat messages across various demographics as opposed to sizes of demographic groups within the population. For example, $n = 1336$ in the table below indicates that 1336 of the 4000 chat messages collected were uttered by cats, not that 1336 individual cats were observed.

Species distribution

```
##      species    n percentage
## 1        cat 1336       33.4
## 2        dog  590       14.8
## 3       deer  421       10.5
## 4      mouse  415       10.4
## 5       duck  329        8.2
## 6     rabbit  273        6.8
## 7       bear  205        5.1
## 8  crocodile  165        4.1
## 9     monkey  117        2.9
## 10       pig   78        2.0
## 11     horse   71        1.8
```

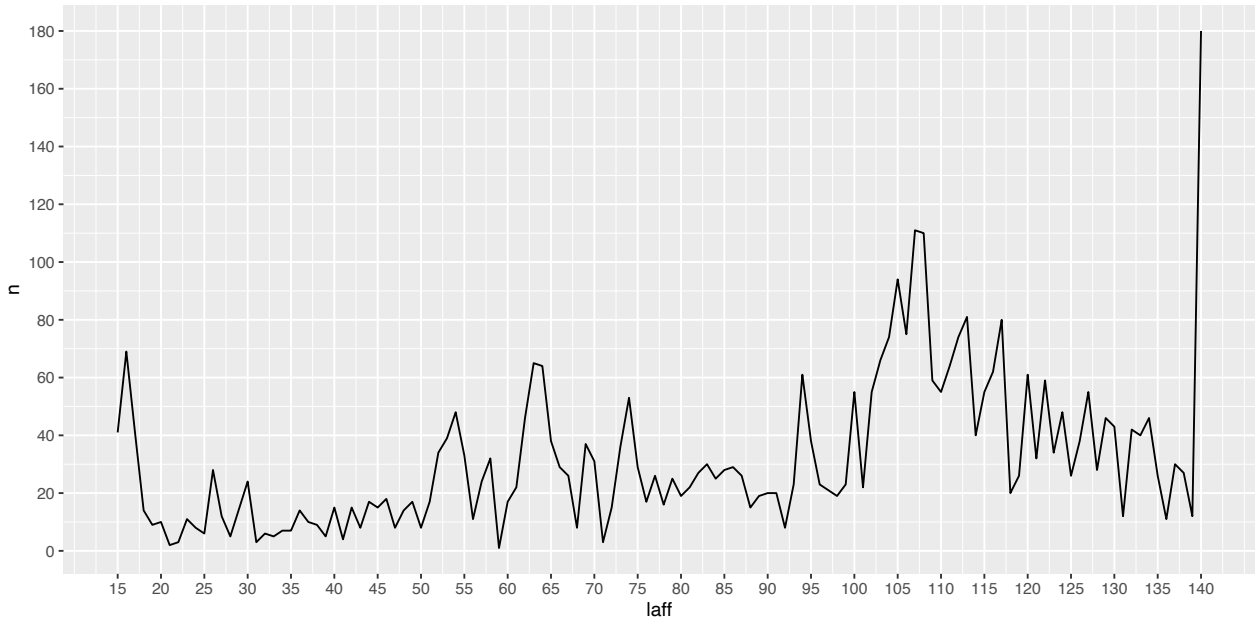
Gender distribution

```
##   gender    n percentage
## 1      f 2157       53.9
## 2      m 1843       46.1
```

Missing Gag track distribution

```
##   missing    n percentage
## 1    trap 1305       32.6
## 2    drop 1220       30.5
## 3      nm  918       23.0
## 4 toon-up  269        6.7
## 5    lure  221        5.5
## 6   sound   67        1.7
```

Laff point distribution These population-wide results displayed below are strikingly similar to those of my original Toontown Rewritten demographic analysis. I had assumed that this would not be the case due to the fact that this analysis contains repeated observations of the same Toon, as each message was treated as a new observation even if the Toon who said it had previously spoken, while the purely demographic analysis did everything in its power to avoid the repetition of individual Toons in the dataset. However, percentages of each species and gender across the population are nearly identical between the two analyses. Missing track percentages diverged due to the inclusion of a new factor level pertaining to Toons who had not yet received their final Gag track, as did Laff point distribution as the maximum Laff points a Toon can obtain has increased from 137 to 140 since the publication of my first demographic analysis. Despite this change, the shape of the line graphs is quite similar, with a large jump between approximately 100 and 115 Laff and a sudden sharp peak at maximum Laff.



NLP metrics

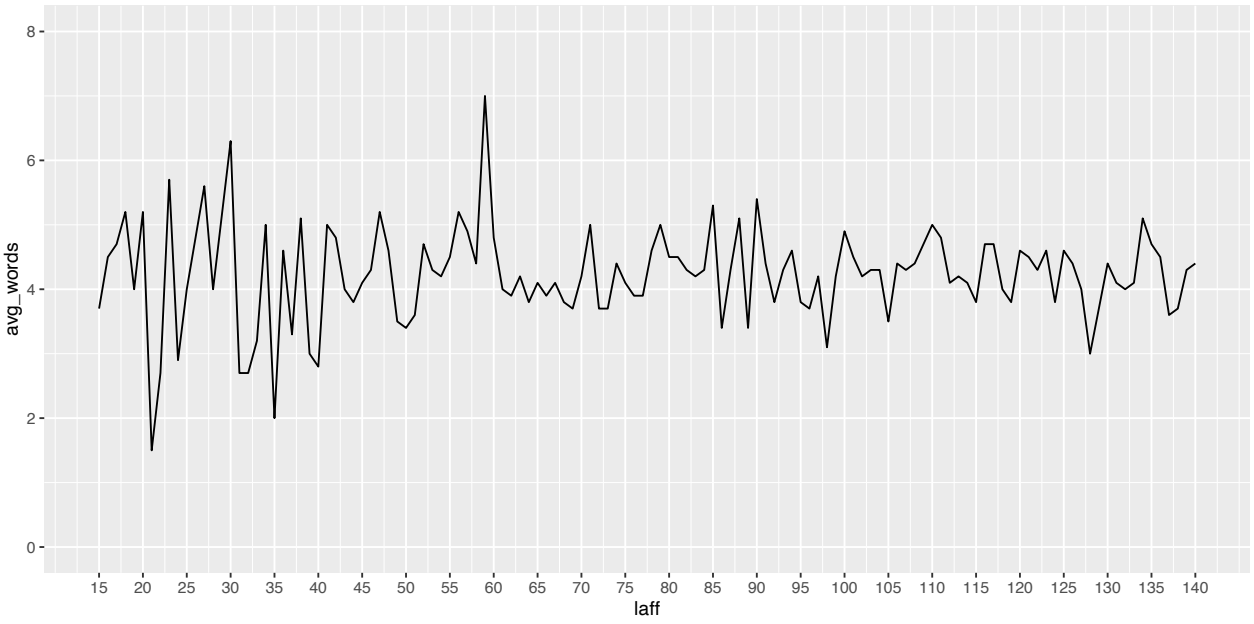
Word and character counts Mean values for character and word count across the levels of **species**, **gender**, and **missing** are nearly identical. The mean word and character counts from my first corpus analysis are similar but slightly lower, sitting at 3.5 and 16.5, respectively.

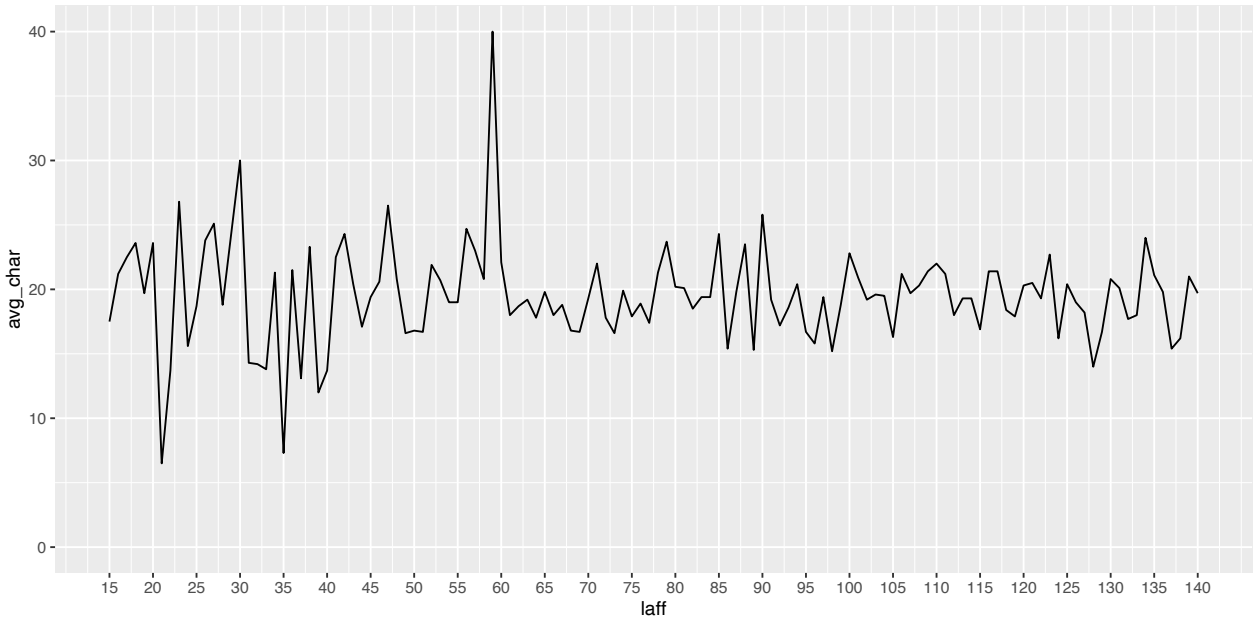
```
##      species avg_words avg_char
## 1       bear       4.4     19.9
## 2        cat       4.3     19.5
## 3  crocodile       4.5     20.7
## 4       deer       4.3     20.0
## 5        dog       4.3     19.8
## 6       duck       4.3     19.7
## 7      horse       3.8     17.6
## 8     monkey       4.0     18.5
## 9      mouse       4.2     19.1
## 10       pig       4.5     20.4
## 11    rabbit       4.3     19.2
```

```
##   gender avg_words avg_char
## 1   male       4.3     19.7
## 2 female       4.3     19.4
```

```
##   missing avg_words avg_char
## 1      nm       4.3     20.3
## 2 toon-up       4.0     18.0
## 3    trap       4.3     19.9
## 4    lure       4.0     18.0
## 5   sound       4.5     20.6
## 6    drop       4.3     19.3
```

Word and character counts and Laff points I was very intrigued by the sudden leap in average characters as displayed on the second plot below. Upon filtering the corpus based on the outlier, I discovered that there was only one single message in the corpus from a Toon with 59 Laff points. Throughout this analysis, when NLP metrics are compared across Laff values, many of them stabilize as Laff increases, presumably due to the amount of messages per Laff value tending to increase as Laff increases.






```
laff_char %>% filter(avg_char==40)
```

```
##   laff avg_char
## 1   59       40
```

```
corpus %>% filter(laff==59) %>% summarise(count=n())
```

```
##   count
## 1     1
```

Sentiment Population-wide results for average message sentiment and percentages for each of the three levels of **sentimentjudgment** are somewhat similar to those observed in my first NLP analysis. Average sentiment for this analysis is higher, 0.111 as opposed to 0.061 from the first analysis, which makes sense as there is a larger proportion of positive messages and a smaller proportion of neutral and negative messages in the current corpus. The first corpus analysis yielded 29.1% positive messages, 57.7% neutral messages, and 13.2% negative messages while the present analysis has yielded 35.0% positive messages, 53.1% neutral messages, and 11.9% negative messages.

```
##   avg_sentiment
## 1         0.111
```

```
##   sentiment sent_percents
## 1  positive          35.0
## 2   neutral          53.1
## 3  negative          11.9
```

Sentiment and species

```
##      species avg_sentiment
## 1       bear         0.143
## 2        cat         0.093
## 3  crocodile         0.178
## 4       deer         0.111
## 5        dog         0.082
## 6       duck         0.122
## 7      horse         0.058
```

```
## 8     monkey         0.109
## 9      mouse         0.122
## 10       pig         0.121
## 11    rabbit         0.177
```

```
##      species positive neutral negative
## 1       bear     35.6    55.1      9.3
## 2        cat     32.2    54.3     13.5
## 3  crocodile     47.9    44.8      7.3
## 4       deer     34.4    54.2     11.4
## 5        dog     32.9    52.9     14.2
## 6       duck     37.4    53.2      9.4
## 7      horse     23.9    62.0     14.1
## 8     monkey     32.5    57.3     10.3
## 9      mouse     36.1    51.6     12.3
## 10       pig     38.5    48.7     12.8
## 11    rabbit     43.6    49.8      6.6
```

Sentiment and gender

```
##   gender avg_sentiment
## 1   male         0.124
## 2 female         0.096
```

```
##   gender positive neutral negative
## 1   male     32.2    55.2     12.5
## 2 female     37.3    51.4     11.4
```

Sentiment and missing track

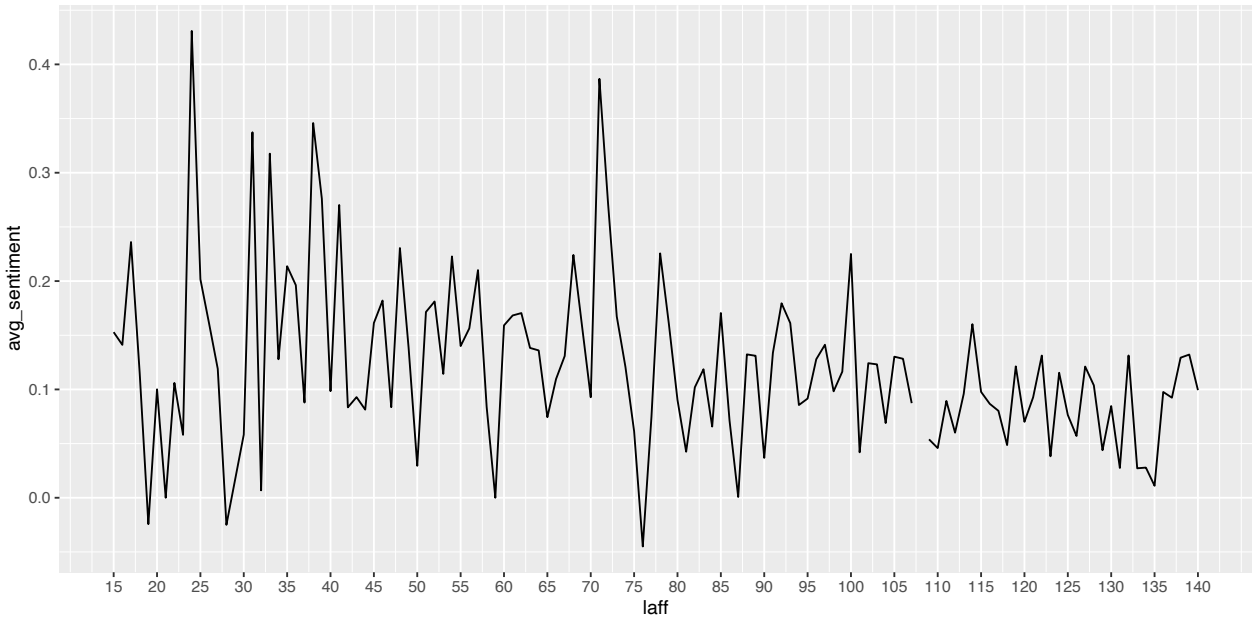
```
##   missing avg_sentiment
## 1      nm         0.152
## 2 toon-up         0.069
## 3    trap         0.101
## 4    lure         0.113
## 5   sound         0.070
## 6    drop         0.101
```

```
##   missing positive neutral negative
## 1      nm     41.6    48.6      9.8
## 2 toon-up     29.0    55.4     15.6
## 3    trap     33.8    54.0     12.2
## 4    lure     36.2    54.3      9.5
## 5   sound     29.9    55.2     14.9
## 6    drop     32.5    54.8     12.6
```

Many demographic groups in this analysis have very few messages in the corpus and thus their results may initially appear to be outliers or notable in some way. For example, messages uttered by horse Toons have by far the lowest average sentiment, sitting at 0.058, but horses make up only 1.8% of the corpus. It is thus impossible to determine with the current data if this divergence is systematic in any way. This observation could also be made for Toons who are missing either Toon-up or Sound, whose average sentiment values are 0.069 and 0.070, respectively, noticeably lower than those of the other levels of **missing**, but these groups combined comprise only 8.4% of the corpus. More data will be necessary to determine if any of the smaller demographic groups show consistently divergent NLP results.

Sentiment and Laff points As stated previously, message frequency tends to increase as Laff increases, so the variance in average sentiment values stabilizes as Laff increases. An identical trend is observed below

for average subjectivity across Laff values.



Subjectivity Strikingly, mean subjectivity from my first Toontown corpus analysis was 0.195 and mean subjectivity from the current analysis is 0.194. It is intriguing that sentiment metrics jumped noticeably in an overall positive direction while subjectivity remained nearly identical. One might suppose that subjectivity would increase as overall sentiment becomes more positive, indicating movement towards more subjective language, but this appears not to be the case for these data.

```
##   avg_subjectivity
## 1            0.194
```

Subjectivity and species

```
##      species avg_subjectivity
## 1       bear            0.198
## 2        cat            0.194
## 3  crocodile            0.213
## 4       deer            0.212
## 5        dog            0.187
## 6       duck            0.200
## 7      horse            0.136
## 8     monkey            0.162
## 9      mouse            0.206
## 10       pig            0.167
## 11    rabbit            0.173
```

Subjectivity and gender

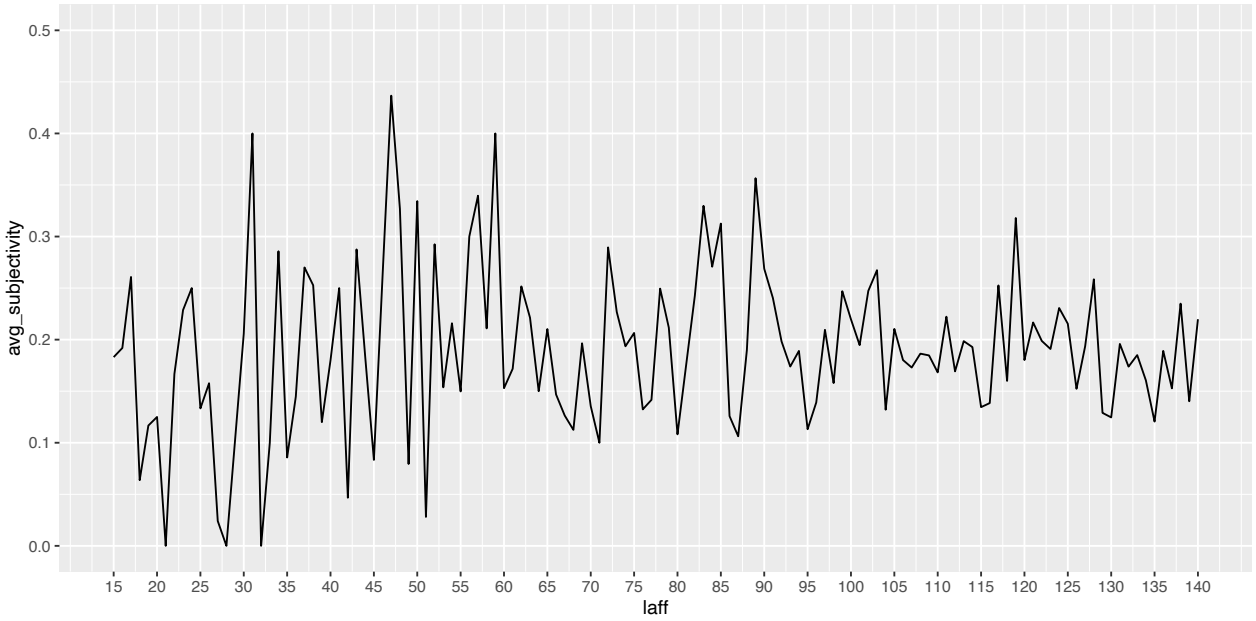
```
##   gender avg_subjectivity
## 1   male            0.191
## 2 female            0.196
```

Subjectivity and missing track

```
##   missing avg_subjectivity
## 1      nm            0.197
## 2 toon-up            0.172
## 3    trap            0.191
## 4    lure            0.157
## 5   sound            0.218
## 6    drop            0.205
```

Average message subjectivity is very consistent across the three factors. The sole result which could arguably be deemed divergent is the average subjectivity for horses at 0.136, but as explained above, horses have very few messages in the corpus so it is not possible at this point to make sweeping judgments about this result.

Subjectivity and Laff points Once again, there appears to be stabilization as Laff increases.



Most frequent tokens

```
##     word count
## 1      i   631
## 2    the   436
## 3  beans   394
## 4    you   377
## 5      a   343
## 6     to   324
## 7    for   273
## 8     is   206
## 9     my   176
## 10    it   164
## 11    ty   155
## 12   lol   154
## 13    in   152
## 14    im   150
## 15    me   150
## 16    do   139
## 17     u   134
## 18   and   133
## 19  need   125
## 20  have   122
```

The most frequently occurring tokens differ very little across the two analyses. Below are the most frequent tokens from my first analysis.

```
##     word count
## 1      i   594
## 2    the   309
## 3    you   301
## 4      a   276
## 5     to   239
## 6      u   201
## 7     my   175
## 8     it   170
## 9     is   167
## 10 beans   160
## 11    me   146
## 12    do   140
## 13    im   139
## 14   and   132
## 15   lol   127
## 16   for   126
## 17   are   119
## 18  like   118
## 19    so   115
## 20  that   115
```

Discussion

Few notable NLP-related results were yielded by this analysis. Mean values for character and word count, sentiment, and subjectivity were nearly identical across **species**, **gender**, and **missing** population groups. These values, as well as the most frequently occurring tokens in the corpus, were also extremely similar between my first Toontown Rewritten NLP analysis which did not take demographic factors into consideration.

Values of NLP metrics across Laff points appeared initially to fluctuate somewhat heavily but became far more uniform as Laff increased. Further inspection revealed that this pattern is likely owing to the fact that the amount of messages per Laff value tended to be larger at higher Laff values. It is quite interesting that population metrics from a purely numerical standpoint are very similar to those of my demographic analysis given that in the analysis at hand the amount of messages per demographic group was being counted as opposed to the group population itself.

The primary limitation of this analysis is the size of the corpus. Many groups of the population had very few messages and it is thus difficult to accept any broad conclusions from these results. Concluding that certain groups display divergent linguistic behaviour will likely lead to type I error when some of the population subgroups in question possess only a handful of data points. Something of which to also be wary is the potential implication of repeating individual Toons within the dataset and the inherent difficulty of not doing so given the nature of the data being collected. It is possible that some bundles of messages were all spoken by only a handful of distinct individuals which thus casts doubt on the reliability of applying obtained results to larger populations. Additionally, Toons increase their Laff points as they complete tasks and progress through the game's storyline, which may raise questions about the difficulty of teasing apart changes in the individual from changes across subgroups of the population. However, as more data is collected, it is possible that these potential changes may somewhat even out, analogously to how various NLP metrics stabilized in the present analysis as amounts of messages per Laff showed a tendency to increase concurrently with Laff.

Significantly larger amounts of data will be an absolute necessity for further investigation into this topic. This is challenging as data must be collected by hand which is quite time-consuming, as well as unpredictable due to the game's fluctuating population and day-to-day changes in player behaviour. Despite these factors, further investigation and future research are planned. I intend to create a perpetually expanding database of Toontown Rewritten NLP data and information and would also like to eventually delve into machine learning and interactive contexts with this data. Specific machine learning directions are uncertain at this point and will likely be shaped by the results of more extensive analyses with expanded corpora. My first demographic analysis also pointed to divergent population groups defined by bundles of minority demographic characteristics; I would like to investigate if these groups also display divergent linguistic behaviour. This shall be done as more data is collected.

Conclusion

The primary goal of this analysis was to integrate demographic and linguistic research related to Toontown Rewritten. Previous demographic research in this sphere has revealed many interesting patterns and trends while NLP research has been less informative and it was thus hypothesized that using demographics to inform NLP research could produce interesting and consequential results. To test this, a corpus consisting of 4000 chat messages and some demographic information corresponding to the speaker of each message was cleaned and preprocessed and population details and a set of NLP metrics were subsequently calculated. Very few notable results came to light and the results in general bore a strong resemblance to results produced by my previous work. The lack of findings in the present analysis nevertheless does hold important and telling implications: much larger corpora are necessary for future work in this sphere, and the lack of results here cannot be understood as any sort of definitive finding or conclusion. Corpus size, I believe, was by far the primary limitation in this analysis, and the next step for future research is to begin the construction of a much larger corpus integrating demographic and linguistic data like that which was analyzed in this study. This is an absolute necessity to continue to investigate that which may lie at the intersection of these two fields as well as to branch out into other avenues using these data, such as machine learning. As mentioned in the previous section, the results yielded by my demographic analysis published at the beginning of 2022 pointed to the existence of various divergent groups of the population bounded by combinations of certain demographic characteristics, and these groups may indeed be somehow linguistically divergent as well. This is my preferred next direction for my research, and I plan to begin data collection once again very shortly in order to investigate further.

References

Ciereszynski, E. 2022a. "An Exploration of Toontown Rewritten Demographics." https://github.com/c-z-c-z/data-analytics/tree/main/Toontown-Rewritten-demographics.

———. 2022b. "Toontown Rewritten Corpus Analysis." https://github.com/c-z-c-z/natural-language-processing/tree/main/Toontown%20Rewritten%20corpus%20analysis.

———. 2022c. "LDA and NMF Toontown Corpus Analysis." https://github.com/c-z-c-z/natural-language-processing/tree/main/LDA%20and%20NMF%20Toontown%20corpus%20analysis.

Loria, Steven. 2018. "textblob Documentation." *Release 0.15* 2.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Roehrick, Katherine. 2020. *vader: Valence Aware Dictionary and sEntiment Reasoner (VADER)*. R package version 0.2.1.

Silge, Julia, and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1 (3).

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 978-3-319-24277-4. Springer-Vertag New York.

———. 2022. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.0.

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.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.