# Predicting Age Group of Social Media Users Based on Language Use

MIS 464 Final Project Report

Libby Merchant
Sydney Huxtable

6 May 2022

## Introduction

Social media has allowed researchers to tap into trends among different demographics, including class, age, and gender, at an unprecedented scale (Sap, et al., 2014). While this has enabled data-driven research in the social sciences (Morgan-Lopez et al.), it has also been leveraged by social media platforms to build profiles on users that are sold for advertising purposes. Our previous research into this topic gave us direction for this project. More specifically, our research found a study that determined the top words and phrases used in each age bin and the highest and lowest correlation between the language and age bins. Some studies we researched extracted ages from self-reports in tweets, while others have analyzed personality through online language use. In a study by Sap et al. (2014) called the World Well Being Project, the authors developed age predictive lexica using penalized linear regression from Facebook users who shared their status updates and age. The lexica consist of the top 100 predictive words and phrases across 5 different age bins, with each word assigned a “weight” (Sap et al., 2014).

## Project Idea

Age is one such piece of information that can be deduced by these algorithms. While this information can be gleaned from many aspects of a user’s online activity, including “friend” connections and posts they interact with, it can also come from something as simple as the user’s language. Our idea for our analysis is to implement a supervised machine learning algorithm to predict the age groups of social media users based on individual language use. In this study, language use will be defined as a person’s lexicon, spelling, and use of emoticons. A lexicon is specifically the dictionary or vocabulary of a person. This information is very applicable to business, sociology, and psychology because there are many interesting patterns of language across age groups. Seeing as age prediction is likely a standard practice for social media companies, there are many opportunities for advertising to different age groups.

Twitter molds an accurate profile based on information users share publicly, which then can be used by third-party companies for advertisement purposes. Twitter’s privacy policy allows the company to make public data from its users available to the world (Twitter, n.d.). This information includes one’s age, gender, and any other demographic information they decide to share on Twitter. In response to this, some users have opted to provide false data in place of the demographic data frequently sold to advertisers (Peersman, et al., 2011). However, social media continues to exercise a frightening level of power over individuals’ online choices, including health behaviors (Antionio, et al., 2017). This is made possible due to ever-advancing and ever-present machine learning algorithms that can deduce all the information needed for advertising simply from the user’s online activity. Our project is intended to show not only how to predict the ages of Twitter users based on their language use, but also to provide a warning about the dangers of public data.

## Methods

### Data Collection

The present study is far from the first to predict age groups of social media users; Nguyen et al. (2021) created a predictive model for both continuous age and age categories, while Sap et al.

(2014) developed a predictive lexica for age. As such, this study will draw from existing lexica made public by the World Well Being Project at the University of Pennsylvania, which is discussed in Sap et al. (2014). The data consists of four lexica in CSV format, corresponding to four age bins: 1) 13-18, 2) 19-22, 3) 23-29, and 4) 30+. Each lexicon includes the top 100 and bottom 100 words and phrases correlated with each age group and corresponding weights (see Appendix B for example).

The model predictions were carried out on datasets of tweets from 15 celebrities in the English language. Celebrities were considered a prime candidate for the model because their ages are public information, and existing datasets are readily available. This allowed us to assess the accuracy of our model without scraping Twitter for age and demographic information, which may or may not be falsely reported. For this study, tweets were sourced from two datasets published on Kaggle by Sakib (2022) and Github user @estorrs (2018).

Datasets were divided into separate CSV files for each celebrity, at which point they entered the data cleaning stage. Tweets were read from the CSV file into a Python program, which then compiled all tweets into a single string. At this point, the string was passed through the clean() function from the CleanText Python library to remove emojis. Then, all emoticons present (e.g. ":D") in the lexica were removed before applying the lower() function to the text. The emoticons were added back to the text afterward to preserve original casing. As words and phrases from the lexica were not lemmitized, we did not apply lemmitization techniques to the tweet data either. Lastly, the cleaned data was written to a new text file. This process was repeated for each celebrity.

**Analysis Methods**

Next, the text files produced from data cleaning were analyzed according to the lexica. First, the program iterated through the list of top 200 words and phrases from the lexicon, and counted the frequency of each within the cleaned text file. The model then implemented a form of linear regression, using the weights provided in the lexica as coefficients and the frequency of each word or phrase as the input variable. Our operations were based on the formula below:

Figure 1: *Linear Regression Formula*

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}$$

For our purposes, the coefficient $\beta$ shown in Figure 1 represents the weights from the lexicon, while X represents the count of each word as observed in the tweets. Both the weights and the word frequencies were converted to 200-dimensional matrices or "tensors" using the Pytorch library (see Appendix C for code excerpts). The program then performed element-wise multiplication of the two tensors using the torch.mult() method. The elements of the resulting tensor were summed using the torch.sum() method; this sum corresponds to the output Y in Figure 1. This process was repeated for each of the four lexica corresponding to four age bins. In total, the function produced four Y outputs.

Lastly, the outputs from our linear regression analysis were put through a softmax function using the Numpy library. The purpose of the softmax is to normalize the data so that they summed to one and could thus be treated as probabilities for each age bin. It is a generalization of the logistic regression (Sigmoid) function to accommodate multi-dimensional matrices. Our Python implementation of this was based on the formula below:

Figure 2: *Softmax Formula*

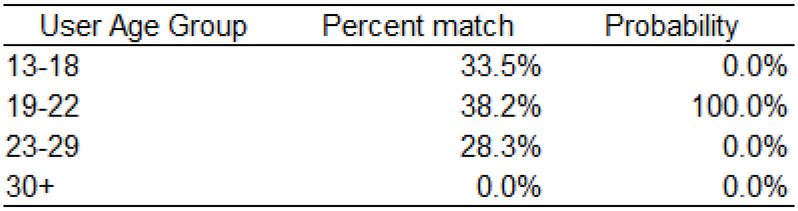
$$\boxed{s\ (x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} \quad e^{x_j}}}$$

Thus, the program generated four outputs, each representing the probability that the tweets belong to users of one of the four predefined age groups (see Appendix A for output table). The age group with the highest probability was interpreted to be the model's prediction. This was then compared with the actual age group of the celebrity at the time the Tweet was uploaded.

## Results

The tweets were originally assessed by percent match with each age group lexicon, but this was later exchanged for a measure of probability as calculated by the softmax function. A comparison of outputs for one user are displayed below:

Table 1: *Comparison of Percent Match and Softmax Function Results for a Single Tweet Set*



| User Age Group | Percent match | Probability |
|---|---|---|
| 13-18 | 33.5% | 0.0% |
| 19-22 | 38.2% | 100.0% |
| 23-29 | 28.3% | 0.0% |
| 30+ | 0.0% | 0.0% |

The softmax function produced less ambiguous output than the percent match prediction and also improved model performance. The model made correct predictions for nine of the fifteen celebrities, resulting in an accuracy rate of 60%, albeit on a small sample size (see Appendix A). The actual age makeup of celebrities was 13-18 (n=2), 19-22 (n=4), 23-29 (n=3), and 30+ (n=6). In contrast, the predicted age makeup was 13-18 (n=2), 19-22 (n=7), 23-29 (n=2), and 30+ (n=4).

The model's incorrect predictions were composed almost entirely of underestimations of the users' age. An exception to this was Harry Styles, who the model placed in a higher age bin than the actual age bin. Nevertheless, this pattern suggests the existence of a systematic error.

## Conclusion

There are several plausible origins of the model’s relatively high error rate. According to our analysis, the most likely issues are cross-platform transferability and the nature of public figures’ tweets. First, the original lexica were created from Facebook status data and published in late 2014. Many phrases in the lexica reflect these origins: attributes such as “repost,” “to your status,” and “fb friends” appeared frequently in the lexica, particularly for the upper age ranges. Such language is characteristic of Facebook statuses and may not transfer well into competing platforms such as Twitter, where terms such as “retweet” and “followers” are preferred. However, considering the time between when the lexica were published and our study was completed, it is also possible that the words and phrases have become outdated due to linguistic shifts.

Previous research has delineated the role of the Internet in driving change in written language (Eisenstein et al., 2014). In our study, some of the most commonly used elements in the lexica across age groups were smiley faces (e.g. “:)”), something that exhibited an inverse correlation with age according to the lexica. Other textual expressions of emotion (e.g. “hahaha”) were common across age groups, but had strong associations with younger age groups in the lexica. We suspect this disconnect to be a small contributor to the overall error rate of the model; however, it may also be evidence for a change in preferences of social media users. Those who grew up with social media and Internet access are now advancing into the older age bins, bringing along their innovative expressions of tone and emotion that were once impossible to convey through text. As such, older age bins are no longer restricted to the formal grammatical rules of written language.

Lastly, the choice to use celebrity tweet datasets to assess model accuracy may have, ironically, become a contributor to the model’s incorrect predictions. As celebrities and other “verified” Twitter accounts create tweets for a targeted audience and purpose, their content may differ fundamentally from tweets created by the average user. Language patterns were often more formal and less personal than in the lexica; the high visibility of a celebrity’s social media account compared to the general public likely restricts their ability to discuss personal topics.

Despite all these differences, the model correctly predicted age groups at times with startling accuracy, in some cases even assigning the user to their age bin with 100% chance (see Appendix A). With exploding advancements in algorithms over the years since the lexica were created, age prediction models have no doubt advanced far beyond the capabilities of our model. We believe that a sense of awareness surrounding predictive demographic models is of great importance for those who participate in social media, and we hope users will take these findings into consideration before posting their next Tweet.

## References

Eisenstein J, O'Connor B, Smith NA, Xing EP (2014) Diffusion of Lexical Change in Social Media. PLOS ONE 9(11): e113114. https://doi.org/10.1371/journal.pone.0113114

Giorgi, S. (2018). Word and Phrase Correlations [Data file]. Retrieved from https://github.com/wwbp/word_and_phrase_correlations/blob/master/csv/age_bins.13_18.gender_adjusted.rmatrix.top100s.csv

Morgan-Lopez AA, Kim AE, Chew RF, Ruddle P (2017). Predicting age groups of Twitter users based on language and metadata features. PLOS ONE 12(8): e0183537. https://doi.org/10.1371/journal.pone.0183537

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2021). *"How Old Do You Think I Am?" A Study of Language and Age in Twitter*. Proceedings of the International AAAI Conference on Web and Social Media, 7(1), 439-448. https://ojs.aaai.org/index.php/ICWSM/article/view/14381

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. 2011. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (SMUC '11). Association for Computing Machinery, New York, NY, USA, 37–44. https://doi.org/10.1145/2065023.2065035

Sakib, A. (2022). Top 1000 Twitter Celebrity Accounts [Data file]. Retrieved from https://www.kaggle.com/datasets/ahmedshahriarsakib/top-1000-twitter-celebrity-tweets-embeddings?resource=download

Sap, M., Park, G.J., Eichstaedt, J.C., Kern, M.L., Stillwell, D., Kosinski, M., Ungar, L.H., & Schwartz, H.A. (2014). *Developing age and gender predictive lexica over social media.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1146–1151. http://wwbp.org/papers/emnlp2014_developingLexica.pdf

Statista Research Department. (2022). Global Twitter user age distribution 2021. Statista. Retrieved May 5, 2022, from https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

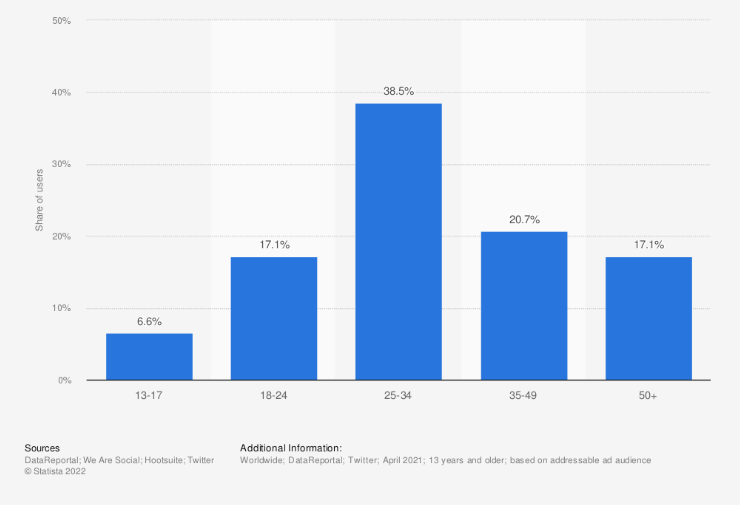
Statista Research Department. (2022). US Twitter reach by age group 2021. Statista. Retrieved May 5, 2022, from https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

Twitter Privacy Policy. (n.d.). Twitter. Retrieved May 5, 2022, from https://twitter.com/en/privacy

GitHub [@estorrs] (2018). Twitter Celebrity Tweet Sentiment [Data file]. Retrieved from https://github.com/estorrs/twitter-celebrity-tweet-sentiment/blob/master/results/celebrity_tweets_results.csv

# Appendix A

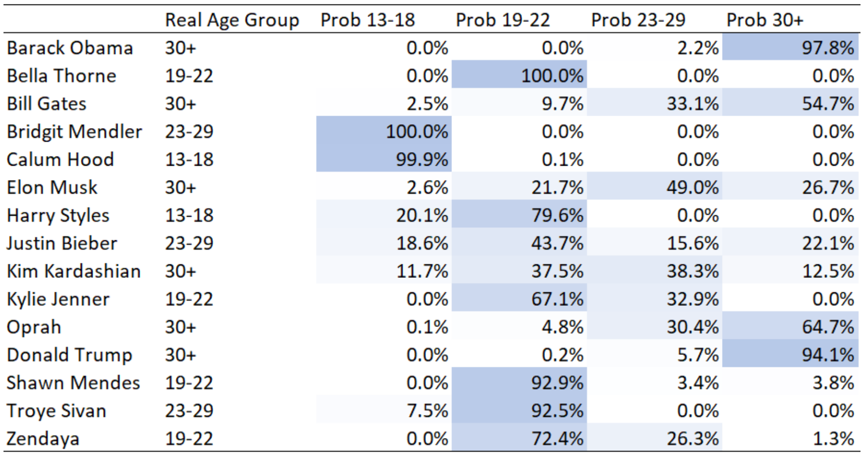
## Charts and Graphs



Sources
DataReportal; We Are Social; Hootsuite; Twitter
© Statista 2022

Additional Information:
Worldwide; DataReportal; Twitter; April 2021; 13 years and older; based on addressable ad audience

Figure A1. Distribution of Twitter users worldwide as of April 2021, by age group

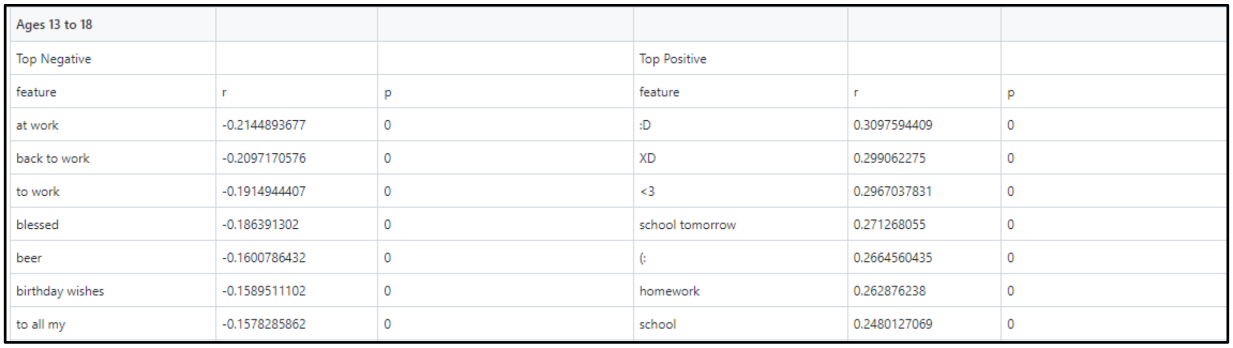


| | Real Age Group | Prob 13-18 | Prob 19-22 | Prob 23-29 | Prob 30+ |
|---|---|---|---|---|---|
| Barack Obama | 30+ | 0.0% | 0.0% | 2.2% | 97.8% |
| Bella Thorne | 19-22 | 0.0% | 100.0% | 0.0% | 0.0% |
| Bill Gates | 30+ | 2.5% | 9.7% | 33.1% | 54.7% |
| Bridgit Mendler | 23-29 | 100.0% | 0.0% | 0.0% | 0.0% |
| Calum Hood | 13-18 | 99.9% | 0.1% | 0.0% | 0.0% |
| Elon Musk | 30+ | 2.6% | 21.7% | 49.0% | 26.7% |
| Harry Styles | 13-18 | 20.1% | 79.6% | 0.0% | 0.0% |
| Justin Bieber | 23-29 | 18.6% | 43.7% | 15.6% | 22.1% |
| Kim Kardashian | 30+ | 11.7% | 37.5% | 38.3% | 12.5% |
| Kylie Jenner | 19-22 | 0.0% | 67.1% | 32.9% | 0.0% |
| Oprah | 30+ | 0.1% | 4.8% | 30.4% | 64.7% |
| Donald Trump | 30+ | 0.0% | 0.2% | 5.7% | 94.1% |
| Shawn Mendes | 19-22 | 0.0% | 92.9% | 3.4% | 3.8% |
| Troye Sivan | 23-29 | 7.5% | 92.5% | 0.0% | 0.0% |
| Zendaya | 19-22 | 0.0% | 72.4% | 26.3% | 1.3% |

Figure A2. Real vs. predicted age by model for Twitter users

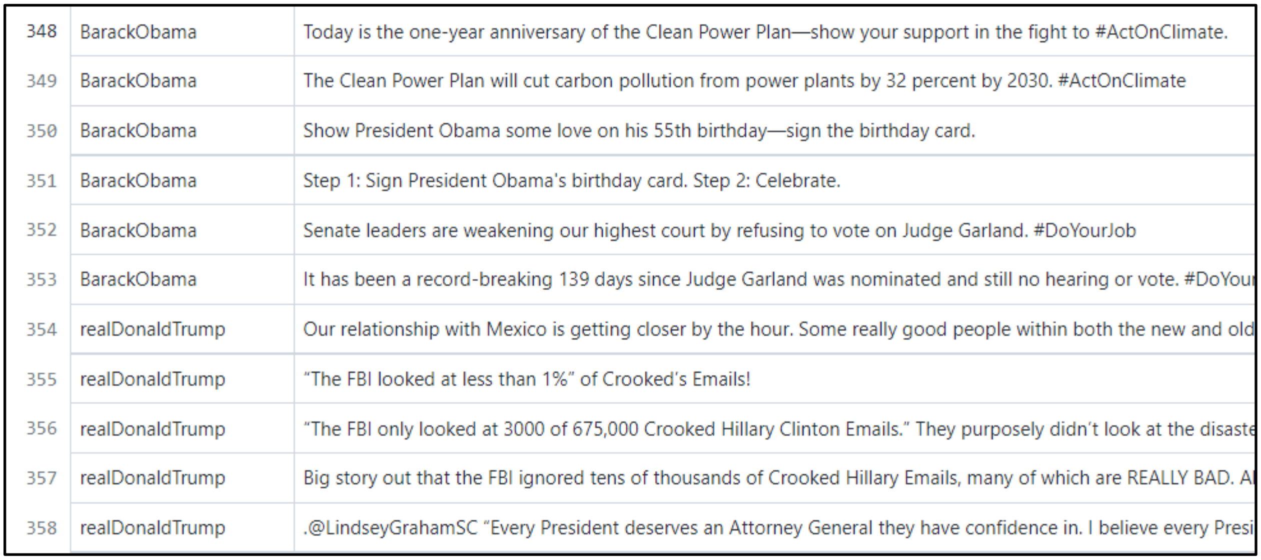
# Appendix B

Tweet and Lexica Data Examples



| | | | | | |
|---|---|---|---|---|---|
| Ages 13 to 18 | | | | | |
| Top Negative | | | Top Positive | | |
| feature | r | p | feature | r | p |
| at work | -0.2144893677 | 0 | :D | 0.3097594409 | 0 |
| back to work | -0.2097170576 | 0 | XD | 0.299062275 | 0 |
| to work | -0.1914944407 | 0 | <3 | 0.2967037831 | 0 |
| blessed | -0.186391302 | 0 | school tomorrow | 0.271268055 | 0 |
| beer | -0.1600786432 | 0 | (: | 0.2664560435 | 0 |
| birthday wishes | -0.1589511102 | 0 | homework | 0.262876238 | 0 |
| to all my | -0.1578285862 | 0 | school | 0.2480127069 | 0 |

Table B1. Excerpt of lexica for ages 13-18. Adapted from World Well Being Project Github: https://github.com/wwbp/word_and_phrase_correlations/blob/master/csv/age_bins.13_18.gender_adjusted.rmatrix.top100s.csv

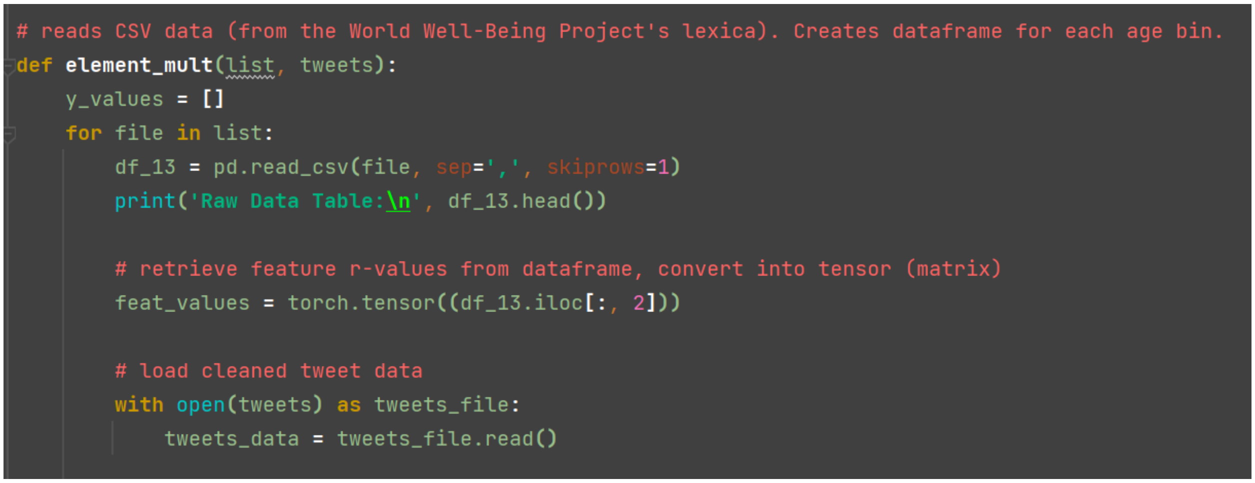


| | | |
|---|---|---|
| 348 | BarackObama | Today is the one-year anniversary of the Clean Power Plan—show your support in the fight to #ActOnClimate. |
| 349 | BarackObama | The Clean Power Plan will cut carbon pollution from power plants by 32 percent by 2030. #ActOnClimate |
| 350 | BarackObama | Show President Obama some love on his 55th birthday—sign the birthday card. |
| 351 | BarackObama | Step 1: Sign President Obama's birthday card. Step 2: Celebrate. |
| 352 | BarackObama | Senate leaders are weakening our highest court by refusing to vote on Judge Garland. #DoYourJob |
| 353 | BarackObama | It has been a record-breaking 139 days since Judge Garland was nominated and still no hearing or vote. #DoYour |
| 354 | realDonaldTrump | Our relationship with Mexico is getting closer by the hour. Some really good people within both the new and old |
| 355 | realDonaldTrump | "The FBI looked at less than 1%" of Crooked's Emails! |
| 356 | realDonaldTrump | "The FBI only looked at 3000 of 675,000 Crooked Hillary Clinton Emails." They purposely didn't look at the disaste |
| 357 | realDonaldTrump | Big story out that the FBI ignored tens of thousands of Crooked Hillary Emails, many of which are REALLY BAD. Al |
| 358 | realDonaldTrump | .@LindseyGrahamSC "Every President deserves an Attorney General they have confidence in. I believe every Presi |

Table B2. Twitter Celebrity Tweet Sentiment data snapshot

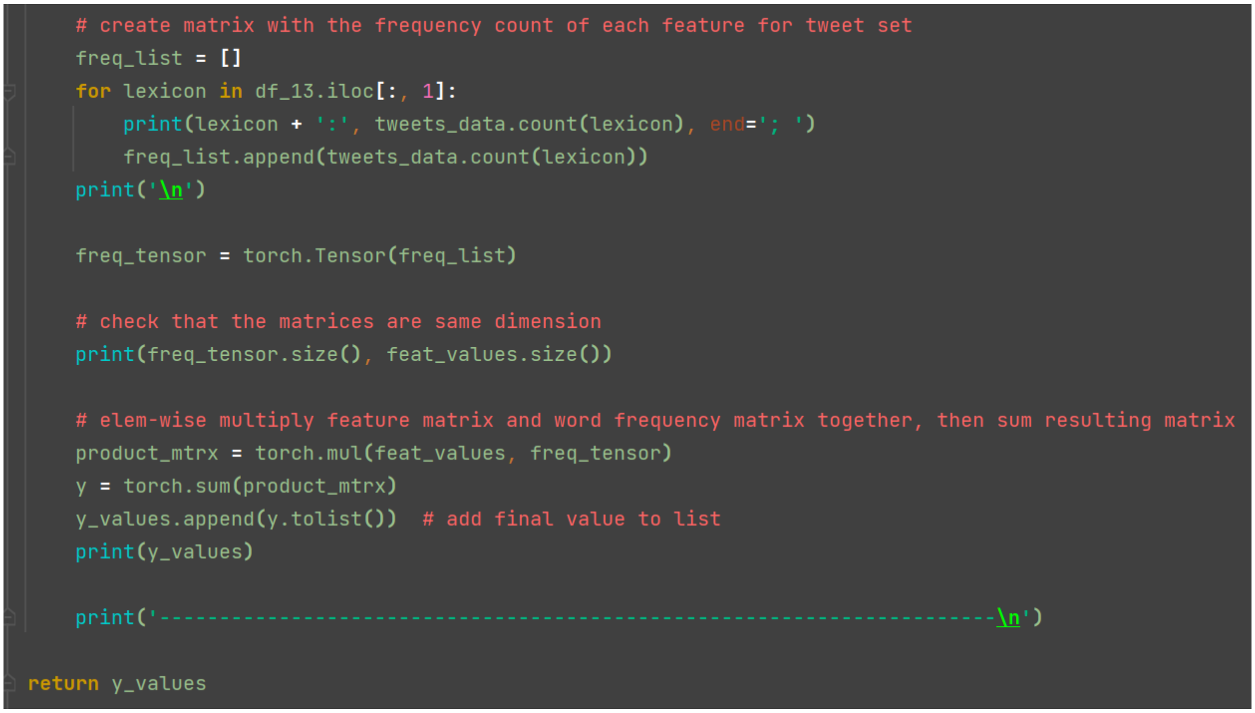
## Appendix C

Code Excerpts



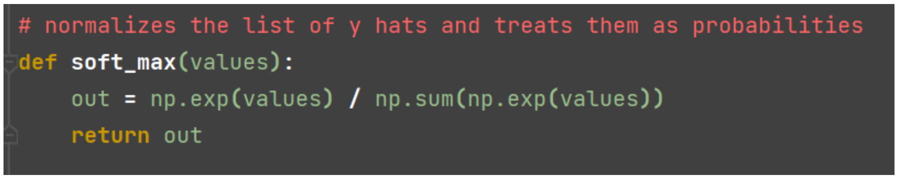
```python
# reads CSV data (from the World Well-Being Project's lexica). Creates dataframe for each age bin.
def element_mult(list, tweets):
    y_values = []
    for file in list:
        df_13 = pd.read_csv(file, sep=',', skiprows=1)
        print('Raw Data Table:\n', df_13.head())

        # retrieve feature r-values from dataframe, convert into tensor (matrix)
        feat_values = torch.tensor((df_13.iloc[:, 2]))

        # load cleaned tweet data
        with open(tweets) as tweets_file:
            tweets_data = tweets_file.read()
```

Figure C1: Python function for element-wise multiplication of two matrices



```python
        # create matrix with the frequency count of each feature for tweet set
        freq_list = []
        for lexicon in df_13.iloc[:, 1]:
            print(lexicon + ':', tweets_data.count(lexicon), end='; ')
            freq_list.append(tweets_data.count(lexicon))
        print('\n')

        freq_tensor = torch.Tensor(freq_list)

        # check that the matrices are same dimension
        print(freq_tensor.size(), feat_values.size())

        # elem-wise multiply feature matrix and word frequency matrix together, then sum resulting matrix
        product_mtrx = torch.mul(feat_values, freq_tensor)
        y = torch.sum(product_mtrx)
        y_values.append(y.tolist())  # add final value to list
        print(y_values)

        print('--------------------------------------------------------------------\n')

    return y_values
```

Figure C2: Continuation of Python function for multiplication of two matrices



```python
# normalizes the list of y hats and treats them as probabilities
def soft_max(values):
    out = np.exp(values) / np.sum(np.exp(values))
    return out
```

Figure C3: Python function for applying softmax to a list of values