Alexis Hancz
Korede Ajogbeje
QMSS 301
4 April 2023

Web Scraping and Sentiment Analysis of LGBTQ+ Demonstration Reddit Post

Introduction

This project's social issue of interest is student activism against Florida's "Don't Say Gay" bill, which bans public school teachers from educating kindergarten through third grade students on "sexual orientation or gender identity" (Diaz, 2022). The controversial bill, signed in March 2022 by Governor Ron DeSantis effective July 1, 2022, sparked protests in Florida highschools and others across the country. These efforts, which face conservative opposition, are towards protecting LGBTQ+ rights amidst an era of far-right initiatives to restrict freedom of gender and sexuality expression (Lavietes, 2022). This issue is interesting for its polarizing nature and forms the background of my analysis.

I scraped and conducted a sentiment analysis on a Reddit post thread titled, "For all those LGBTQ+ that are worried right now about the Right Wing anti-LGBTQ+ rhetoric, you have tons of support and allies…" . Attached is a video of a congregation of students walking out of their school in response to the "Don't Say Gay" bill. The post has an overall score of 40,200. I conduct a sentiment analysis analyzing the post's 2,600 comments to understand the digital conversations of importance to this social issue.

I. Scraping Reddit: Text Collection and Preprocessing

Using the PRAW library, I passed in my Reddit API credentials to access the post. Then I stored the post in an object. Using a for-loop, I appended each comment's body (skipping nested comments) to a list stored as a Pandas dataframe of 293 comments. Using a for-loop, I then appended the nested comments' bodies to the list, and the comment scores, IDs, and date created, and stored these properties in a new dataframe of 2,153 comments. Using the datetime module, I converted the date created to a Pandas datetime format. Then I deleted duplicate rows in the dataframe and stored the cleaned dataframe in a CSV file.

II. Sentiment Analysis: Preprocessing

To prepare for the sentiment analysis, I dropped rows with missing values and dropped unnecessary columns to keep only the comment body, ID, score, and the converted date created. Then, using the regular expression and Natural Language Toolkit modules, I removed unnecessary symbols and stopwords from the comments' bodies. Next, I defined and applied a function to clean the comment body column. I tokenized the comments using a lambda function and saved the result as a new dataframe. Finally, using PorterStemmer, I stemmed the tokenized words with the stemmer.stem(i) function and used a lambda function to iterate over each token. I saved the stemmed tokens as a new column in the original dataframe.

Polarity and Subjectivity Analysis

To begin my analysis, I created a WordCloud using matplotlib, WordCloud and the collections module. I split a concatenated list of comment bodies into individual words and saved them in a dictionary. I used the dictionary to generate the WordCloud and displayed it in a plot with the .imshow method. Then I applied Textblob's polarity and subjectivity calculation functions to each comment iteratively using a lambda function. I stored the results in two new columns in my original dataframe. I defined an if-elif function that classifies each comment's polarity calculation as positive (polarity > 0), neutral (polarity =0), or negative (polarity <0). I applied the function to my original dataframe and stored

the results as a new column. Finally, I classified each comment's sentiment calculation as subjective (sentiment >0.5), otherwise objective (sentiment <0.5) using a conditional statement. I stored the results in a new column in the original dataframe.

Descriptive Statistics

1. I ran the summary statistics of my polarity and subjectivity calculations and stored the results in a new object (**Figure 1, see appendix**). While most comments are objective and positive, there is considerable diversity. For example, while the average comment is slightly positive (mean= 0.079), sentiment varies significantly across comments (std= 0.284). Similarly, while the average comment leans moderately objective (mean= 0.369), objectivity varies significantly across comments (std=0.319).

2. I broke down all subjectivity (**Fig. 2**) and polarity classifications (**Fig. 3**) by frequency. TextBlob classified 1353 comments as objective (62.84%) and 800 comments as subjective (37.16%). There are almost twice as many objective comments as subjective comments, which surprised me due to the controversial nature of the post, which I thought likely to induce subjective opinions. TextBlob classified 937 positive comments, (43.52%), 794 neutral (26.88%), and 422 negative (19.60%). Positive comments occur about 1.6x more frequently than neutral comments, and about 2.2x more frequently than negative comments. This confirmed my hypothesis that the majority of comments would express positive sentiments because the post is housed in the "Made me Smile" subreddit.

4. I created a crosstab comparing polarity and subjectivity classification frequencies **(Fig. 4)**. The overwhelming majority of neutral comments are considered objective, which confirmed my hypothesis that it is unlikely for a comment to express neutrality and subjectivity. On the other hand, there is a relatively even split between subjectivity and objectivity for positive comments, while negative, subjective comments occur 1.85x more often than negative, objective comments.

5. I created frequency bar charts of polarity **(Fig. 5)** and subjectivity scores **(Fig. 6)**. Reflective of summary statistics, the majority of positive comments score between 0.01 and 0.24, while the majority of negative comments fall between -0.01 and -0.23. Also reflective of summary statistics, the majority of subjective comments score between 0.45 and 0.6, while the majority of objective comments score between 0.0 and 0.1.

6. I created a bar chart of the 15 most common words **(Fig. 7)**. The top three most common words, "people", "kids", and "like", have a range of 222, while the remaining top 12 words have a smaller range of 74, suggesting that frequencies vary more greatly between the top three than between the other top 12.

7. I created frequency tables displaying the subjectivity and sentiment classifications of the top 50 scoring comments **(Fig. 8)**. There is a relatively even split between objective (52%) and subjective (48%) top scoring comments. Subjective comments have an 11% greater share of top-scoring comments than in all comments overall, suggesting that audiences favor more opinionated, emotive comments. The sentiment breakdown is very similar to the breakdown in all comments overall.

Exploratory Analysis

1. I created a scatterplot comparing subjectivity and polarity scores using matplotlib. The red-blue gradient communicates the distance of the comment's polarity score from 0 (**Fig. 9)**. We are more likely to see high variance in polarity for comments with subjectivity scores between 0.8 and 1.0, perhaps because subjective comments have flexible interpretations or express strong opinions that are more likely to be divisive. Likewise, we see increasingly uniform polarity scores clustered in the -0.25 to 0.25 range as the subjectivity score decreases, perhaps because more objective comments are less likely to express strong opinions or emotions.

2. I created a scatterplot comparing subjectivity and comment score. The red gradient communicates the distance of the comment's subjectivity score from 0 (**Fig. 10**). Comments with scores between 300 and 1000 tend to lean strongly towards subjectivity, while a tiny minority of comments with scores above 3000 lean even more strongly towards objectivity. The audience of this post may favor comments expressing stronger, more emotive language contributing to subjectivity.

3. I created a scatterplot comparing polarity and comment score. The blue-purple gradient communicates the distance of the comment's polarity score from 0 (**Fig. 11**) Comments with scores between 0.01 and 1000 vary significantly in polarity, generally ranging from -0.60 to 0.80. However, comments with scores >1000 are increasingly associated with a polarity score from -0.20 to 0.40. The audience of this post may be split between favoring positive or negative comments.

4. Words that interest me in the WordCloud **(Fig. 12)** include "change", "generation", "love", "vote", "world", "rights", "children", and "hope", which suggest that the majority of comments are related to topics such as such as generational change, youth, social and personal values, and political rights, consistent with what I would expect in a discussion of LGBTQ+ rights. Overall, there is a clear message of optimism, determination and support for young people standing up for social transformation and civil rights, and that the Don't Say Gay issue is a major reflection of the functioning of society and the world at large.

Conclusions

I expected the comments on this post to be overwhelmingly positive and subjective because it lives in a subreddit called "Made me Smile". I was surprised to find that the majority (62.84%) are actually objective, and negative and neutral comments make up 56.49%, so a subreddit's focus on positive content doesn't necessarily determine the tone of the comments on a post. The controversial nature of the post may have encouraged this effect; perhaps some users engaged with the post to offer a different perspective, and/or the topic may have triggered a debate, leading to a wider range of opinions.

If I were to advise a researcher who studies online discussions of the "Don't Say Gay" bill, I would tell them that highly subjective opinions are likely to vary widely in their sentiment. Further, there is not a clear audience favorite between positive and negative sentiments regarding the issue. On the flip side, the more objective a statement is, the more likely it is to express neutrality. I would also advise that online audiences are likely to favor comments that either lean strongly subjective or occasionally strongly objective, and that powerful concepts such as "change," "generation," and "rights" are particularly important to commenters.

There are several implications for activists to consider here. One, Reddit users are more likely to engage with content that expresses a very strong opinion than factual information, which may exacerbate the polarization of the Don't Say Gay issue and prevent engagement with more flexible perspectives. Two, based on the WordCloud, users associate the Don't Say Gay walkouts with closely held political and social beliefs about the world and the future for children. While it may prove difficult to change people's opinions on such a matter, the concepts frequently mentioned reflect energy, passion and determination, all of which are useful to drive forward an organized response to Don't Say Gay.

There are a number of limitations in this sentiment analysis study. First, further research should investigate factors that may cause high variability in sentiment and objectivity scores. Second, more sophisticated analysis techniques are needed to analyze the strength or directionality of the relationship between sentiment and polarity. Third, the sentiment analysis tools used may be subject to biases that do not always capture the social and emotional nuances of the comments.

Appendix

Fig. 1

Summary statistics for polarity and subjectivity scores:

	polarity	subjectivity
count	2153.000000	2153.000000
mean	0.079288	0.368854
std	0.284455	0.319322
min	-1.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.400000
75%	0.200000	0.600000
max	1.000000	1.000000

Fig. 2 Polarity breakdown for all comments

Polarity	Positive Sentiment	Neutral Sentiment	Negative Sentiment
Count	937	794	422
Percentage	43.52	36.88	19.60

Fig. 3 Subjectivity breakdown for all comments

Subjectivity	Subjective	Objective
Count	800	1353
Percentage	37.16	62.84

Fig. 4 Crosstab comparing polarity and subjectivity classification frequencies

subjectivity2	Objective	Subjective
sentiment		
Negative Sentiment	148	274
Neutral Sentiment	772	22
Positive Sentiment	433	504

Fig. 5 Frequency bar chart of sentiment/polarity scores

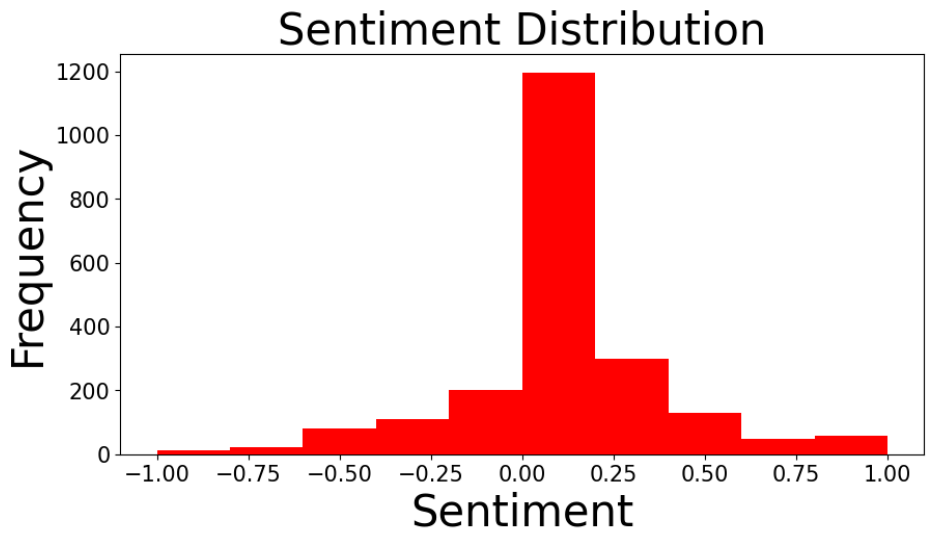


Fig. 6 Frequency bar chart of subjectivity scores

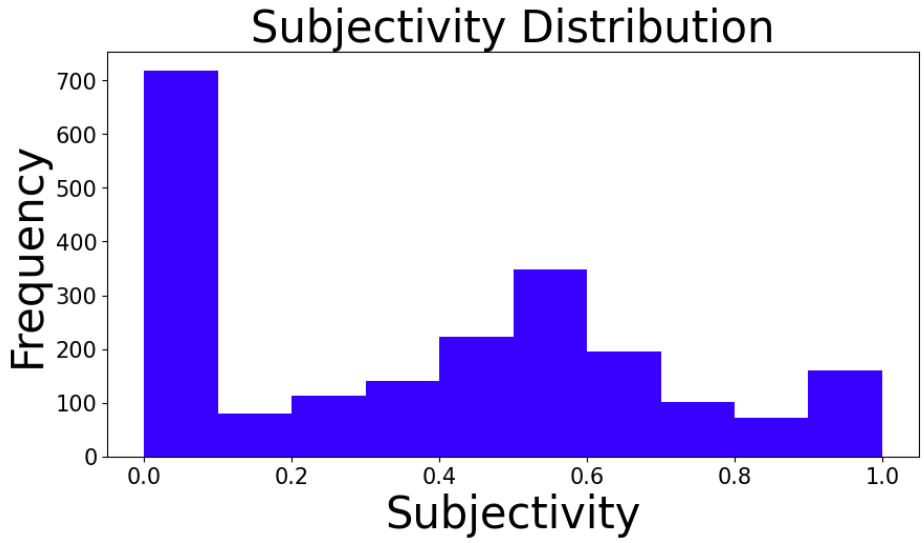


Fig. 7 Bar chart of the 15 most common words

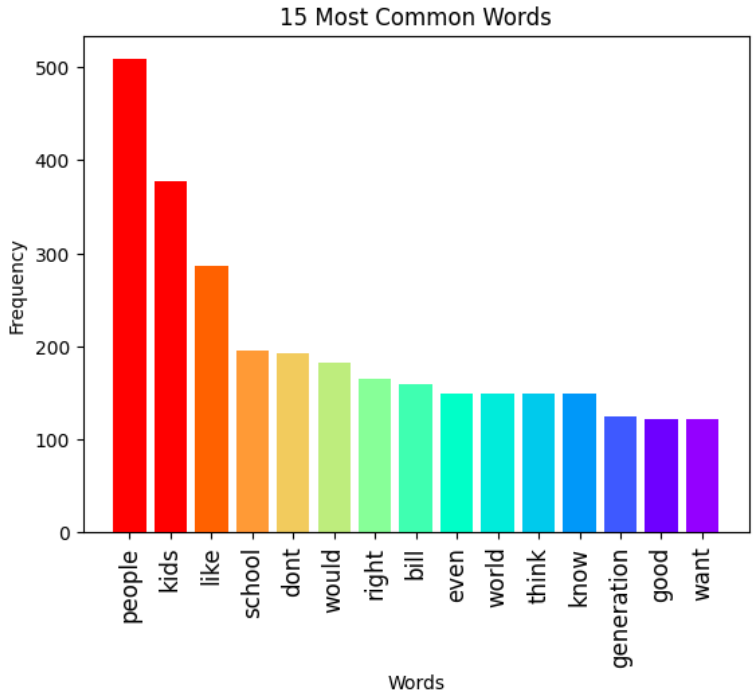


Fig. 8 Frequency tables displaying subjectivity and sentiment classifications of top 50 scoring comments

```
Polarity Frequency of Top 50 Words:
Positive Sentiment    24
Neutral Sentiment     15
Negative Sentiment    11

Polarity Frequency of Top 50 Words (%):
Positive Sentiment    48.0
Neutral Sentiment     30.0
Negative Sentiment    22.0

Subjectivity Frequency of Top 50 Words:
Subjective    26
Objective     24

Subjectivity Frequency of Top 50 Words(%):
Subjective    52.0
Objective     48.0
```

Fig. 9 Scatterplot comparing subjectivity and polarity scores

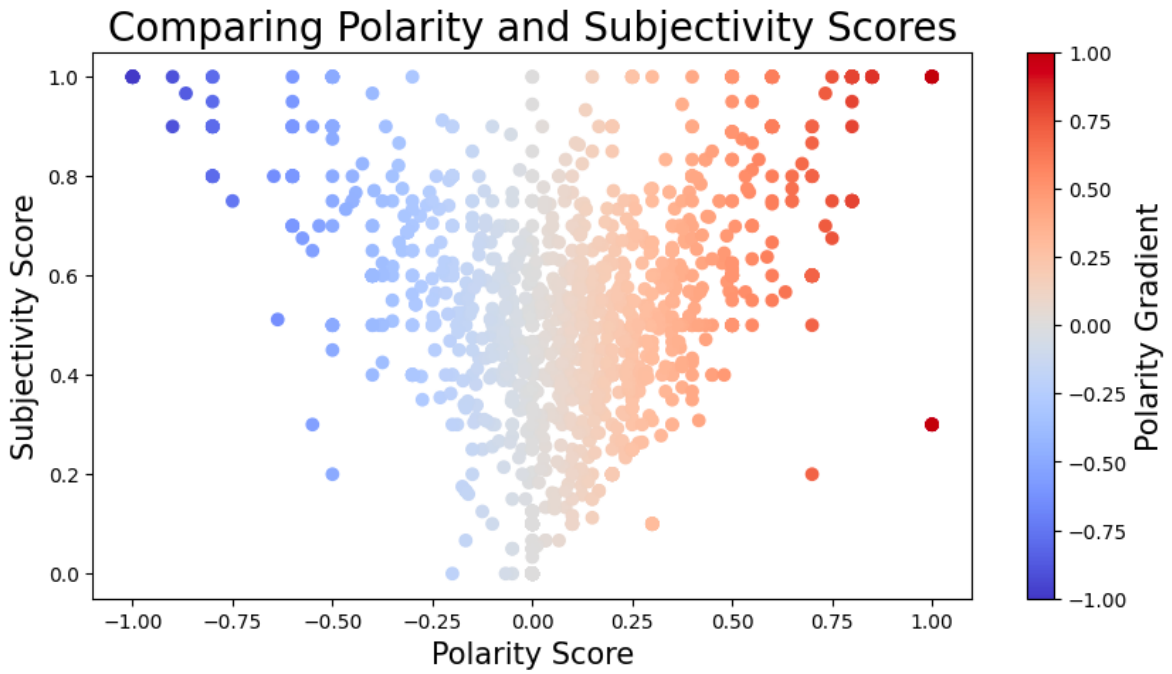


Fig. 10 Scatterplot comparing subjectivity and comment score

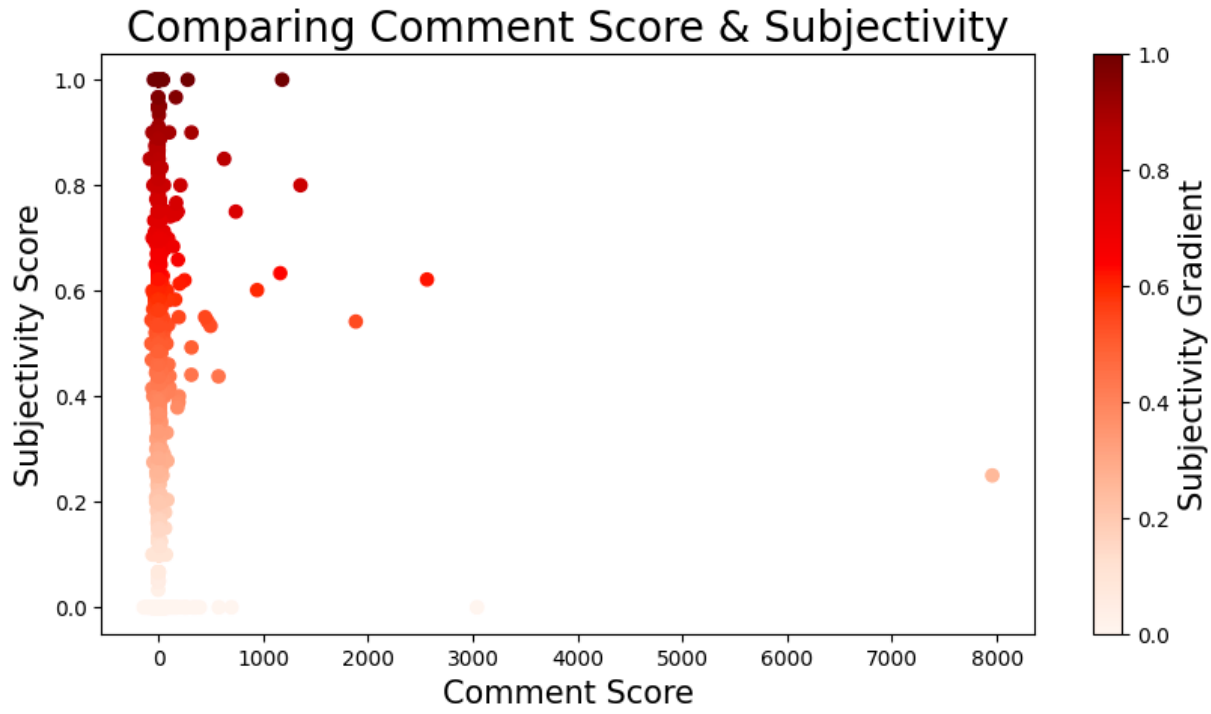


Fig. 11 Scatterplot comparing polarity and comment score

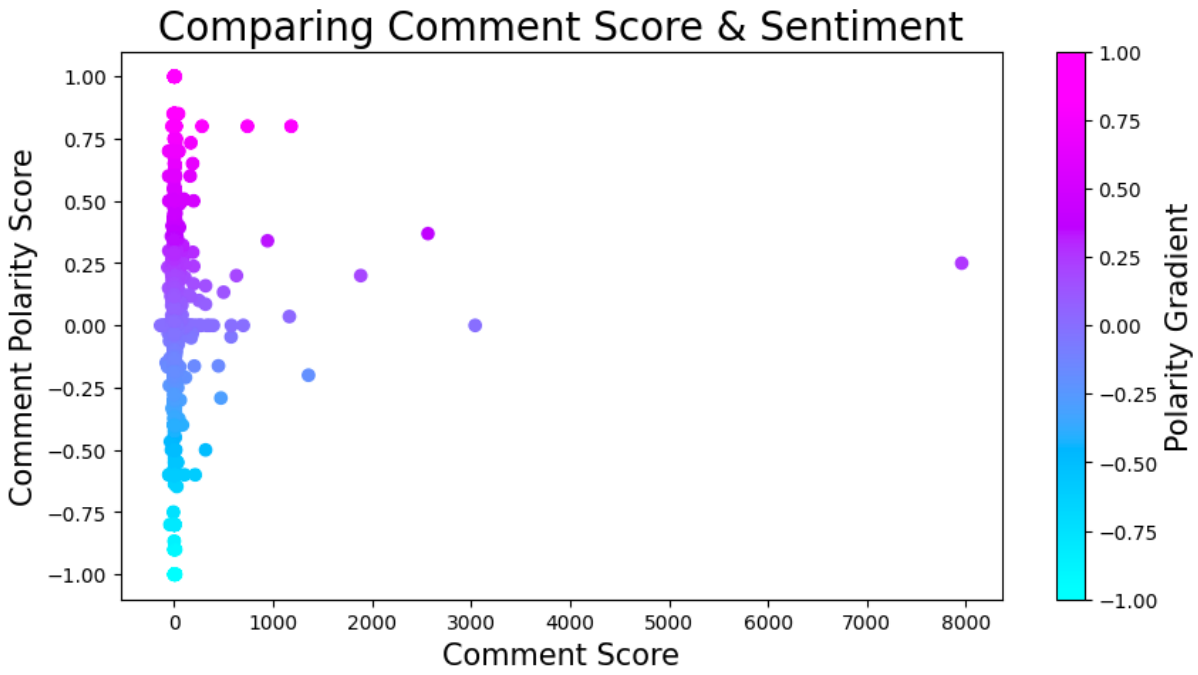
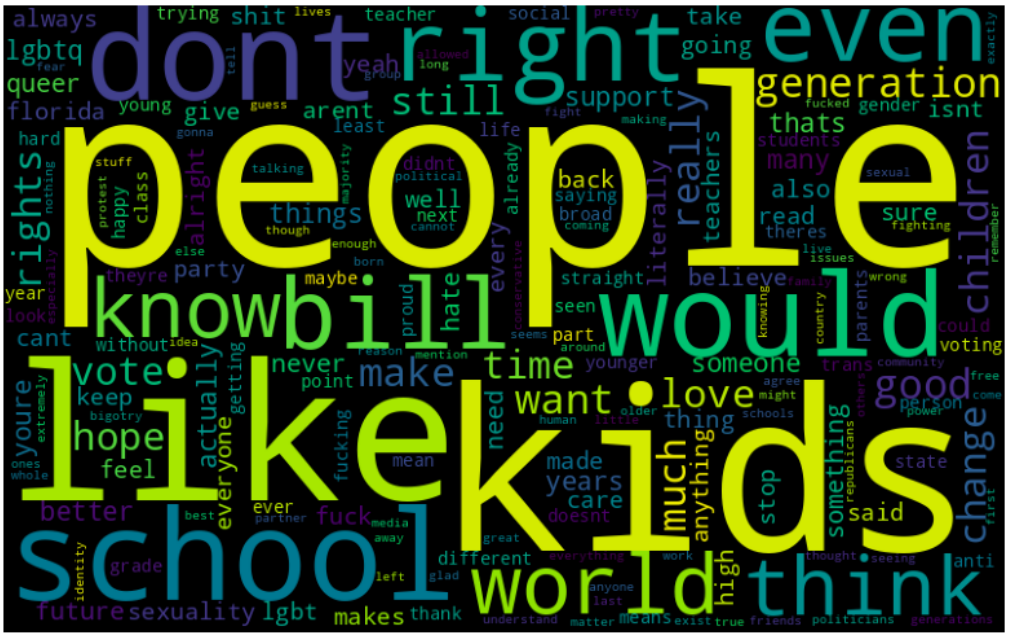


Fig. 12 WordCloud



References

Diaz, J. (2022, March 28). Florida's governor signs controversial law opponents dubbed 'don't say gay'. *NPR*. Retrieved April 4, 2023, from https://www.npr.org/2022/03/28/1089221657/dont-say-gay-florida-desantis

Lavietes, M. (2022). As Florida's 'don't say gay' law takes effect, schools roll out LGBTQ restrictions. *NBC News*. Retrieved April 4, 2023, from https://www.nbcnews.com/nbc-out/out-news/floridas-dont-say-gay-law-takes-effect-schools-roll-lgbtq-restrictions-rcna36143