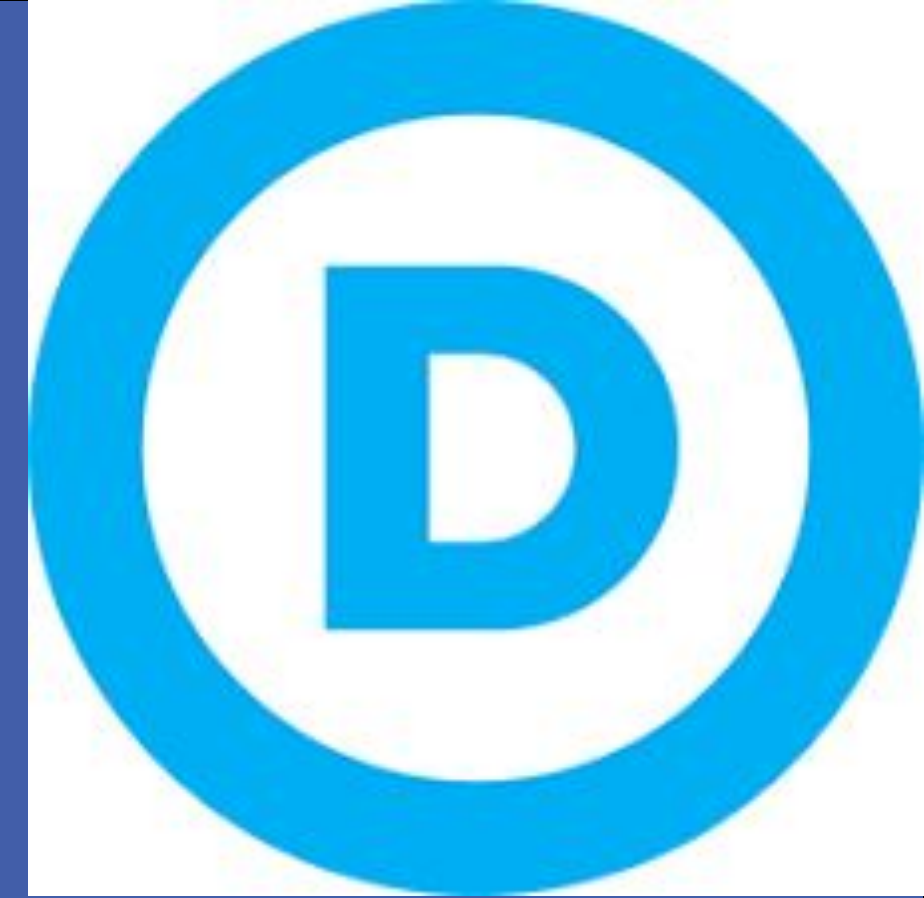# A Study of Language on Reddit in /r/democrats and /r/republicans

Completed By: Quinn He

University of Massachusetts at Amherst

Introduction

Social media offers users a platform for political discourse and discussion with people from all over the world or country. Facebook, Reddit, and Twitter are the primary social media platforms that all for the most text based posts. I chose to study Reddit over other text based social media apps because Reddit enables long form discourse on specific topics within a subreddit's post.

I want to study political discussion on this past year's political events and observe the different communication style of user's political discourse.

Research Questions

1. Do different political parties communicate the same way about similar or different topics?
2. What topics are discussed by subreddits?
3. How do subreddits use different words
4. Do subreddits tend to use more negative language than positive language

Data

- The two data sets I am using were collected with RedditExtractoR
- /r/democrats data set spans one year 11/08/2021 to 11/08/2022 with 26,764 comments
- /r/republican is from 11/10/2021 to 11/06/2022 with 2,054 comments
- Large discrepancy between the data collected because /r/democrats is significantly larger than /r/republicans with almost 200,000 compared to 17,600 respectively
- Variables in each dataset include url, author, date, timestamp, score, upvotes, downvotes, golds, comment, comment_id.

Methods

Preprocessing:

I removed comments with [deleted] or [removed] since that is a user's comment could have been deleted or removed by a moderator. Each subreddit was then converted into a corpus and tokenized, where I removed stopwords, punctuation, urls, numbers, symbols, and turned to lowercase. DFMs were used for further analysis.
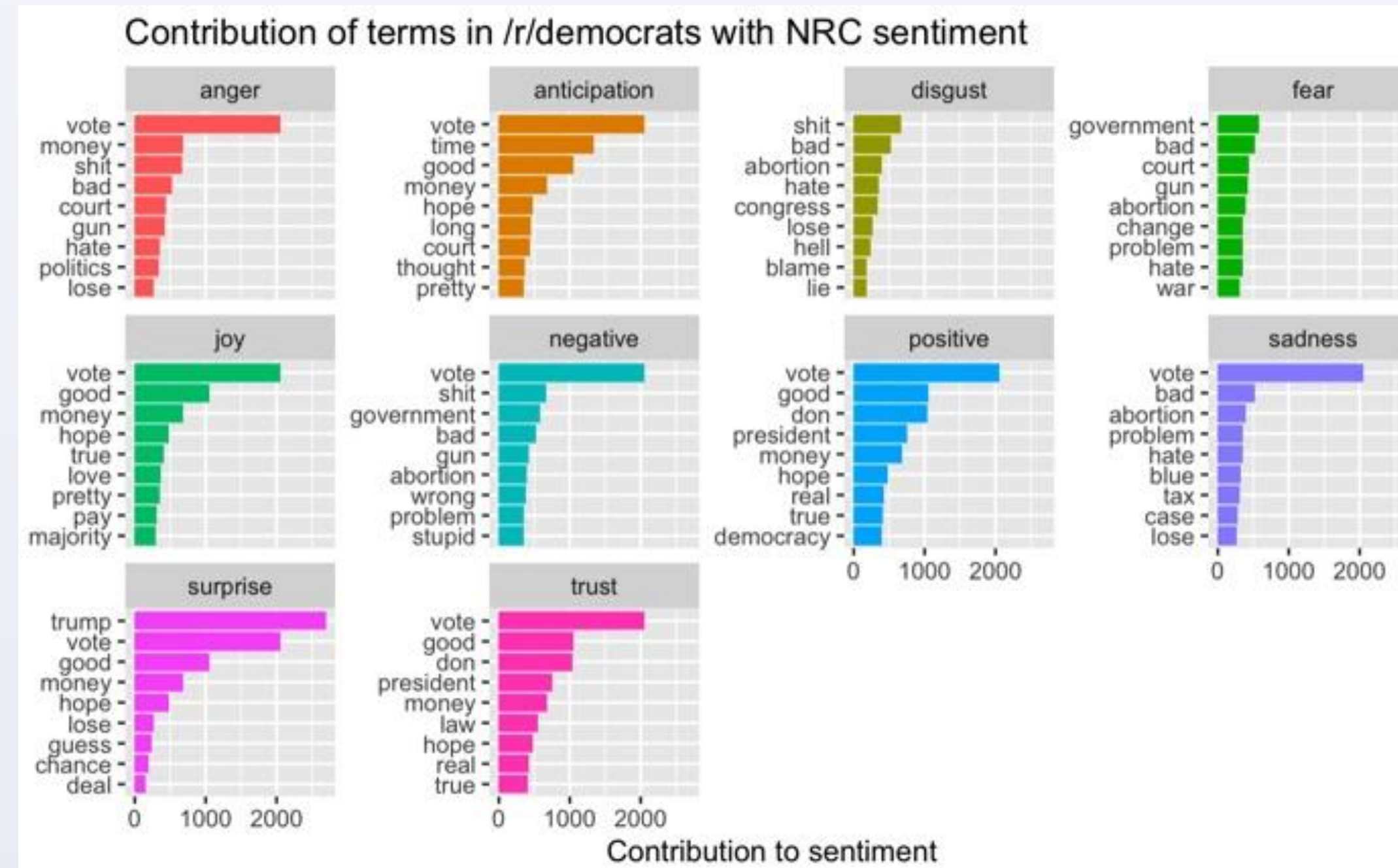
Analysis:

I used the NRC dictionary because it uses 8 basic emotions to measure sentiment. I used the positive and negative sections to create a polarity score by taking the different of the two.

Next, I used LDA topic modeling to examine what topics were most discussed amongst each subreddit and compared to see if there was overlap. I used a K of 5 because any more caused very jumbled topics with no clear link.
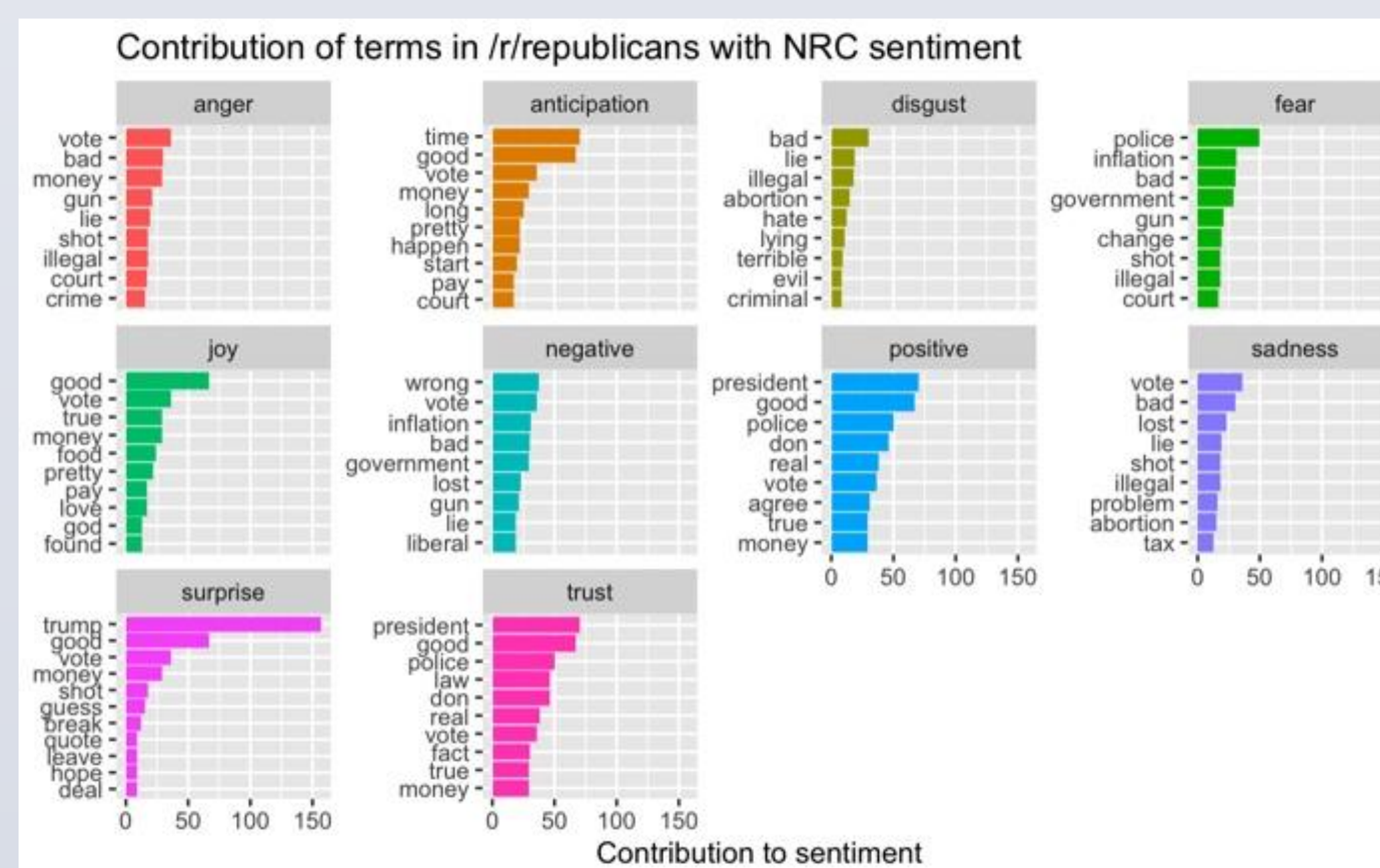
After preprocessing, Zipf's law was applied to the corpuses to determine word frequency. Top words in each subreddit were "people", "just", "trump", "vote" in /r/democrats and "people", "just", "trump", "biden" in /r/republicans. This was used to get an idea of what words are used the most by users.

Results

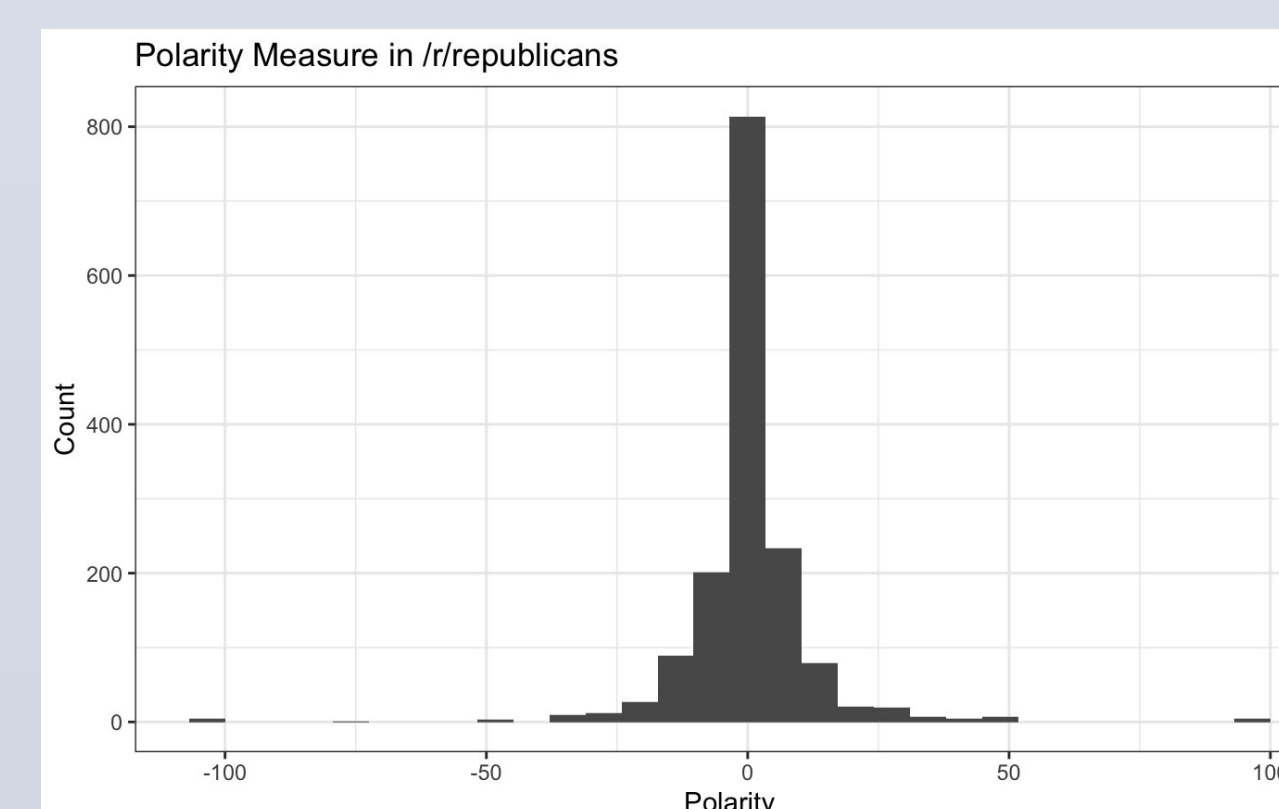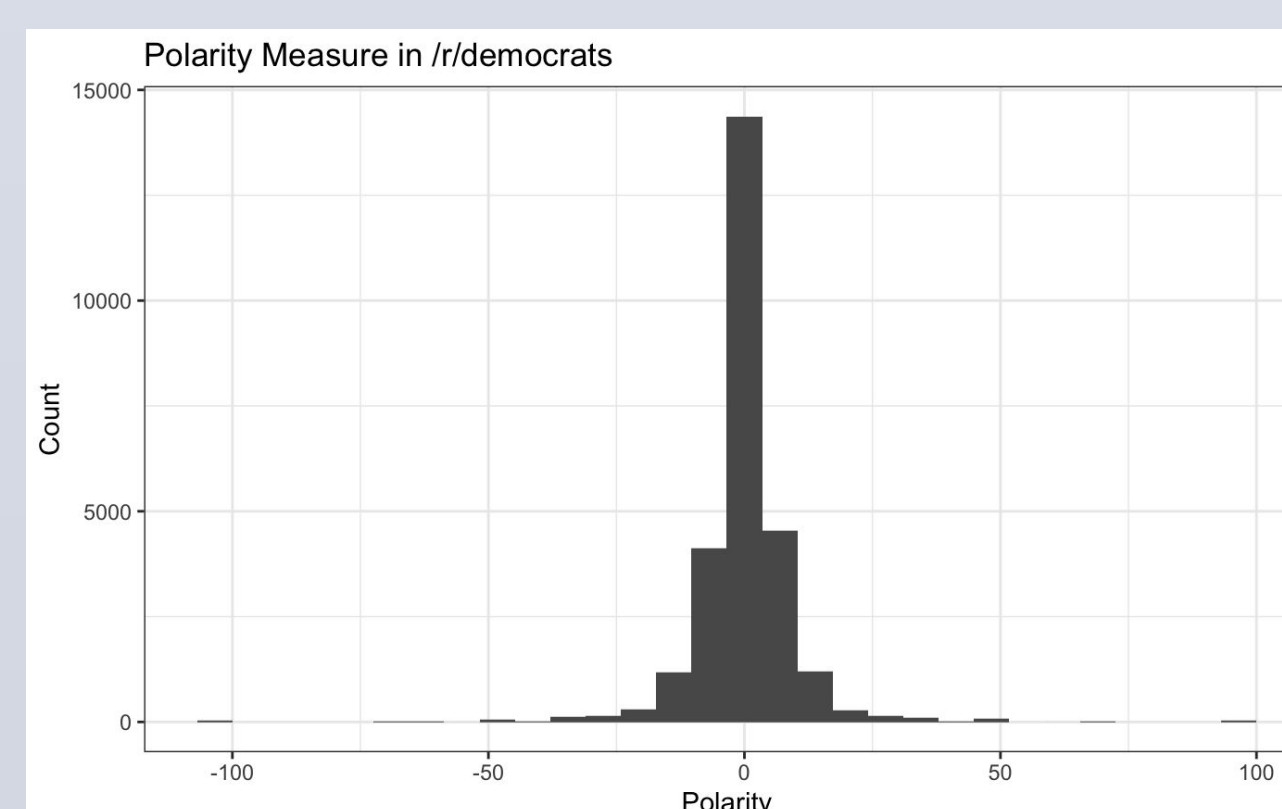Contribution of terms in /r/democrats with NRC sentiment

- Voting is of primary concern for democrats because people are using it under many different kinds of sentiment
- Focus on Fear, Positive, Negative, Anger
- Prevalent issues include: gun rights, abortion, and voting

- Both subreddits share similar discussion points, but different sentiments
- A few words find themselves in shared topics
- I can observe discussion of topics in sentiment analysis
- "Abortion" is found in both subreddits, but primarily used in /r/democrats where it's used in the context of Sadness, Negative, Disgust, and Fear. While liberals are typically pro-choice, the sentiment of the word is very negative. The word's used in /r/republicans makes sense as republicans are typically pro-life and have ill views of abortion
- "Gun" appears in both corpuses under Anger, Fear, and Negative.
- "Police" appears in Trust on /r/republicans, which is expected. Trust in police tends to have more right-wing support

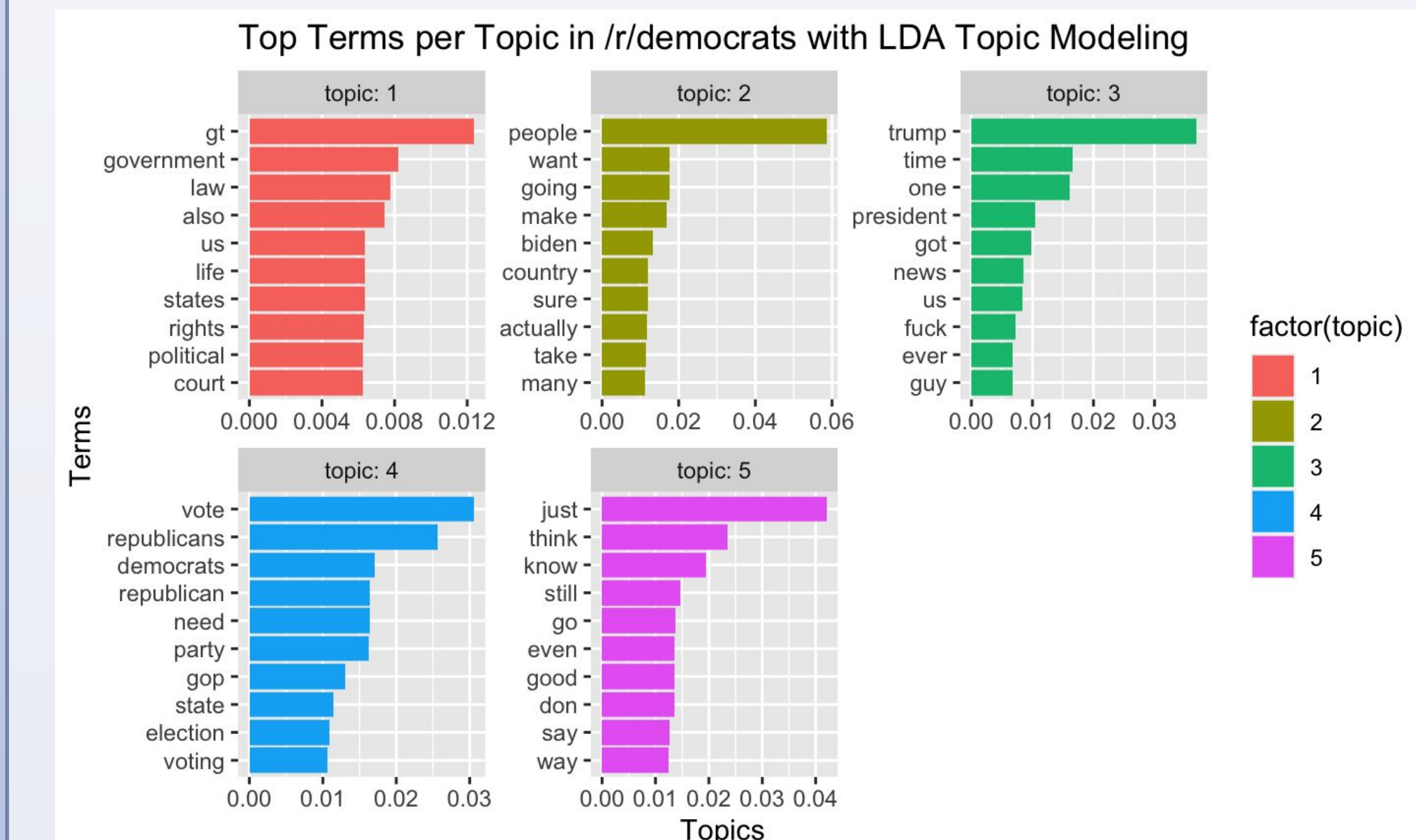Contribution of terms in /r/republicans with NRC sentiment

- Republicans speak negatively of inflation and liberals. Inflation is not seen in the graph for democrats
- Trump is used most frequently in the surprise section but I do not know why
- Anticipation, Trust, Positive, and Surprise are the most prevalent sentiments

Polarity Measure

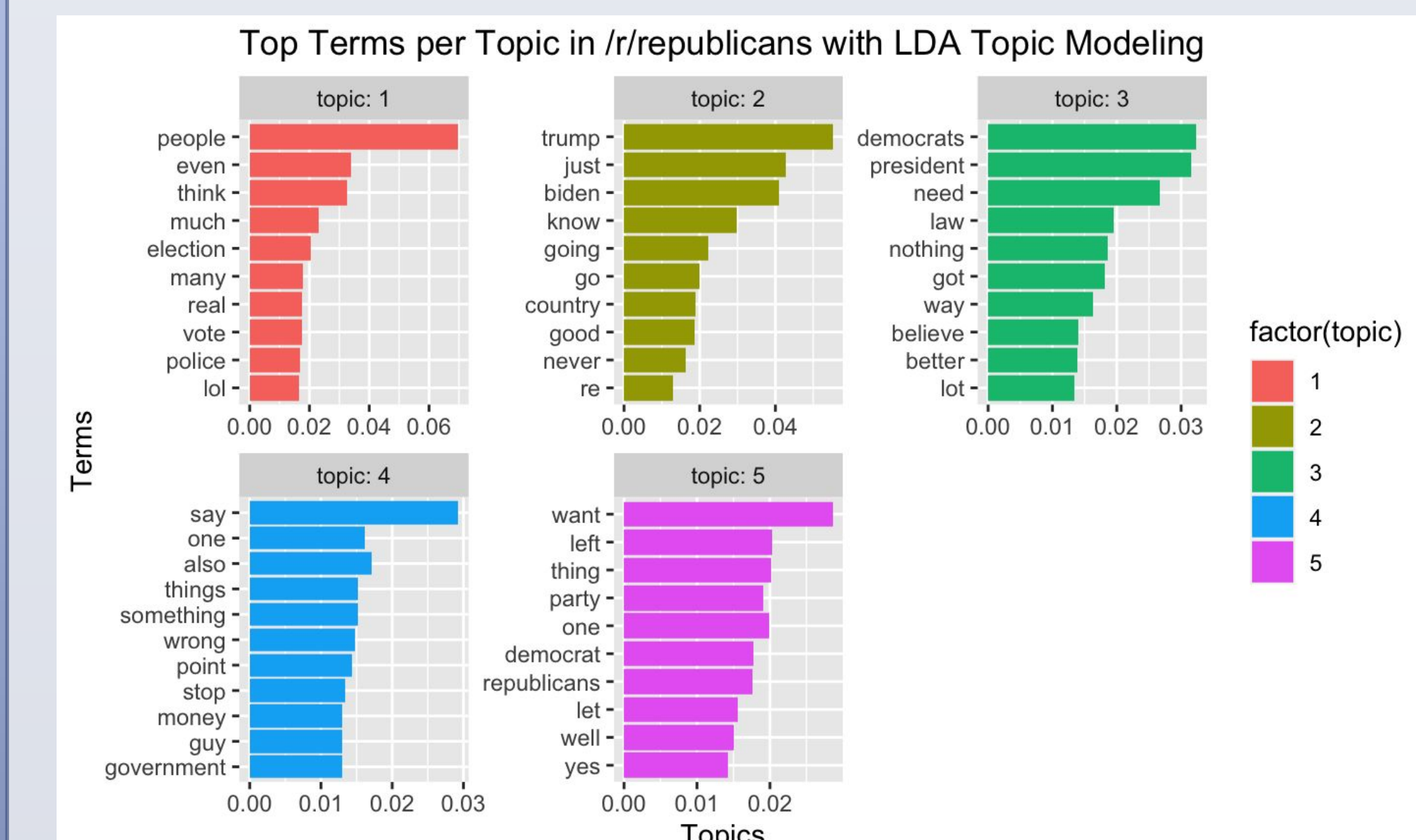Polarity Measure in /r/democrats

Polarity Measure in /r/republicans

The polarity score for each subreddit was 0.12 for /r/democrats and 0.27 for /r/republicans, under the NRC dictionary. I expected negative speech to be more prevalent in the subreddits, but their scores indicate a very neutral polarity score. Overall, comments tend to not have positive or negative speech, which leads me to believe discussion on these important issues is very civilized among users.

LDA Topic Modeling

Top Terms per Topic in /r/democrats with LDA Topic Modeling

LDA topic modeling was used to determine what subjects the subreddits discuss. Topic 4 in /r/democrats heavily features words related to voting. The word "need" in the topic stresses the importance. It's the most clear topic. Other topics are not as clear. I am not sure what gt means but it is the top word of Topic 1 which appears to discuss life and states rights.

- "Trump" is 1 of 2 shared top words in topic 3 of /r/democrats and topic 2 of /r/republicans
- Democrats talk of Trump very negatively
- When republicans talk about Trump, Biden is also typically referenced

Top Terms per Topic in /r/republicans with LDA Topic Modeling

- /r/republicans talks more about democrats than /r/democrats talk about republicans
- Some topics are difficult to decipher
- Low volume of comments makes topic modeling more difficult

Conclusion and Takeaways

Overall, both subreddits discuss some different issues, but there is a good bit of overlap as demonstrated by LDA and word frequencies. They share similar sentiment towards topics found in both subreddits. The subreddits are neutral in their discussion with very little negative speech overall. There was not much discussion about the midterm elections, except for the stress of voting among liberals.

The main limitation with this study is the fact I had such an unbalanced data set between the two subreddits. /r/democrats had an exceptionally larger set of comments compared to /r/republicans, which made it difficult to do a true 1 to 1 comparison. Reddit is seen as a more liberal minded platform, which may suggest the high activity in /r/democrats.

In a future study, I would look for more substantial subreddits right-wing politics. The inclusion of /r/politics or /r/worldpolitics may also help fill in any gaps in political representation. Also, it would be interesting to plot the prevalence of top words over time. I think other types of topic modeling would benefit this study with different values of K.

References

De Francisci Morales, G., Monti, C. & Starnini, M. No echo in the chambers of political interactions on Reddit. Sci Rep 11, 2818 (2021). https://doi.org/10.1038/s41598-021-81531-x.

Oliver, A. (2019, June 29). Determining presidential approval rating using Reddit sentiment analysis. Medium. Retrieved December 17, 2022, from https://towardsdatascience.com/determining-presidential-approval-rating-using-reddit-sentiment-analysis-79120b56cc7

Shen, Q., & Rosé, C. P. (2022). A Tale of Two Subreddits: Measuring the Impacts of Quarantines on Political Engagement on Reddit. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 932-943. https://doi.org/10.1609/icwsm.v16i1.19347