AN ANALYSIS OF POLITICAL SUBREDDITS

DAVID CHANIN | GRAY BITTKER | MOKSH MEHTA | NAVYA SAHAY

Background

This project examines the political landscape of Reddit on the subreddit level using adjacency data to map how connected different communities are. We use a variety of statistical methods and models to analyze the sentiments of different communities, clusters, and how political user content was based on the distances from a set of “seed” subreddits.

Hypothesis

We devised the following three hypotheses for our project:

1. Hypothesis A: There is a positive correlation between the distance from a political community and the negative sentiments of those communities.
2. Hypothesis B: Republican connected subreddits have a more positive sentiment than Democrat connected subreddits.
3. Hypothesis C: There is a negative correlation between the distance between the root political nodes and how politically biased they are.

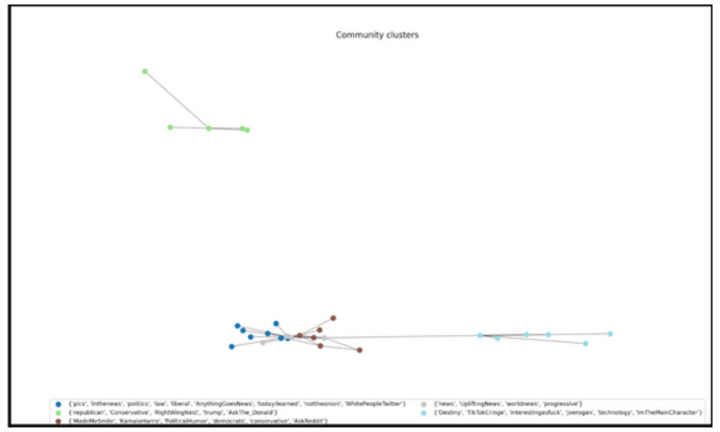
Data

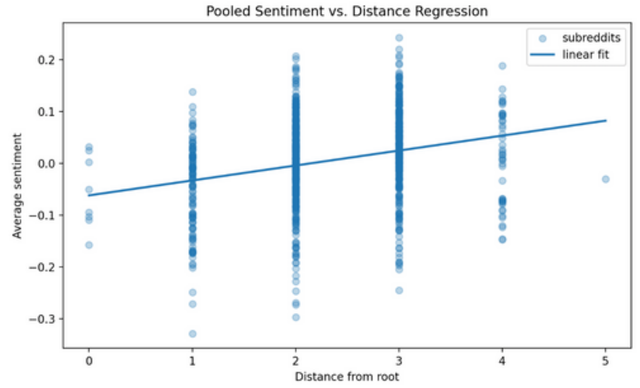
To analyze the relationships between our communities we used Reddit’s API to collect adjacency data on 904 different communities. Each community had approximately 50 adjacent ones based on where its top users also posted/commented, giving us nearly 45,000 adjacency vectors that demonstrated how strong the relationship between one community and another was. We also collected text data for the top posts we analyzed to engage in sentiment analysis on each community.Our scraping script collected data for a queue of subreddits, starting with our seeds, and expanded the queue if 5 or more top users all interacted with the same other community.

Methodology & Results

We utilized the Muddassar model to find the sentiment of the top 10 posts for the current year, month, week, and day of 223 subreddits; we considered the average sentiment of all these posts within a subreddit to be that subreddits approximate sentiment. We calculated the distance in terms of the number of nodes for each subreddit with respect to the seed nodes.To address our last hypothesis we used a transformer-based BERT model to return how left, center, and right each post was in different communities. We used the ratio of left to right to get a political bias score.

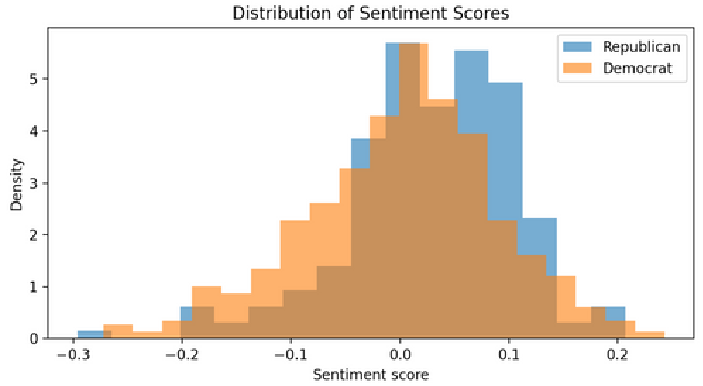
To inform our hypotheses and explorations, we wanted to explore the communities of subreddits that formed naturally based on the data of connections between subreddits based on number of common users and karma count.General subreddits based on news or other topics have proximity to both left-wing and right-wing subreddits. Specifically, extreme right-wing subreddits are mostly not connected to other communities or subreddits in the network.


Community clusters
{'pics', 'inthenews', 'politics', 'law', 'liberal', 'AnythingGoesNews', 'todayilearned', 'nottheonion', 'WhitePeopleTwitter'}
{'republican', 'Conservative', 'RightWingNest', 'trump', 'AskThe_Donald'}
{'MadeMeSmile', 'KamalaHarris', 'PoliticalHumor', 'democrats', 'conservative', 'AskReddit'}
{'news', 'UpliftingNews', 'worldnews', 'progressive'}
{'Destiny', 'TikTokCringe', 'interestingasfuck', 'joerogan', 'technology', 'ImTheMainCharacter'}



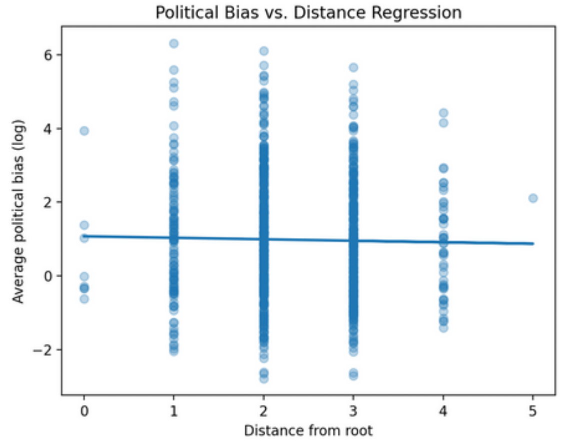
Hypothesis A: We fail to accept hypothesis one: instead, there is a statistically significant positive correlation between distance of a subreddit from the roots (['conservative', 'politics','republican', 'liberal','democrats', 'progressive', 'joerogan', 'trump'] and the positivity of the sentiment of the text within the subreddit (and not the negativity of the sentiment as hypothesized).

Hypothesis B: We used overlaid histograms and a Two-Sample T-Test to compare sentiment scores for Republican and Democrat subreddits. This method highlighted both the differences and overlaps between the two distributions. The Republican distribution had a higher mean and lower standard deviation, while the Democrat distribution showed more spread. This visualization supports the statistical significance of the difference in sentiment and allows for intuitive comparison.



Two-Sided T-Statistic:	3.2687
Two-Sided P-Value:	0.0013

Hypothesis C: To explore the hypothesis that subreddit political bias weakens with distance from ideological “root” subreddits, we visualized a regression between network distance and the log of political bias. There is a statistically significant negative correlation between distance of a subreddit from the roots (['conservative', 'politics','republican', 'liberal','democrats', 'progressive', 'joerogan', 'trump'] and the political bias of the text within the subreddit (becomes more right as the distance increases).



Conclusion

While we were able to validate our second hypothesis as well as our third hypothesis, we failed to accept hypothesis one. Rather, we instead found a positive correlation between the distance from a root political community and the positive sentiments of those communities.

Our results show that political subreddits differ in both sentiment and structure. Republican subreddits tend to have more positive sentiment than Democrat ones, and this difference is statistically significant. We also found that subreddits closer to political roots tend to be more biased, and that users cluster into clear communities based on political interest, with little overlap between opposing sides.

Challenges

Due to the main nature of the data being categorical (i.e. subreddits and not a numerical value) , many of the methods we learned in class were not directly applicable since they were focused on continuous data.

We also struggled with Reddit’s API call limit which significantly reduced the speed of our data collection script. To fix this we implemented time-based checks to ensure we had enough API calls for our requests. This allowed our script to run continuously without issues.

One of the main challenges in visualizing the results was choosing a clustering algorithm that would best capture and showcase the trends between subreddits when viewed. We initially considered k-means clustering, but realized that the data was already a network which is why we chose the Louvain algorithm.