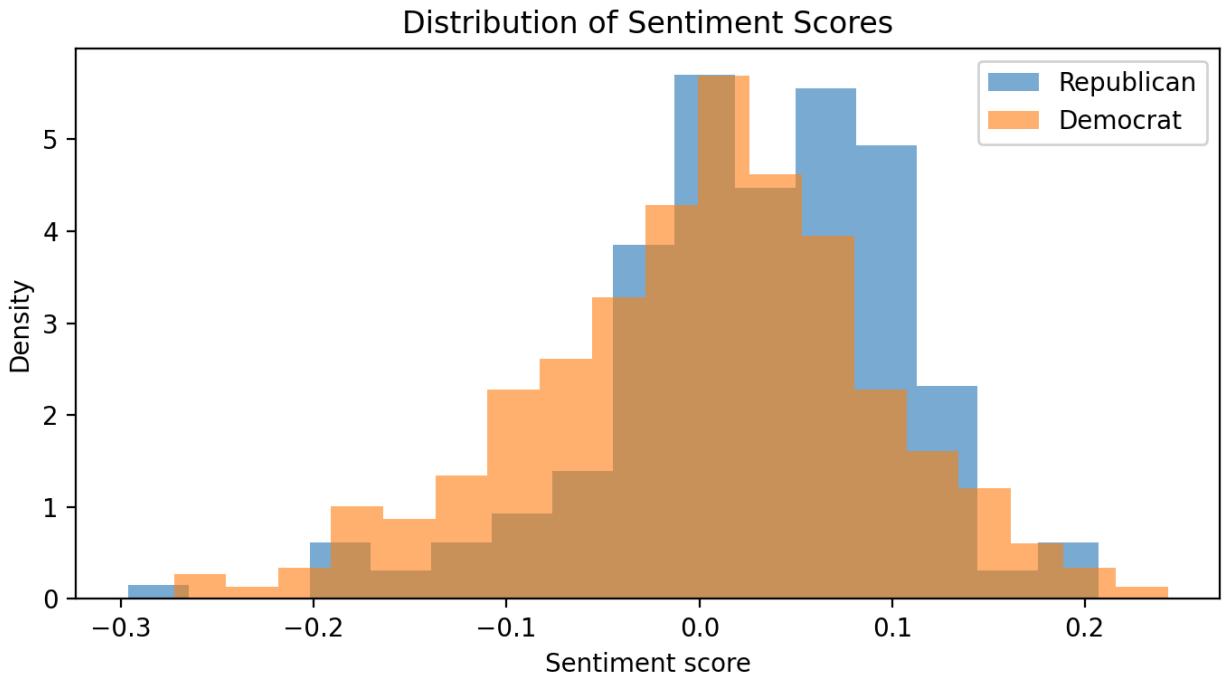
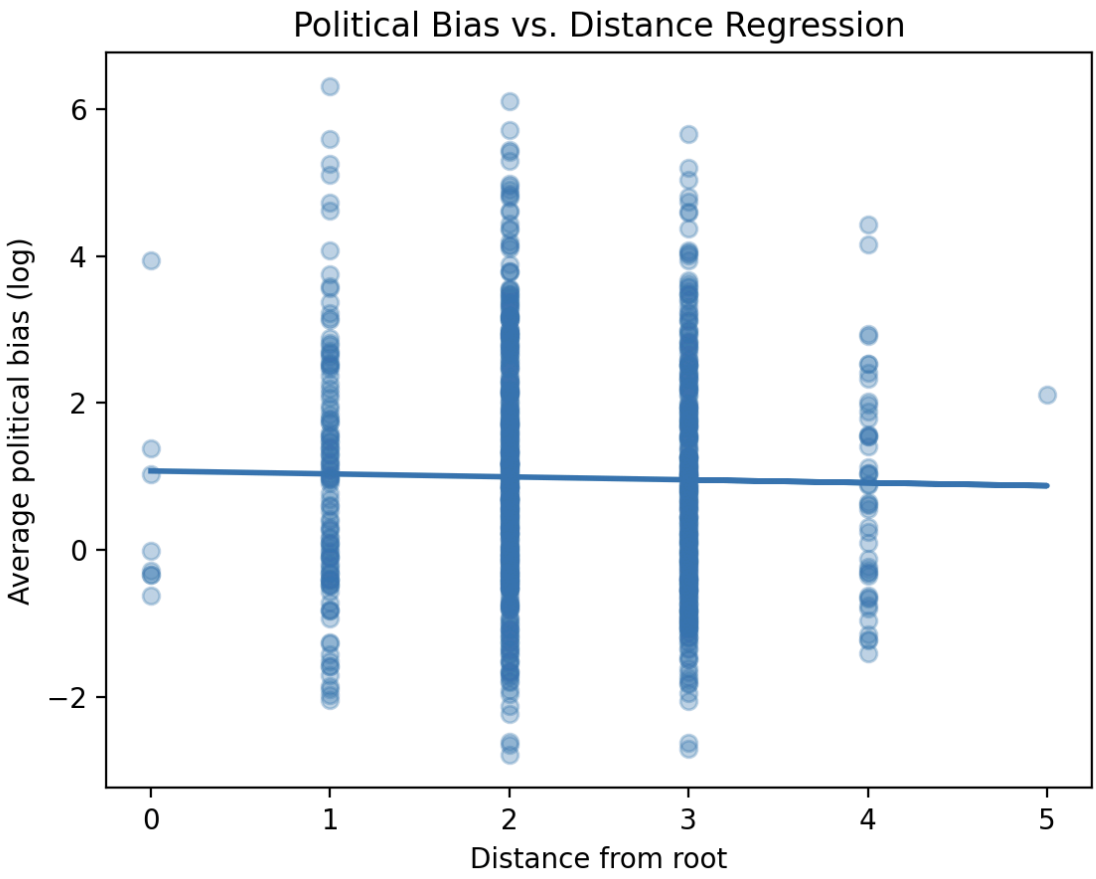


- Why did you pick this representation?
- What alternative ways might you communicate the result?
- Were there any challenges visualizing the results, if so, what were they?
- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

I choose to depict the Republican and Democrat sentiment scores as an overlay of two histograms because it makes it very easier to visually compare the overall shape, spreads, and medians of the distributions. For example, it is immediately apparent that the Republican mean is greater than the Democrat mean, and the Democrat standard deviations is greater than the Republican standard deviation. The apparent visualization of the differences also serves as a common sense validation of the statistically significant (p=0.0013) difference between the sentiment distributions. The semi-transparent bars also emphasize a lot of the similarities within the dataset. Viewers can see clearly that a majority of the sentiment scores for both groups lie within a similar region around 0.

Side by side histograms or box plots are also good alternative candidates to visualize the data. Side by side histograms would make it easier for the viewer to see and understand the complexities around each individual distribution, but it would be visually harder to directly compare them. A box plot would make it easier to understand the median and spread of the distribution, but the importance of the non-uniform densities would be lost to the diagram. Because of the strong tradeoffs, choosing the write visualization was a major challenge. We ended up picking the overlaid two histograms because we wanted to emphasize the importance of the differences between the distributions.

Lastly, this visualization is mostly self-contained. The clear axis labels and legend make the histograms pretty easy to understand without additional information. However, it is important to add a brief explanation of how the sentiment scores were generated. A description of our groups breadth first political subreddit technique and how the model classified sentiment on each subreddit is necessary to fully understand the dataset.



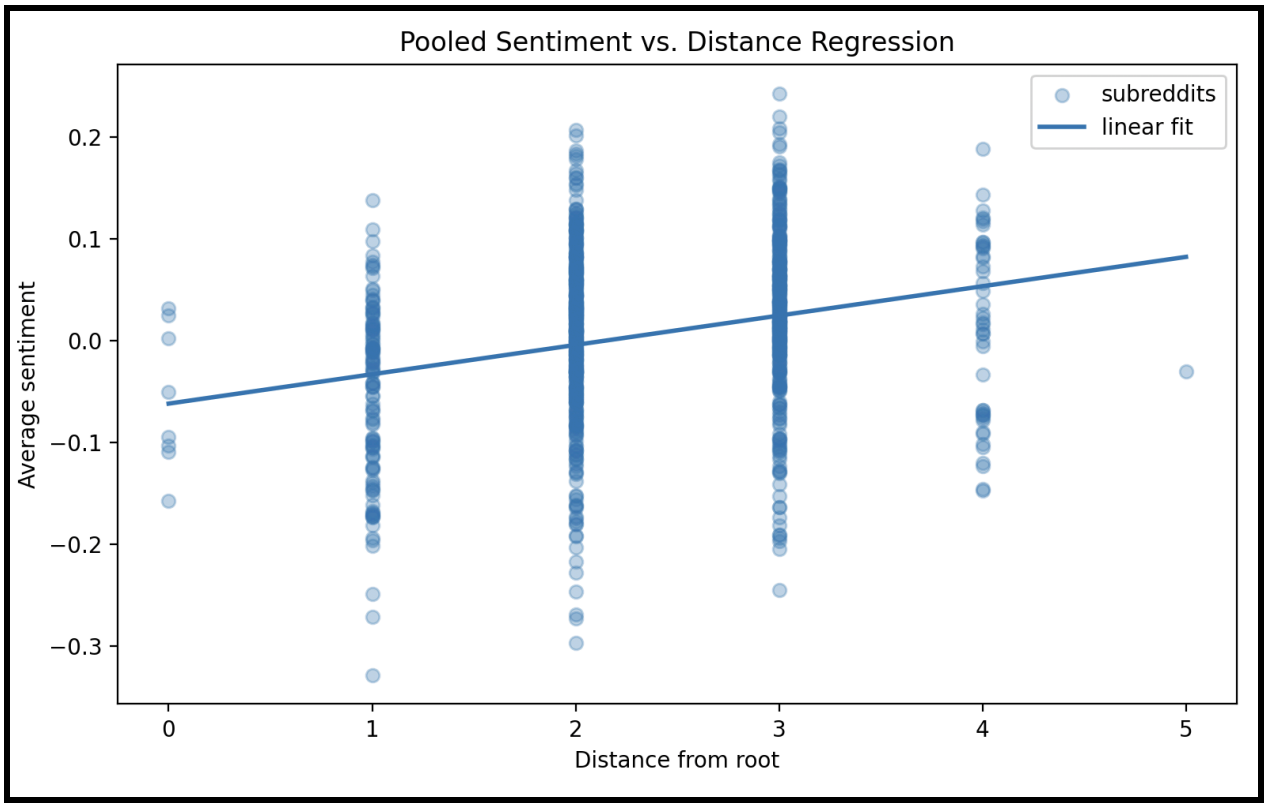
- Why did you pick this representation?
- What alternative ways might you communicate the result?
- Were there any challenges visualizing the results, if so, what were they?
- Will your visualization require text to provide context or is it standalone (either is fine, but it's recognized which type your visualization is)?

I picked this visualization because one of our hypotheses was "There is a negative correlation between the distance between two nodes and how politically biased they are". To confirm or deny this correlation (as seen in the graph there is a slight negative correlation visible) , a regression analysis was necessary such that we could discern the relationship between the two variables. An alternative way to communicate the result would be to have a basic scatter plot

without the regression line that just shows the points based on distance from the node and the subreddits' political bias. However, to make the correlation evident, we needed the regression line. Another alternative way we could have communicated this result would have been to have a heat map of the network of subreddits with different color gradients to indicate political bias. However, as each subreddit's distance was measured from different roots at a time (though subreddits could be neighbours of several roots), this may have led to several colors merging on the network and a clear trend would be hard to identify with the visualization.

The main challenge to visualize the results was to convert the data in "centers-to-text.csv" to a coherent pandas dataframe that could then directly be used to train and test a regression model. In the csv, the data was structured by the roots, such that each root had one row with a dictionary to capture all the different political biases of subreddits based on their distance from the root. However, to represent each point on a scatter plot and then do the regression analysis, it was necessary to loop through each row and convert the dictionaries to a list of dictionaries each with the keys: "root", "distance", "bias". Each of these small dictionaries represented one point on the graph and this could then be easily converted to a dataframe of rows needed for the regression analysis.

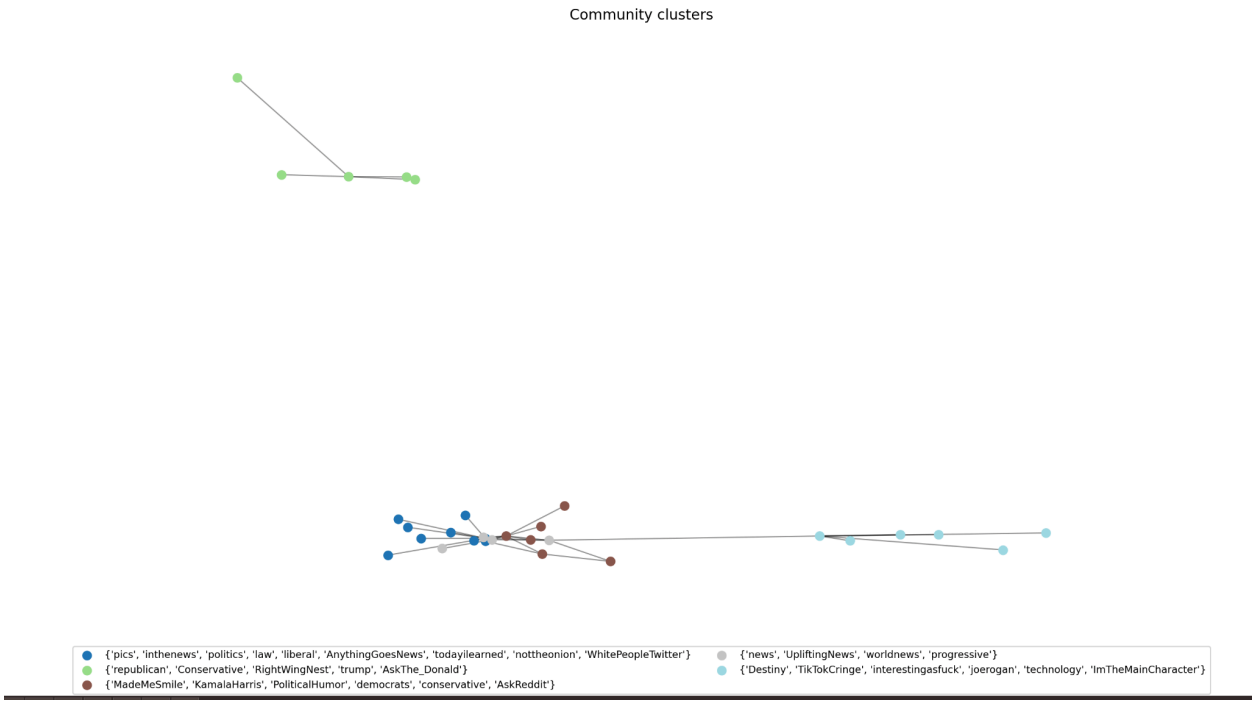
This visualization would require text to provide context as to how each of the variables were calculated as they are not variables directly available in the raw data.Additionally, distance itself could mean various things and an explanation that the distance refers to network distance is necessary. To transform the data such that we could create this visualization, we first used BFS to find the distance between our root/seed subreddits ('conservative', 'politics', 'republican','liberal', 'democrats','progressive', 'joerogan', 'trump') and all the other subreddits in our raw data. This distance was essentially calculated based on the number of nodes one would have to travel from the root to reach a particular subreddit. The political bias was calculated using a BERT-based model that took the textual data – the titles of pages within each subreddit and returned a political bias percentage of the form [left, center, right] for each textual example. We took the ratio of left to right percentages to calculate how polarized towards either end of the spectrum a subreddit was. We then took the log of this ratio just to normalize the scale for visualization purposes. Due to the multitude of calculations that were used to get the two variables, textual explanations like the one just provided must accompany the visual.



I picked this visualization because one of our hypotheses was “There is a positive correlation between the distance from a political community and the negative sentiments of those communities.”. To confirm or deny this correlation (as seen in the graph there is a positive correlation visible) , a regression analysis was necessary such that we could discern the relationship between the two variables. An alternative way to communicate the result would be to have a basic scatter plot without the regression line that just shows the points based on distance from the node and the subreddits’ sentiment. However, to make the correlation evident, we needed the regression line. Another alternative way we could have communicated this result would have been to have a heat map of the network of subreddits with different color gradients to indicate sentiment positivity level. However, as each subreddit’s distance was measured from different roots at a time (though subreddits could be neighbours of several roots), this may have led to several colors merging on the network and a clear trend would be hard to identify with the visualization.

The main challenge to visualize the results was to convert the data in “centers-to-text.csv” to a coherent pandas dataframe that could then directly be used to train and test a regression model. In the csv, the data was structured by the roots, such that each root had one row with a dictionary to capture all the different political biases of subreddits based on their distance from the root. However, to represent each point on a scatter plot and then do the regression analysis, it was necessary to loop through each row and convert the dictionaries to a list of lists containing distance and sentiment.

This visualization would require text to provide context as to how each of the variables were calculated as they are not variables directly available in the raw data.Additionally, distance itself could mean various things and an explanation that the distance refers to network distance is necessary. To transform the data such that we could create this visualization, we first used BFS to find the distance between our root/seed subreddits ('conservative', 'politics', 'republican','liberal', 'democrats','progressive', 'joerogan', 'trump') and all the other subreddits in our raw data. This distance was essentially calculated based on the number of nodes one would have to travel from the root to reach a particular subreddit. 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.



I picked this clustering visualization because 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. This visualization helped us formulate our hypotheses and explore the trends within the data as to the interests of users on a particular political subreddit. For instance, the network shows that users interested/active in the joe rogan subreddit are likely to be interested/active in the technology subreddit as well. The cluster/network visualization also shows that users active on the republican subreddit are also active on “Conservative”, “RightWingNest”, “trump” and “AskThe_Donald” subreddits. This community of subreddits don’t seem to be connected to the other communities at all suggesting

that there aren't many users in common with these, especially right-wing subreddits and the other subreddits recorded.

An alternate way we could have communicated the results could be a series of area charts with each area chart corresponding to one of the seed (E.x: "liberal") and the top other subreddits that its users are also active on. The area chart would capture the level/percentage of activity for each of these top subreddits.

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 realised that the data was already a network as we knew the top connections for each subreddit based on user_count and karma_count. We then chose the Louvain clustering algorithm which is a greedy algorithm that optimizes for modularity – the density of edges connections within a community versus the edges outside the community. We felt that this algorithm would best capture the relationships in the already-existing network and help show the communities that naturally form. This did prove to be the case as the figure highlights how the communities naturally form around different political ideologies or focuses in some cases. Another challenge when making this visualization was the placement of the legend which required a lot of experimentation with the loc and bbox_to_anchor features in plt.legend() function to ensure that the legend did not overlap the plot. This was mainly achieved by running the code several times with different positions for the legend till we were successful in achieving a clear, distinct legend and plot.

The visualization would require text to provide context as the distance between nodes within a community as represented by the edges needs to be explained. We added weights to each edge connecting two subreddits. These weights were calculated by dividing the common karma count between the two subreddits by the number of common users between the two subreddits. We felt this edge weight helped show how strong a connection between two subreddits was as it highlighted the level of activity common users had on both subreddits. Consequently, the higher the activity, the closer the subreddits were on the graph. This explanation is necessary to understand the spatial aspects of the plot.