HYPOTHESIS:

This project sets out to analyze the relationships between political and non-political communities on Reddit using a variety of different data analysis techniques. We have three primary hypotheses we tested using statistical methods such as a two-sample T-test and linear regression. Our research hypotheses are as follows:

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 two nodes and how politically biased they are.

DATA:

To answer these hypotheses we used an automated script that interacted with Reddit's API through the Python Reddit API Wrapper (PRAW) to collect a large sample of user-generated content from different communities. Our strategy revolved around using a careful selection of "seed communities" that included some of the most political liberal and conservative subreddits. For each of these communities our algorithm first obtained a representative sample of the community by collecting the names of the authors of the top posts over the past year, month, week, and day. If over 5 of these top authors also had top or recent posts/comments in another community that subreddit was then added to a queue for analysis. By crawling in this automated manner our data grew organically without researcher bias and gave us a large sample.

FINDINGS:

To address our first two hypotheses 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. We used a linear regression model to determine if there was a correlation between distance from the root nodes and the sentiment score. The mean squared error we got for the regression model was 0.0072, which was very low indicating a high accuracy. We got a p_value of 0.0 <= 0.05. Hence, we can reject the null hypothesis that there is no correlation between the two variables as the low p_value indicates statistical significance for the correlation. As we got a correlation slope of 0.0288, it suggests a positive correlation between the distance from the root and positivity of the sentiment. As a result, we must reject our first hypothesis as subreddits further from the root are actually more positive in sentiment. However, the R^2 value was 0.0054 which is very low and suggests that this correlation has little variance and is not necessarily the best predictive model for the relationship between the two variables.

Correlation slope	0.0288
p_value	0
R^2 value	0.0554

Next, we ran a Two-Sample T-Test for our second hypothesis on the Republican connected and the Democrat connected subreddits to determine whether there was a difference between their sentiment distributions. We found a very statistically significant difference (p=0.0065), suggesting that Republican mean sentiment is truly greater than Democrat mean sentiment. We can thus reject the null hypothesis for B, there is no difference between the mean political sentiments of Republican connected and Democrat connected subreddits.

	Republican	**Democrat**
Sampling Size	199	104
Mean	0.0228	-0.0096
Standard Deviation	0.0787	0.0836

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

To address our last hypothesis we used a transformer-based BERT model to determine how left, center, and right each post was in the different communities. We used the ratio of left to right to get a political score. The lesser the score was the more right the text was and the higher the score/ratio was the more left the text was. We used a linear regression model to determine if there was a correlation between distance from the root nodes and the sentiment score. The mean squared error we got was 3926.62 which given the scale of the political bias ratio (with the maximum political bias being approximately 550 indicating extreme bias towards left) was relatively low, indicating a reasonably high accuracy. We got a p_value of 0.034 <= 0.05. Hence, we can reject the null hypothesis that there is no correlation between the two variables as the low p_value indicates statistical significance for the correlation. As we got a correlation slope of -1.6686, it suggests a negative correlation. However, the R^2 value was 0.0019 which is very low and suggests that this correlation has little variance and is not necessarily the best predictive model for the relationship between the two variables. On the poster we utilize a log-transformed version of the y-axis (the political sentiment) in order to better control political outliers and provide a more understandable visualization. Hence, we will continue our analysis utilizing the statistics of the original regression with the below values.

Correlation slope	-1.6686
p_value	0.034
R^2 value	0.0019