Our project analyzes the online political landscape at a pivotal time in America. Numerous studies have shown that political polarization is increasing amongst American voters, and both Democrats and Republicans are more likely than ever to view the other party as an "existential threat" to America.[1] It is well documented that internet communities have played a critical role in this polarization by creating online "echo chambers" where opinions go unquestioned, and many companies are reluctant to combat these issues as politically involved users drive engagement numbers up.[2] Our data was collected with this background in mind, and we deliberately "seeded" our crawling script with some of the most polarized communities on social media app Reddit to see if there are statistically significant divisions in the online political ecosystem. This research has implications for any user of Reddit, particularly the 48% of users who are based in the United States and especially those who use Reddit as their primary source of news.[3] With nearly 8 million users subscribed to the r/politics community alone, it is clear that Reddit is a major source of political discourse in America, and we hope our findings help these users better understand the online political environment they are contributing to.[4]

Numerous other academics have already done research on Reddit communities, and we relied on this previous work to guide our design process. One such study comes from the University of Michigan and uses a much larger dataset than our project, focusing on millions of comments and offering a valuable look at how moderators remove content.[5] Another smaller study looks at the political bias of the top posts on Reddit in a weeklong window, giving us some valuable context as to what Reddit's general political leanings are.[6] Finally, we found a source that uses Louvain clustering which is a model we use in our final conclusions, giving us a baseline to compare our results against.[7] All of these sources give us some previous results to compare ours too and help us understand our findings in the broader context of broader literature. In particular, the questions that they ask are ones we can consider in our final report, and their model presentation can also be replicated as it is tried and true.

Given our data is user generated it will contain any of the underlying biases of the users that made it, and with how broad Reddit is that will include a wide range of potential biases. While the biases in the user-generated content itself aren't of particular concern for our project, we did consider how to address the more systemic biases that are inherent to the community

[1] https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/
[2] https://www.brookings.edu/articles/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/
[3] https://explodingtopics.com/blog/reddit-users
[4] https://www.reddit.com/r/politics/
[5] https://michiganross.umich.edu/news/new-study-reddit-explores-how-political-bias-content-moderation-feeds-echo-chambers
[6] https://rsisinternational.org/journals/ijriss/articles/reddits-american-political-left-wing-bias-a-study-of-the-top-100-posts-from-september-12-21-2024/
[7] https://snap.stanford.edu/class/cs224w-2018/reports/CS224W-2018-85.pdf

upvote system, something that is especially important given our research revealed that Reddit is biased against conservatives. To mitigate this bias we used an equal selection of liberal and conservative Reddit communities for our seed subreddits, though due to the nature of our data collection script and crawling methods it is likely our data as a whole is biased against conservatives. When we interpreted our data we kept this in mind and ultimately made it a major part of our conclusion, doing sentiment analysis on the results to see which seed communities might be the most biased. In the grand scheme of our project this bias is a feature rather than a bug, and we tried to examine it in detail with our analysis methods.

While bias wasn't a major concern for us we were quite worried about the privacy implications of having an automated crawler collect user generated content, particularly because people often post personal information on Reddit and could be reasonably identified by their data. To alleviate these concerns we did not include any fields related to usernames in our final data deliverable, making it impossible for anyone to determine who made a post without manually looking up the content on Reddit. While we didn't obtain explicit consent from users to use their data we did go through Reddit's official API which all users consent to when they use Reddit, and this assuaged many of our concerns about explicit participant consent. Given our data does include lots of post details it is entirely possible someone could use it maliciously, though we don't view this as a major risk given all data we scraped is publicly available and anyone could access it by simply using the Reddit app or website.

WORKS CITED:

https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

https://www.brookings.edu/articles/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/

https://explodingtopics.com/blog/reddit-users

https://www.reddit.com/r/politics/

https://michiganross.umich.edu/news/new-study-reddit-explores-how-political-bias-content-moderation-feeds-echo-chambers

https://rsisinternational.org/journals/ijriss/articles/reddits-american-political-left-wing-bias-a-study-of-the-top-100-posts-from-september-12-21-2024/

https://snap.stanford.edu/class/cs224w-2018/reports/CS224W-2018-85.pdf