INFO 440: Final Report

Title: Analyzing Political Opinion Dynamics on Reddit in Response to Geopolitical Events

Group 8: Claudia Adam (cca57), Andrew Rogers (ar3933), Thien Hoang (th3227), Jason Hicks (jrh446), Jeffrey Cheung (jc4759)

Introduction and Motivation

Social media platforms play a central role in shaping political discourse, offering a valuable lens through which to study public opinion, polarization, and the spread of misinformation. Reddit, in particular, provides real-time, unfiltered insights into how people react to evolving geopolitical events. Additionally, the narratives that form in these online communities can influence broader political attitudes and perceptions of legitimacy. Understanding these dynamics can help develop tools that foster healthier political dialogue and mitigate the harms of polarization and misinformation. Our project addresses this challenge by analyzing how politically aligned Reddit communities respond to major geopolitical events. We selected subreddits representing communities with strong, divergent ideological leanings, and tracked shifts in sentiment and discourse to study how polarized groups process the same information differently. Do their reactions converge toward shared narratives, or do they diverge further into echo chambers? To explore this question, we collected posts and comments from both subreddits and applied natural language processing techniques to capture not only the content of political conversations but also the emotional tone and temporal trends that define them. The results have practical implications, helping social media platforms anticipate when and where misinformation is likely to spread, while also providing policymakers with faster, data-driven insights into public opinion. Ultimately, understanding how online political communities react to geopolitical events helps navigate the information landscape of the digital age.

Methodology

To investigate how polarized communities on Reddit respond to geopolitical events, we used a methodology that combined data collection, text preprocessing, feature extraction, clustering, and sentiment analysis. Using PRAW, the Reddit API, we gathered posts and comments from politically aligned subreddits, specifically r/Republican and r/Democrats. The raw text was cleaned by removing stop words, punctuation, and other noise, then tokenized for analysis. Then, we applied TF-IDF vectorization, which allowed us to capture the importance of words relative to the broader corpus. Posts and comments were then grouped into topics through unsupervised clustering, enabling us to map discourse around recurring political themes. For sentiment analysis, we used TextBlob to assign polarity scores, which provided a measure of whether discussions leaned positive, negative, or neutral. Finally, we visualized the results to track how sentiment shifted over time and across subreddits. This pipeline was designed not only to capture the substance of political conversations but also to reveal temporal patterns and divergences in emotional tone. Overall, our project offers insights into how online political communities process unfolding geopolitical events.

Results

Our pipeline collected posts and comments from r/Republican and r/Democrats, cleaned the text, grouped posts into simple topics, and scored sentiment over time. Daily average sentiment in both communities stayed close to neutral most days, with short swings but no clear, lasting gap between subreddits.

We tried collecting posts using both Top and New listings. Neither approach solved coverage or balance. The main blocker was the ~1,000-post cap per listing imposed by the Reddit API. This cap limited how much we could pull per subreddit and time window, which in turn limited our ability to form stable, mixed-subreddit topics. Using TF-IDF with K-Means, we ended up with one very large “catch-all” topic and several tiny, often single-post topics. Because most topics did not contain posts from both subreddits, we could not make fair, side-by-side sentiment comparisons within the same issue. As a result, we were not able to determine any echo-chamber effects in this dataset. This reflects a data/collection limitation rather than a finding about the communities themselves.

Conclusion

Reddit remains a strong setting for studying fast-moving political discussions, but our study shows that data collection is the area needing the most work. With the current API limits, our topics were either too broad (everything in one bucket) or too narrow (single-post buckets). Without well-formed, mixed-subreddit topics, we cannot directly test whether the two communities converge or diverge on the same issues over time. Our results therefore do not claim convergence; they show that our current pipeline could not measure the question well enough to decide.

Future Directions

To make the echo-chamber test feasible, we recommend the following steps:

1. Data collection and framing (top priority)
    a. Acknowledge the ~1,000-post API cap as the central constraint.
    b. Pull on a recurring schedule (e.g., daily/weekly) and append to a local store to build our own historical database over time.
    c. Evaluate third-party data providers/data brokers for historical Reddit coverage to overcome API caps and fill gaps.
    d. Prefer time-bounded windows (recent months) rather than only “Top,” and merge titles + bodies; include high-quality comments to strengthen each topic.
2. Topic grouping

a. Allow short phrases (bigrams/trigrams) and basic lemmatization so topics reflect issues rather than single words.

b. Use a grouping method that doesn’t force everything into a few large clusters or many one-post clusters; set minimum size and require posts from both subreddits before a topic is used.

3. Sentiment and stance
   a. Try a social-media-oriented sentiment tool and, where feasible, add a simple stance label (for/against/neutral) so we can compare tone and position within the same topic and week.
4. Quality checks
   a. Spot-check a small sample of posts per topic; track easy metrics (topic size, subreddit balance) and drop topics that do not pass these checks.
5. Notebook clean-ups
   a. Apply the comment-cleaning function to the correct DataFrame; merge titles and bodies before vectorizing; keep date filters consistent across posts and comments.

With more complete and balanced data, either by scheduled collection or via trusted data providers, we expect to form clearer, shared topics. Then with this we can measure whether sentiment within those topics moves together or pulls apart over time. This will put us in a much better position to answer the echo-chamber question directly.