

Political Views on Reddit - 02805 Social graphs and interactions

link to explainer notebook: <https://github.com/OscarWulff/Political-Views-on-Reddit-02805-Social-graphs-and-interactions>

Oscar Wulff Mandrupsen, s216163^{a,1}, Gustav Risager Clausen, s214940^{a,2}, and Jakob Astrup Fischer, s214411^{a,3}

Online political discourse on Reddit often unfolds in environments where anonymity lowers social consequences, enabling more extreme expression and fostering polarized echo chambers. In this work, we analyze political discussion across Reddit political communities using a combination of network analysis and natural language processing. Our results reveal three large and well-defined communities that closely correspond to political orientations, with left- and right-leaning groups separated by a mixed, ideologically diverse cluster. Although users across these orientations engage with similar political issues, they do so within distinct, like-minded spaces.

Sentiment analysis further demonstrates that communities with greater ideological diversity exhibit more positive discourse, suggesting potential pathways for fostering healthier political interactions online. Finally, we find that politically active users express more positive sentiment when posting outside political subreddits, while political topics and sentiment frequently spill over into their activity on non-political forums.

Introduction. Political polarization has become one of the defining challenges of our time. As more political conversations move online, understanding how people interact across ideological boundaries matters for the health of our democracies. Social media platforms like Reddit give us a unique window into these dynamics because they preserve detailed records of how millions of people talk to each other across thousands of different communities.

Reddit is organized into user-created communities called subreddits, each focused on a specific topic. What makes Reddit especially interesting for studying political behavior is that users keep the same identity across all these spaces. This means we can see how someone who argues about politics in one subreddit behaves when they post about their hobbies somewhere else, whether it be a political or non-political forum. Unlike Twitter where political posts flood everyone's feeds, Reddit users actively choose when to engage with political content, making it easier to separate political engagement from other activities.

Echo chambers, where people with similar views reinforce each other's beliefs and rarely encounter opposing viewpoints, have become a common concern in online political discussions. While we know these chambers exist, we understand less about how they form and what makes them stronger or weaker.

In this study, we investigate how political communities connect to each other through shared users, measure not just whether echo chambers exist but how isolated they really are, their language and topics of discussion, and examine how users behave when they cross political boundaries.

Data & Preprocessing. We use the TL;DR Reddit dataset, which contains 3.8 million posts (over 18 GB) from before 2018 (1). After cleaning out deleted accounts, duplicates, and inconsistent text fields, we keep about 3.45 million posts. To identify political communities, we used Gemini 2.5 Flash to classify subreddits as left, right, centrist, or non-political based on subreddits names. Then we use Beautiful Soup (2) to scrape descriptions for the political subreddits we just classified. Afterwards we used a combination of Gemini and manual reviews to remove misclassified, non-relevant subreddits and only keep "left" and "right" based on these scraped descriptions. We then create two datasets. One which consists of only 37 clearly political subreddits and 15,475 posts, and another which consists of all posts in non-political subreddits from those users classified as being political. This forms the basis for all analyses in this study. We only selected subreddits whose primary topic is American politics or American users.

Significance

We analyze political behavior on Reddit and show that users form three clear communities: left, right, and a mixed group where mainstream political subreddits coexist. Although left and right groups rarely interact, they often discuss the same political topics, suggesting that polarization is driven by social separation rather than differences in issues. We also find that isolated left communities express more positive emotion, while right communities show more negative tone, and mixed communities show the most positive climate overall.

Politically active users become more positive outside political forums.

These findings highlight how online platforms can reinforce political division and show that improving cross-group interaction may be more effective than moderating content.

Author affiliations: ^aDTU

Contributions:

While each person had more responsibility for one section than others, we have all contributed to each section. We also did a lot of pair programming for the individual sections as well as had group discussions regarding how to use each individual method and tie it to our primary research question.

¹Oscars largest contributions lie in the Data processing, network statistics, sentiment distributions and Louvain communities.

²Gustavs largest contributions were with the work surrounding echo chamber scores, sentiment homogeneity as well as the work on network-based tf-idf vectors.

³Jakobs largest contributions were in the section on political users influence on non-political forums, including pre-processing of the new dataset, sentiment scores and finding topics for each political orientation.

AI Declaration:

We have used generative AI tools throughout the project. We have used ChatGPT and Claude in large part for coding and formatting of plots. It has also been used for guidance on structure of report sections, and how to efficiently deliver our points. The generative AI has been used throughout the whole project, especially in the coding/model-building stage, in which we have used the output for our notebook. Any code suggested by the tool was adapted, integrated, and tested by us. All text-based generated output was treated as suggestions. We reviewed and edited the outputs and verified all technical content and equations ourselves.

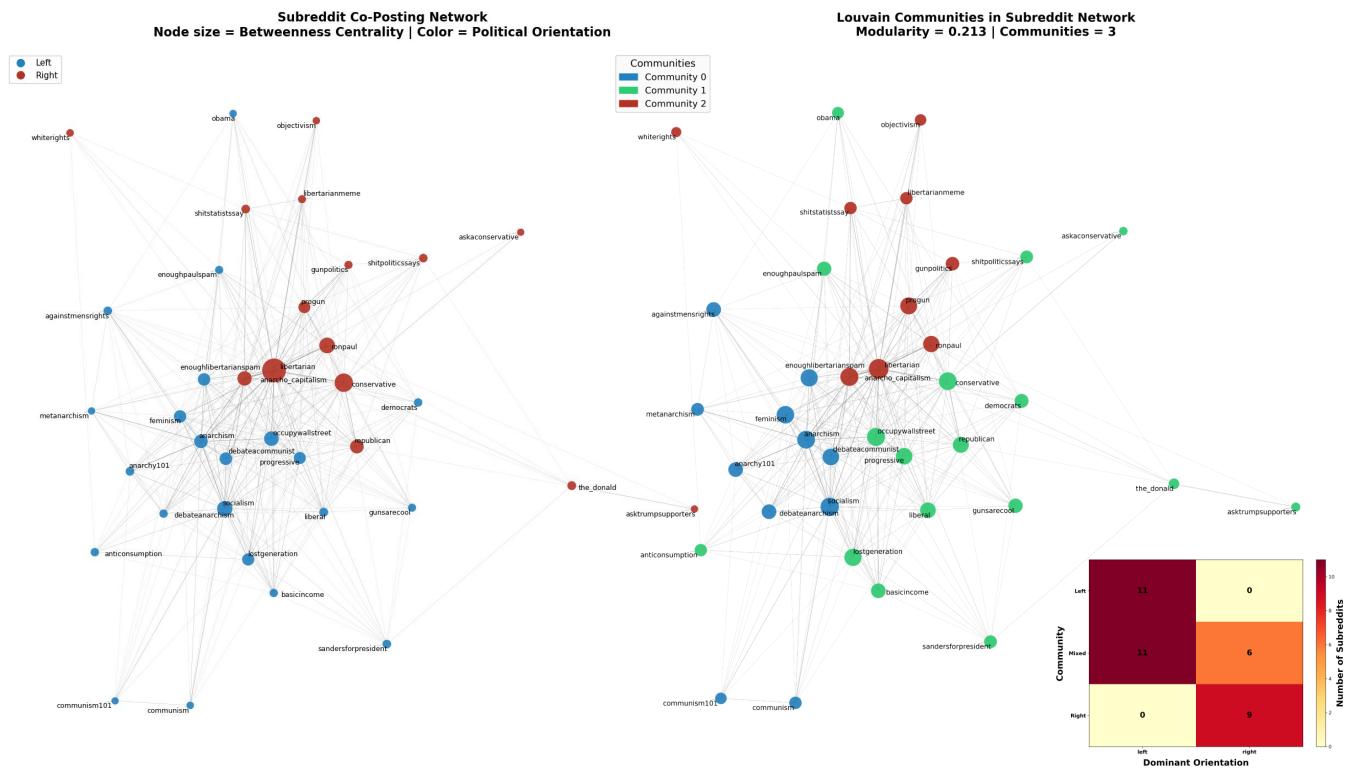


Fig. 1. Both networks show connections between political subreddits. The nodes are subreddits, and edges represent shared users who posted in both subreddits. Node size represents the betweenness centrality of each subreddit in the network. The distance between nodes illustrate if they have many shared co-posting users. and The left graph shows our classification of subreddits into left and right. The right graph shows the Louvain Community Algorithm results. The heatmap shows the distribution of left and right subreddits in the communities.

Results. The Louvain community algorithm identified three distinct communities, as shown in Figure 1. Community 0 contains only left-leaning subreddits, Community 2 contains only right-leaning subreddits, and Community 1 contains a mixture of left- and right-leaning communities. The algorithm achieves a modularity of 0.21. The partitioning of left & right graph seen in left side of Figure 1 has a modularity of 0.19, which shows the Louvain had a stronger partitioning than our labeling. The moderate level of community structure (modularity) shows political polarization is present, but cross-ideological interaction still occurs. The network has density of 0.452, meaning that 45% of connections exists, however, the graph only has 301 edges, meaning that a low amount of users actually post in different subreddits. The median connectivity is 16, and the degree Associativity is -0.2 meaning that highly connected subreddits link to many smaller ones rather than to each other.

This three community structure aligns with the core tensions of the US two-party system, where political identity often becomes an “either-or” choice. At the same time, the mixed community shows that many users still interact across political lines. Interestingly, community 1 contains “democrats”, “conservative” & “republican” subreddit, illustrating a more diverse community of traditional US politics, while community 0 and 2 contains more extreme left or right wing, where 0 contains “Against Mens Rights”, “Anarchy101”, “socialism”, “communism”, and 2 contains “whiterights”, “libertarian” and “progun”.

We can also see that nodes such as Libertarianism, Socialism and conservative have high betweenness centrality, meaning they share a lot of users with other

subreddits. These communities act as bridges between otherwise separated groups. This likely occurs because these subreddits represent broad political ideologies rather than specific political parties, which makes them natural spaces for debate and cross-ideological interaction.

We also introduce our own metric, the Echo Chamber Score, which measures how isolated a subreddit is based on user composition, sentiment homogeneity, and structural separation (see Methods). Figure 2 reveals a clear contrast between political sides. Right-wing subreddits exhibit the highest level of isolation, and they also show a lower median sentiment than left-wing communities.

In Plot B of Figure 2, we observe a clear reverse bell shape in sentiment distribution, with most posts being either very negative or very positive. This indicates that political discussions on Reddit rarely remain emotionally neutral and instead polarize toward extremes. We find that mixed communities show the most positive average sentiment at 0.50, with left at 0.27 and right communities the most negative at 0.23. The lowest sentiment in right-leaning communities aligns with previous findings on Donald Trump’s language use in political speeches (3), which show a more negative and confrontational tone that may have influenced broader right-leaning online political discourse. The mixed community with the highest sentiment, suggests that diverse political spaces create more balanced emotional tone overall.

Plot C in Figure 2 reveals an asymmetry in how sentiment relates to echo chamber intensity. We see that mixed communities display a strong positive correlation where more positive sentiment corresponds with stronger echo chambers. Left and right communities show con-

stant sentiment score. This reveals that the tone has the strongest role in mixed communities where diverse viewpoints interact, where more isolated subreddits inside the mixed community, display higher sentiment. However, overall the mixed community have higher sentiment as seen in [Figure 2](#) plot B when looking at all posts within the community.

In Plot D of [Figure 2](#), we plot user purity against sentiment homogeneity. We observe that most subreddits concentrate at high user purity values above 0.9, confirming ideological separation dominates Reddit's political landscape. However, we find that sentiment homogeneity varies widely. Subreddits like *sandersforpresident* show both high user purity and high sentiment homogeneity, but still fairly low echo chamber score since it is part of the mixed community. Whereas, *the_donald* scores very low in sentiment homogeneity illustrating high level of conflicted discussion.

Network based on TF-IDF vectors. If we change the edges to instead represent semantic similarity between subreddits, based on cosine similarity of TF-IDF vectorized text, we obtain a fundamentally different type of network. Rather than connecting communities through shared users, this semantic network connects communities through shared language and topics of discussion. In this network, an edge exists only if two subreddits use sufficiently similar language (cosine similarity ≥ 0.3), meaning that they discuss closely related subjects. See [Figure 3](#)

This allows us to separate who interacts with whom from what they talk about. While the structural network captures behavioral overlap between users, the semantic network captures topical overlap between communities, without implying that users actually move between these spaces. By applying Louvain community detection to this semantic graph, we identify communities of subreddits that group together purely based on similarity in language use.

[Figure 3](#) shows the resulting semantic network together with word clouds for each semantic community. Each community clearly reflects a dominant political topic. One community centers on gender and feminism, another on anarchism and anti-capitalism, a third on gun politics and shootings, a fourth on corporations and regulation, and a fifth on electoral politics centered around figures such as Bernie Sanders and Donald Trump. These results show that political discussion on Reddit naturally organizes around issues, rather than only around party identity. It only shows which topics have a high interest among the Reddit users.

The left-right distribution across these semantic communities, shown in the heatmap in [Figure 3](#), further supports this interpretation. Some semantic communities show a strong ideological skew, while others contain a more balanced mix of left- and right-leaning subreddits. Importantly, this does not mean that users from different sides interact directly with each other. Instead, it shows that both sides engage with the same political topics, but often in separate ideological spaces.

Political users in non-political contexts. As an extension of our analysis of sentiment and common topics across communities, we turn to a larger-scale semantic investigation. We would like to investigate common political user activity

Reddit Political Network: Structure and Community Dynamics

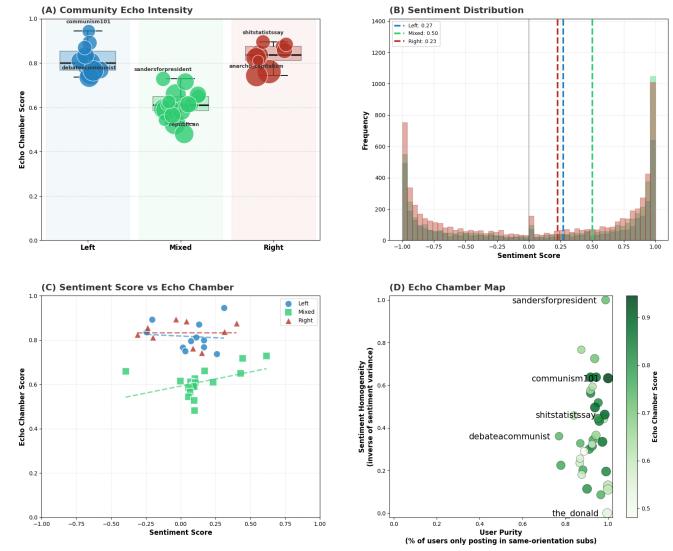


Fig. 2. (A) Distribution of echo chamber scores across left, mixed, and right communities. Mixed communities show the highest echo intensity, followed by right and left communities. Each point represents a subreddit. Point size proportional to the number of posts in each subreddit (B) Sentiment distributions across political communities. All three groups show a polarized, reverse bell-shaped pattern with posts concentrated at emotional extremes. Mixed communities have the highest average sentiment, while right communities show the lowest. (C) Sentiment score vs echo chamber intensity. Mixed communities show a large positive trends between sentiment extremity and echo strengths. (D) Echo chamber map of all subreddit. Node size is correlated with isolation score. User orientation purity (x-axis) vs sentiment homogeneity (y-axis), with node size showing structural isolation and color indicating echo chamber strength.

in general and whether their political sentiment spills into non-political forums. Unlike the previous dataset, which was used specifically to identify communities within purely political subreddits, we consider a larger dataset, which focuses on politically active users that are defined as users who have posted in politically-oriented subreddits. A user's political orientation is determined by whether they have posted more frequently in left- or right-oriented subreddits. Users with equal activity in both, or most in centrist subreddits are labeled as "mixed" and "centrist", respectively. We then compile all posts authored by these politically active users and focus on the 22,625 posts written by the 2465 unique left- and right oriented users.

With this new dataset, we assign a sentiment score to each post based on a weighted version of the LabMT sentiment scores. We would like to determine whether left- and right oriented users switch tones according to the subreddit in which they post. Therefore, we look into the distribution of sentiment scores in non-political subreddits vs. political subreddits. The results from [Figure 4](#) show that regardless of the subreddit politically active users post in, they tend to have a slightly positive sentiment (median > 0.5). With sentiment scores ranging primarily between -2 and 2 , the median difference in political vs. non-political sentiment is slightly negative at -0.094 for left-oriented users and -0.241 for right oriented users. This indicates that politically active users tend to have a more positive tone in non-political subreddits compared to political ones, where right-oriented users tend to have the largest difference. The distribution in [Figure 4](#) shows that sentiment between left- and right oriented users

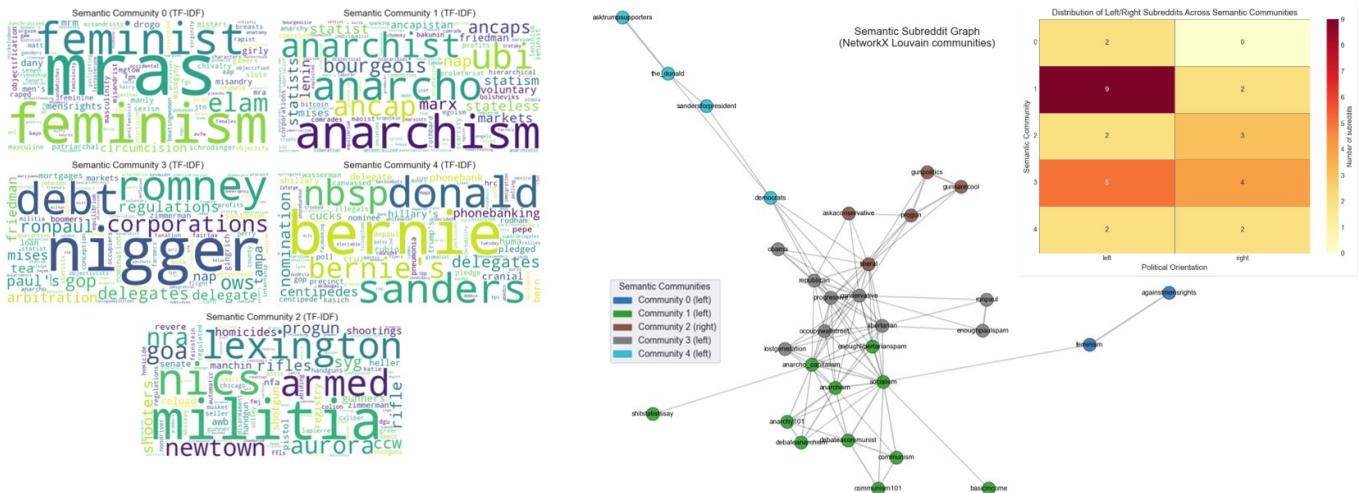


Fig. 3. Semantic subreddit network based on TF-IDF vocabulary and Louvain community detection. The network (center) shows five semantic communities with characteristic vocabulary revealed in word clouds: Community 0 (blue) emphasizes feminist discourse, Community 1 (green) focuses on anarchist themes, Community 2 (brown) centers on gun rights and militia topics, Community 3 (gray) discusses electoral politics, and Community 4 (cyan) features Bernie Sanders and progressive politics. The heatmap (bottom right) shows the distribution of political orientations across semantic communities. Community 1 is heavily left-leaning, while Communities 0, 2, 3, and 4 show more balanced or left-leaning compositions.

outside of political subreddits is comparable, but right oriented generally have a lower sentiment within political subreddits, thus creating a larger difference in tone.

Overall, analysis of political user activity shows that political users largely has a slight tendency to use a higher sentiment in non-political posts.

As in previous section, we used TF-IDF vectorization to find the most common words in the extended dataset, however this time we used it in combination with Non-negative Matrix Factorization (NMF) in order to find words used in patterns. This allows us to find words mentioned in context of similar topics through data-driven methods, thus reducing the somewhat subjective attempt of categorizing words into topics. We extracted the list of words from the 5 most common patterns for left- and right oriented users, which left us with the topics and words shown in Figure 4.

Both orientations share topic clusters related to everyday life (e.g., friends, school, hobbies) as well as broader ideological and societal issues (e.g., religion, rights), reflecting a mix of common interests and politically grounded themes.

However, there are also topics which separate the two orientations. Left-oriented top categories include themes related to economy, education and personal financial circumstances, such as loan, school/college, debt, and work. This pattern indicates that discussions among left-oriented users frequently involve student finances and the economic burden of education. This aligns with well-documented political priorities within the U.S. left, where policies addressing student debt relief and expanded public funding of higher education are prominent. One prominent topic cluster centers on gender rights, gender-based violence, and sexuality, indicated by terms such as gender, rape, and gay. This also reflects a well-established political trend in which left-aligned groups more frequently emphasize issues related to gender equality and LGBTQ+ protections. In contrast, right-aligned groups within U.S. politics tend to place less emphasis on these issues due

to generally more conservative positions on gender and sexuality.

One of the most prominent topics among right-oriented users relates to guns and gun regulation. Frequent use of terms such as firearm, permit, and defense suggests discussions centered on the U.S. Second Amendment and policies surrounding gun ownership. This aligns with the well-documented political position of the American right, where strong support for individual firearm rights is a core stance. Furthermore, they seem to have a common topic of Vehicles and driving. As there are words such as "accident", "gas" and "cop" this might be related to either law enforcement or gas prices, however most words just indicate common discussion regarding vehicles.

There seems to be a general trend where key political topics concerning each political orientation flow into posts on non-political subreddits, by politically active users, such as gender-based rights and gun control.

Methods

Semantic Network Construction. Subreddits are represented as TF-IDF vectors of their aggregated text, and pairwise cosine similarity is computed: $\text{sim}(i, j) = \frac{\vec{v}_i \cdot \vec{v}_j}{\|\vec{v}_i\| \|\vec{v}_j\|}$. Edges are created when similarity exceeds a fixed threshold, forming a semantic network.

Post-level sentiment is estimated using the VADER compound score in the range [-1,1]. Community and subreddit sentiment are analyzed using the mean and variance of these scores.

LABMT sentiment used the labmt data, which assigns a sentiment score to 10,000+ words, ranging from 1-9, where 9 is the happiest sentiment. Initially the sentiment was calculated using the average score of each word in the "content" feature, however upon reviewing we found that just taking the average word score was insufficient to accurately capture sentiment. Therefore, we filtered out some of the neutral filler words and applied weights in order to allow stronger emotions to have a larger impact. The sentiment score was calculated as $s_{lab} = \text{labmt score}, w = |s_{lab} - 5|$

$$\text{sentiment}_{scaled} = \frac{\sum_{i \in \text{words}} (\mathbf{s}_{lab_i} - 5) \cdot w_i}{\sum w_i}$$

Topic Extraction was done by creating a TF-IDF matrix, which weights words by their importance across all documents. Then, we applied a technique called Non-negative Matrix Factorization (NMF), which discovers hidden patterns in the TF-IDF matrix. These hidden patterns are what we consider topics, meaning each pattern of words found by NMF produces a

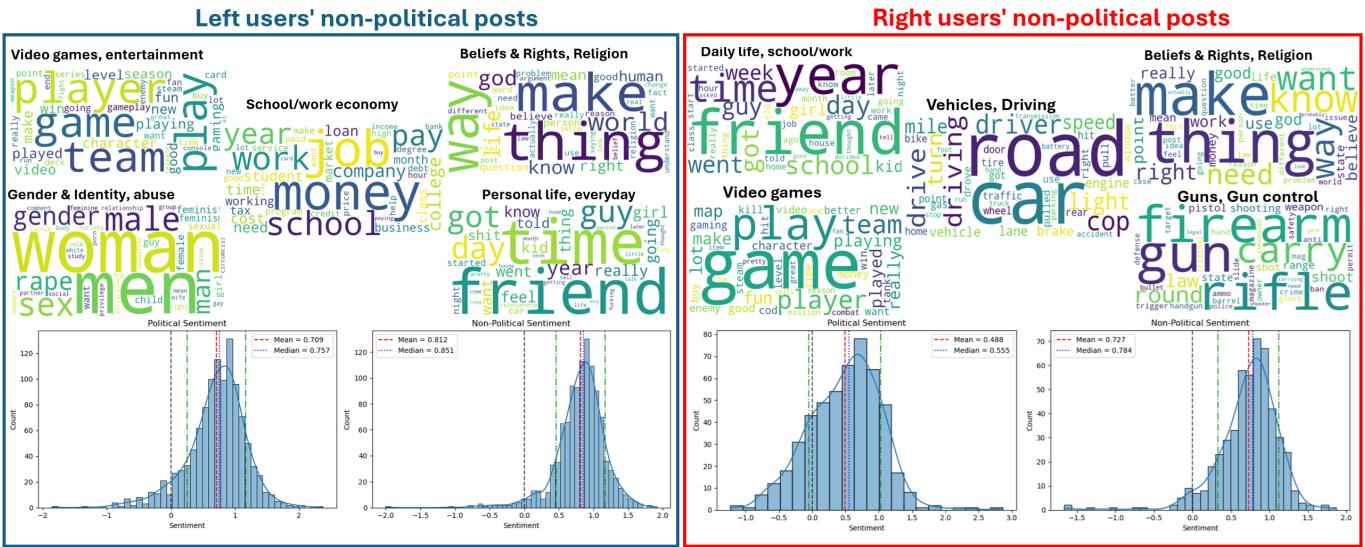


Fig. 4. Top Wordclouds of the 5 most common topics among left- (left-hand side) and right-oriented (right-hand side) users. Bottom) Bar plot over their sentiment scores in- and outside of political subreddits. Mean is marked with a red dotted line, median with a blue dotted line, and mean \pm 1std as green dotted lines.

list of keywords within said topic. We extracted the most relevant patterns in order to capture the main topics among users.

Sentiment Homogeneity is quantified as the inverse variance of sentiment or political orientation within a community, where higher values indicate stronger echo-chamber behavior: $H = 1 - \frac{\sigma^2 - \min(\sigma^2)}{\max(\sigma^2) - \min(\sigma^2)}$

The Echo Chamber Score is a composite metric (0–1) measuring how insular a subreddit is, combining four normalized components: P = Cluster Orientation Purity, T = User trappedness (percentage of users that never post in subreddits of opposite orientation), I = Structural Isolation (node's isolation score is the fraction of its total connection weight that goes to neighbors with the same orientation), H = sentiment homogeneity, and is given by the expression Echo Chamber Score = $0.4P + 0.35I + 0.15H + 0.1T$. Where

$$P = \frac{\max(\text{left}, \text{right})}{\text{left} + \text{right}}$$

Discussion. Our analysis reveals that political polarization on Reddit is more complicated than just left vs right. These patterns suggest that polarization comes from how people interact rather than what they talk about.

The three community structure we found shows, we got a pure left community, a pure right community and a mixed community containing traditional party subreddits like democrats, conservative and republican. This means the real divide isn't between Democrats and Republicans but between these traditional politics and more extreme positions. The mixed community shows that people do still interact across party lines even if they avoid more radical spaces.

It is important to note that our partitions of left- and right oriented subreddits and especially users can't be considered as a truth label of actual orientation, as it is based on classification of the general content in the Subreddit. However, as the Louvain communities support that there are clean splits between left- and right-oriented subreddits, we deemed the labeling relevant.

The political orientations and statements regarding relevance are all centered around U.S politics, however reddit is an international media, meaning there are likely non-american users posting in subreddits regarding American politics that could influence our results. Americans make up by far the largest portion of reddit users, accounting for 43% of reddit traffic in 2025(4), with UK coming second with 5.5%, which adds to why the data fits with American politics, especially as the subreddits deemed political after manual inspection are primarily global (socialism, economy, worldnews) or purely American ("Republican", "Democrats", "Obama").

Since the dataset is from 2018, updating it with newly scraped Reddit data would be a valuable future extension. We attempted to collect data from reddit, but restrictive API access requirements limited us to scraping subreddit summaries only.

The patterns we observe might reflect something deeper about American political culture. Political isolation in the US isn't new, it just takes different forms. Before social media, people were limited to local newspapers and radio stations, now we see digital isolation where people sort themselves into ideologically homogeneous online communities. Reddit didn't create political echo chambers, it just made them visible and easier to join. Our analysis shows that ideology drives network structure, with three distinct communities emerging from user behavior alone. Most importantly, polarization arises from social separation rather than different topics, both sides discuss the same issues in separate spaces. Fixing this requires more than just exposing people to opposing views, it requires rethinking how online spaces bring people together.

References.

1. TensorFlow Datasets, TensorFlow Datasets: Reddit dataset (<https://www.tensorflow.org/datasets/catalog/reddit>) (2015) Accessed: 2024.
 2. L Richardson, Beautiful soup: A python library for web scraping (<https://www.crummy.com/software/BeautifulSoup/>) (2004).
 3. J. C. M. de Vos, J. R. G. van der Wal, and J. H. den Hollander, “Clinton’s speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context, & Media* 25, 143–152 (2018).
 4. WorldPopulation Review, Reddit users by country 2025 (<https://worldpopulationreview.com/country-rankings/reddit-users-by-country>) (2025).