Decoding Echo Chambers: Analyzing Language, Sentiment and Authorship in Political Subreddits

Author: Mariano Aloiso

Abstract: This study examines the phenomenon of echo chambers by analyzing submissions and comments from seven political subreddits over a four-year period. Term frequency analysis and word clouds were used to explore the use of language. The similarity of subreddits was measured using tf-idf matrices and cosine similarity, leading to the categorization of subreddits into left-leaning, right-leaning, and neutral groups. Topic modeling and sentiment analysis was applied to understand the emotions displayed by users.

1 Introduction

The phenomenon of echo chambers has garnered significant attention in recent years. Echo chambers refer to environments - both virtual or physical - where individuals are predominantly exposed to information that amplifies their opinions and ultimately reinforces their beliefs. This is seen as a key contributor to polarization in society, since social media platforms allow for the creation of communities that can further entrench ideological divisions.

This study aims to analyze the characteristics and impact of echo chambers by applying text mining techniques to data from seven political subreddits on Reddit. Reddit is an ideal platform for this study as it consists of topic-specific communities, known as subreddits, where users can engage with content that revolves around specific topics or interests. We seek to understand how echo chambers influence users' sentiment and vocabulary, to gain insights into how they can shape individual opinions. This research addresses the following questions:

1) What are the defining features and characteristics of an echo chamber?
2) How do echo chambers function within online communities, particularly in political subreddits?

2 Methodology

This analysis was based on submissions and comments from 7 political subreddits spanning from January 2019 to December 2022. The official Reddit API was initially used during the exploration phase. However, its slow access speed and the limited retrieval capacity of a maximum of 1000 records at a time posed significant challenges. To overcome these obstacles, the data was sourced from an archive maintained by users within the Reddit community r/datasets. The archive encompassed the entire content of the 7 subreddits since their inception until December 31st, 2022. To ensure enough volume of data and focus the analysis on the most recent information, the dataset was restricted to the period from January 1st, 2019 to December 31st, 2022.

Term frequency analysis and word clouds were utilized to examine the

prevalent terms and get visual representations of the language used within each subreddit. The similarity of the subreddits was measured by creating a tf-idf matrix of the seven communities and using the cosine similarity between these corpora. This enabled the categorization of subreddits into three distinct groups: left-leaning, right-leaning, and neutral.

Latent Dirichlet Allocation (LDA) was applied to identify seven topics within the submissions across all subreddits. Sentiment analysis was conducted using the NRC lexicon to represent the polarity of sentiment and the emotions shown by the users in the comments. These methodological approaches contributed to a comprehensive understanding of the dynamics within the chosen echo chambers.

3 Data Processing

All the data processing and analysis was run on a MacBook Pro equipped with 32 GB of RAM. The initial dataset used 22 GB of storage space, in zst format at a ratio of 5:1 (the uncompressed data required an estimated 97 GB of memory, more than the available RAM). It was not possible to load all the data into memory. Instead, the data was read and processed in streaming fashion. The subsetting, cleaning and processing steps were implemented in a line-by-line reading strategy. Each line was sequentially read, decoded, processed, re-encoded, and subsequently saved to a new file. This approach allowed for an efficient use of system resources and the analysis of the large-scale dataset.

The first step required reviewing the available data to select the relevant keys and choose the appropriate time. The original dataset included the complete historical records of each subreddit, spanning as far back as 2009. The analysis was restricted to the period between January 1, 2019, and December 31, 2022, to ensure that the data was recent and that there was enough information for conducting the study. Each dataset had a varying number of keys, according to the evolution of features of the social media platform. For the submission files, the selected keys comprised "id", "author", "downs", "ups", "title", "num_comments", "created_utc", "selftext", and "score". Similarly, for comments, keys retained were "link_id," "author," "created_utc," "body," "score," "ups," "downs," "controversiality," and "gilded." The resulting subsets were up to 96% smaller than the original datasets.

The submissions were prepared for analysis through a series of data cleaning and data normalization steps. First, submissions that were marked as deleted or removed were excluded, along with any submissions made by moderators, in order to focus on user-generated content. Submissions with a low number of comments (less than 3) or low score (less than 5 votes) were not included. Hyperlinks, subreddit names (/r), and URLs were removed, along with emojis. Finally, non-alphabetical characters (except for ! and ?) were removed and extra whitespaces were stripped.

Submissions Files Size			
	Original	**Subset**	**Clean**
r/politics	903.5 MB	61.4 MB	797 KB
r/Conservative	252.4 MB	34.5 MB	743 KB
r/PoliticalDiscussion	54.3 MB	6.1 MB	260 KB
r/Republican	50.3 MB	6.8 MB	178 KB
r/democrats	35.1 MB	5.6 MB	1.4 MB
r/NeutralPolitics	24.5 MB	1.1 MB	91 KB

r/progressive	5.3 MB	871 KB	287 KB

The same steps were applied to comments. Comments with less than 5 interactions (either total number of up/down votes, total score or number of replies) were excluded to reduce the size of the final dataset.

Comments Files Size			
	Original	**Subset**	**Clean**
r/politics	18.05 GB	7.07 GB	6.37 GB
r/Conservative	1.45 GB	849.1 MB	751.4 MB
r/PoliticalDiscussion	867.8 MB	227.2 MB	202 MB
r/Republican	119.1 MB	53.1 MB	47.7 MB
r/democrats	101.8 MB	51.4 MB	45.3 MB
r/NeutralPolitics	87.7 MB	17 MB	13.3 MB
r/progressive	41.1 MB	4.8 MB	4.3 MB

The clean datasets were small enough to be loaded into memory for analysis.

4 Exploratory analysis

The first step in evaluating the use of language in each subreddit was creating word clouds of the most frequent terms. Fr a more accurate analysis, common English stop words were removed using the NLTK corpus, along with custom terms that tended to dominate the word clouds, such as "like," "ever," "I've," "Trump," "karma," and the names of the subreddits themselves. Filtering out these common n grams helped focus the analysis on the distinctive vocabulary patterns that characterized each subreddit.

The analysis of frequent terms in r/politics reveals a diverse range of topics and vocabulary. Notably, terms such as "severely," "colluded," and "implicated" suggest a focus on critiquing political figures or institutions. References to "cow," "sport," and "elderly" indicate discussions related to varied subjects beyond politics, a reflection of the diverse interests and the number of users in the community. The presence of terms like "palestinians," "sweden," “indian,” and "ussr" indicates the presence of discussions on international affairs. Furthermore, words such as "nuanced," and "contributing" suggest an appreciation for multiple perspectives and detailed analysis. The combination of these terms in the r/politics subreddit reflect the community’s interest on topics that extend beyond American politics, in contrast with the rest of the subreddits.

Fig 4.1: r/politics

Among the most common unigrams in r/Conservative, words such as “cheaper”, “relief” and “proves” indicate a concern for economic factors and fiscal responsibility. Terms like “tds” (short for Trump Derangement Syndrome) and “republicans” suggest a strong emphasis on partisan politics and ideological loyalty within the subreddit, while the words “incident”, “insult” and “violation” hint at the community’s engagement with contentious topics, another trait of echo chambers.

There is also evidence of political partisanship in r/progressive, reflected on the frequent use of words like “democrats”, “power”, “republican” and “Biden”. Unigrams like “vote”, “make” and “know” indicate an

active engagement with policy-making. Finally, the terms “good”, “bad”, and “better” suggest a focus on contrasting political ideologies.

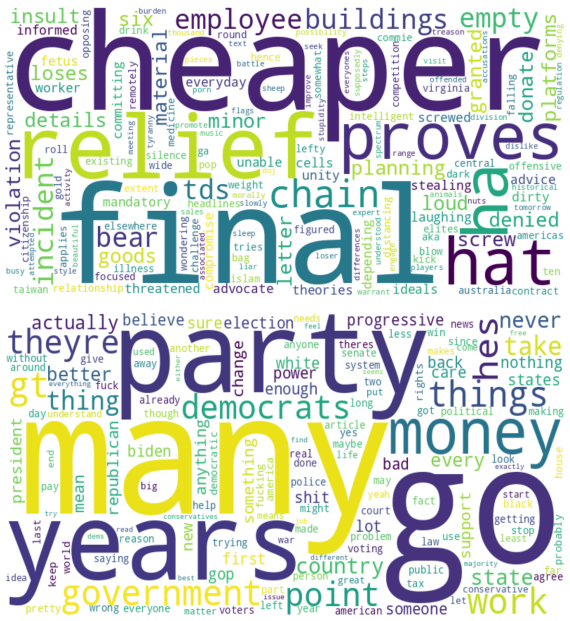


Fig 4.2: r/Conservative and r/progressive

r/democrats and r/Republican subreddits also exhibit characteristics of echo chambers. The informal tone of the conversation (as seen by the frequent use of “fucking”, “shit” and “lol”) fosters group cohesion in the community. In democrats, terms such as ”agree”, ”hope”, ”change”, and “pretty” indicate a strong sense of agreement, as members primarily interact with like-minded individuals. Some of these terms are also seen in the word cloud for r/Republican.

In r/Republican, terms like “democrats”, “dems”, and “left” indicate a focus on partisan politics, as in the case of the subreddit r/Conservative. The repetition of terms like “literally”, “enough”, and “pretty” may reinforce existing narratives.

Some distinctions can be drawn in the echo chamber dynamics of both subreddits. r/democrats appears more focused on agreement, hope, and change, while r/Republican emphasizes problem-solving, support, and evidence.

Fig 4.3: r/Republican and r/democrats

There are noticeable differences in the language used in r/NeutralPolitics and r/PoliticalDiscussion with respect to the more partisan oriented subreddits. The terms “states”, “government”, "election” and “political” are evidence of an emphasis on discussing focus on political issues over giving opinionated comments. Unigrams such as “make”, “point”, “actually”, “well allowed”, “moderators”, “opinion”, and “discussion” are indicative of the communities’ commitment to open dialogue and to exploring diverse perspectives. These linguistics patterns are a departure from the echo chamber dynamics observed in the other subreddits.

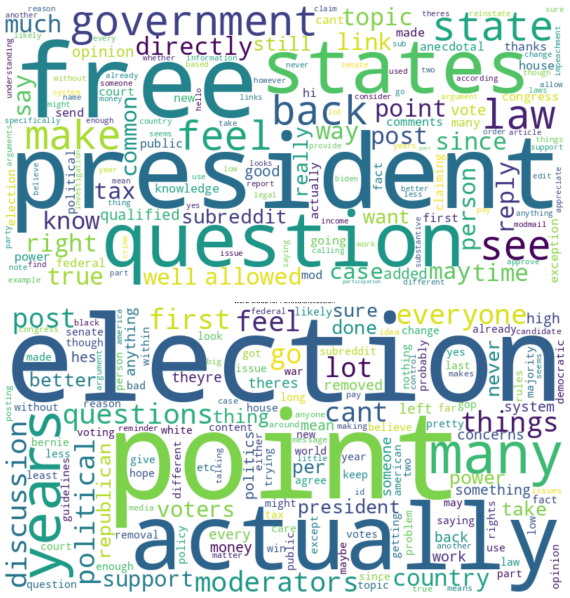
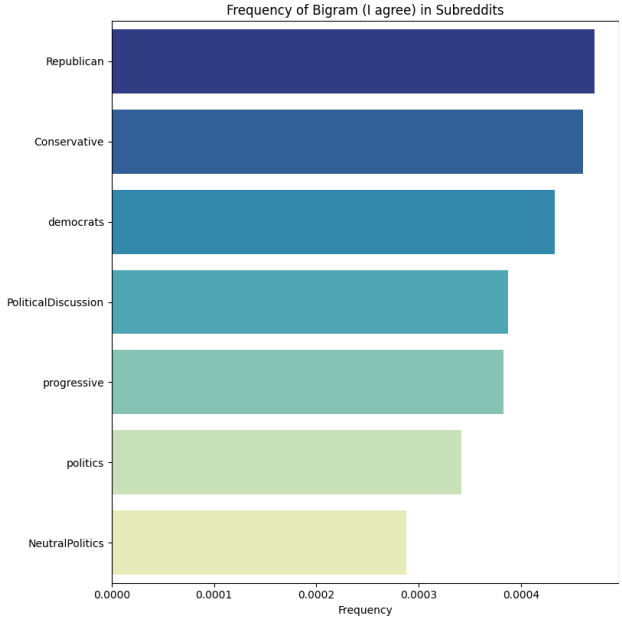


Fig 4.4: r/NeutralPolitics and r/PoliticalDiscussion

Another aspect to consider is the frequency of bigrams that indicate agreement or disagreement. The findings demonstrate that partisan subreddits exhibit a higher frequency of bigrams such as “I agree”, suggesting a greater decency for members to express concurrence. Conversely, the bigram “I argue,” which signifies disagreement, is more commonly observed in the politically neutral subreddits. This disparity contributes more evidence on the different levels of echo chamberness of these subreddit, with partisan communities displaying a higher likelihood of reinforcing shared beliefs and opinions.



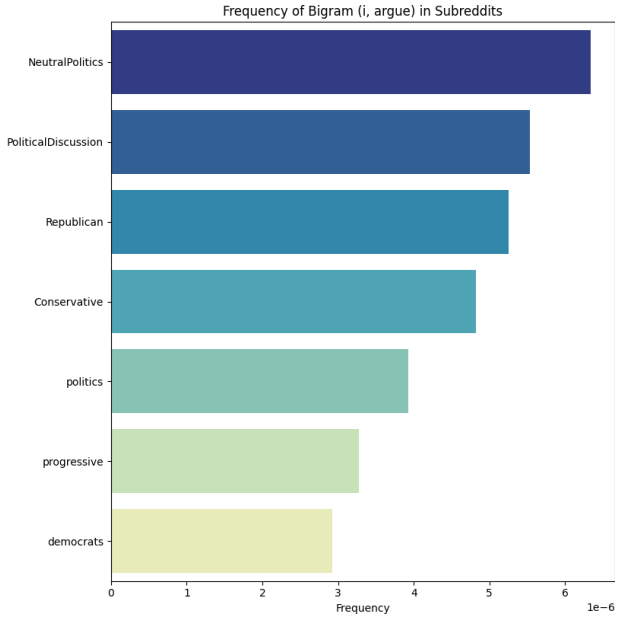


Fig 4.5: bigrams frequency

5 Content similarity

To assess the degree of similarity among the content of the subreddits, the submissions were processed to construct a term frequency-inverse document frequency (tf-idf) matrix of dimensions 1000 by 7. The cosine similarity among the corpora was used as a measurement of comparability.

This metric considers the angle between two vectors and yields a value ranging from 0 to 1, with 1 indicating perfect similarity.

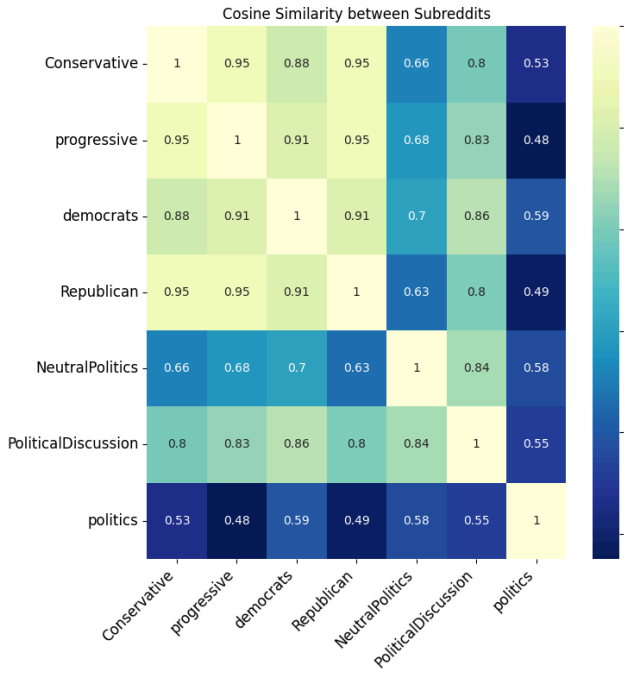


Fig 5.1: Similarity between Subreddits

The figure reveals notable patterns. Conservative shows a high degree of similarity with other political subreddits such as progressive, democrats and Republican. The progressive and democrat subreddits exhibit a strong similarity, implying shared themes and vocabulary within these communities. In contrast, NeutralPolitics displays comparatively lower similarity with other subreddits, potentially due to its rules that attempt to maintain a neutral and fact-based discussion. The politics subreddit shows a relatively low overall similarity, suggesting a broader range of topics compared to the more ideologically focused subreddits. It is worth pointing out that politics has the highest similarity with the democrats and Neutralpolitics subreddit.

6 Topic modeling

Latent Dirichlet Allocation (LDA) was used to identify the common topics across all the subreddits. A set of custom words such as “Obama" and “Trump” was removed from the corpora, as these words appeared consistently across all topics and added little value. LDA was run using 3 to 9 topics and the results were evaluated manually and relying on coherence scores to determine the optimal number of topics.

Num of Topics	Coherence Score
3	0.5667
4	0.5891
5	0.5581
6	0.5001
7	0.5263
8	0.5239
9	0.4957

Coherence score measures the semantic coherence of topics by quantifying the degree of semantic similarity between high-frequency terms within each topic. A higher coherence score indicates more coherent and interpretable topics. The experiments using 3 to 5 topics yielded favorable coherence scores. However, after looking at their terms, it became apparent that the topics lacked breadth for a useful in-depth analysis.

The optimal number of topics was determined to be 7. This configuration had a satisfactory coherence score, and presented easily interpretable terms for each topic. This is the model that was ultimately used to classify the submissions.

Topic	Terms

Judicial System	house, court, supreme, committee, judge, state
Political Figures & Investigations	january, desantis, election, mar-a-lago, donald, fbi, joe
International Affairs	us, ukraine, covid, arizona, war, health, china
State Government	bill, texas, states, law, tax, governor, plan, state
Federal Government	senate, vote, capitol, gop, race, doj, us, general
Social Issues	abortion, climate, party, voting, voters
Other	people, one, dont, opinion, time, like, student, right

The topic "Other" includes submissions that did not align with any of the other topics and were discarded before conducting sentiment analysis.

A Hierarchical Dirichlet model (HDP) was used as an alternative approach for topic modeling. HDP is a Bayesian non-parametric model that estimates the number of models using a Dirichlet distribution. The model suggested 20 topics, which was deemed impractical as it made further analysis challenging.

7 Sentiment analysis

The subreddits were grouped according to their political leaning, which was determined based on their respective community guidelines and their cosine similarity. The grouping was as follows:

- Right-Leaning: r/Conservative and r/Republican.
- Neutral: r/NeutralPolitics and r/PoliticalDiscussion.
- Left-Leaning: r/democrats and r/progressive

The subreddit r/politics was excluded due to the predominance of submissions on international affairs, and its significantly larger user base compared to the other subreddits, which would introduce a bias into the results.

To obtain the community's polarity towards the topic, the negative and positive sentiments of the comments under each submission were calculated using the NRC Lexicon. The sentiment of each individual comment was multiplied by the score, (difference between upvotes and downvotes). This was done to give more weight to those comments that had a lot of interactions. Comments with no votes were removed for efficiency, as it was found that they did not contribute significantly to understanding of the prevailing attitudes in the community. The polarity was calculated by adding the negative and positive sentiments across all submissions under the same topic, and normalizing by the total emotion:

Polarity = (pos - neg) / (pos + neg)

This gives a score between -1 and 1, where 0 indicates a neutral sentiment.

The polarity scores indicate a slightly negative sentiment along the right-leaning subreddits in the topic "Judicial System," with a polarity of -0.0279. Conversely, both the left-leaning and neutral subreddits exhibit a positive sentiment towards this topic, with polarities of 0.0267 and 0.0239, respectively.

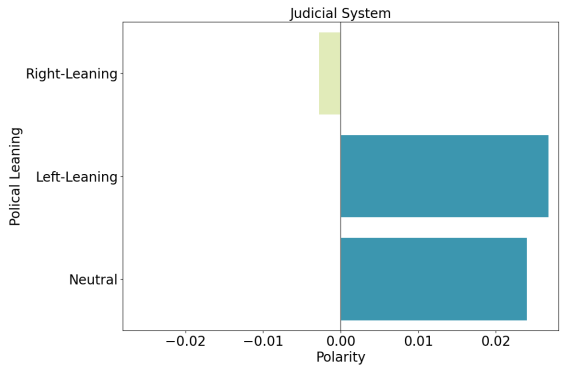


Fig 7.1: Judicial System - polarity

In the topic of "Political Figures & Investigations," the right-leaning subreddits display a polarity of 0.0055, indicating a relatively modest positive sentiment. In contrast, the left-leaning subreddits have a higher positive sentiment towards this topic, with a polarity of 0.0390. The neutral subreddits have a neutral polarity of -0.0015. This can be interpreted as indifference towards submissions that discuss political figures, and indicates a community less focused on partisan politics.

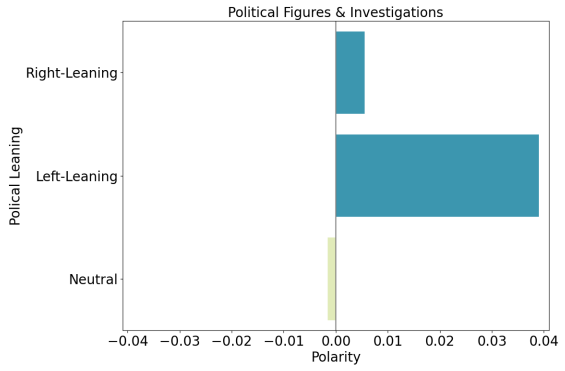


Fig 7.2: Political Figures & Investigations - polarity

Shifting focus to the topic of “International Affairs,” the evidence shows that the neutral-leaning subreddits have a higher positive sentiment of 0.1348 versus the other groups (left-leaning: 0.1255; right-leaning: 0.0277). The fact that all scores are positive suggests that discussions related to international affairs are generally positively received across all groups, in particular among users in the neutral communities.

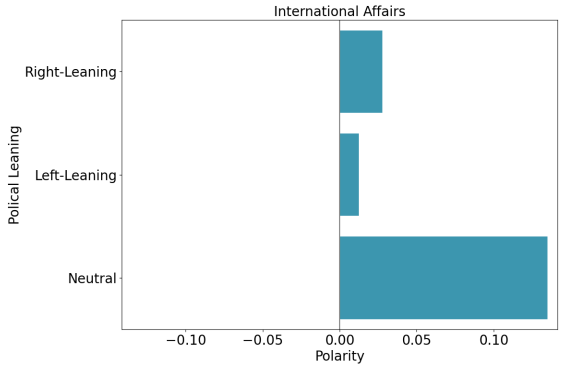


Fig 7.3: International Affairs - polarity

As for "State Government," the sentiment scores highlight interesting dynamics. The right-leaning subreddits display a positive sentiment of 0.0284, indicating a generally favorable view towards state government. In contrast, the left-leaning subreddits exhibit a significantly negative sentiment, with a polarity of -0.0973, while the neutral subreddits have a relatively neutral sentiment of 0.0033. These results indicate that the perception of state government is more favorable within the first group, while the left-leaning subreddit expresses a strong negative sentiment. The high scores may suggest the convergence of opinions.

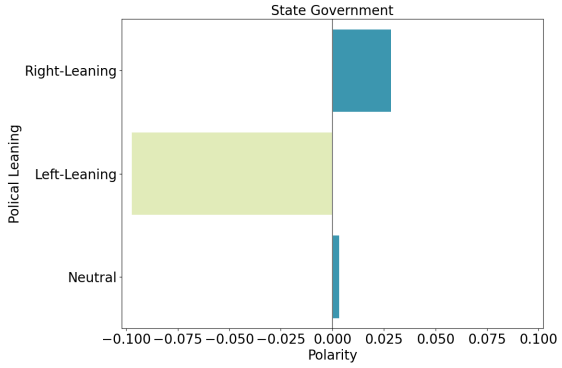


Fig 7.4: State Government - polarity

The scores for the topic “Federal Government” show that all subreddits express positive sentiments, although with different magnitudes. It is important to note that the data encompasses two different administrations on a federal level, and this may distort the analysis.

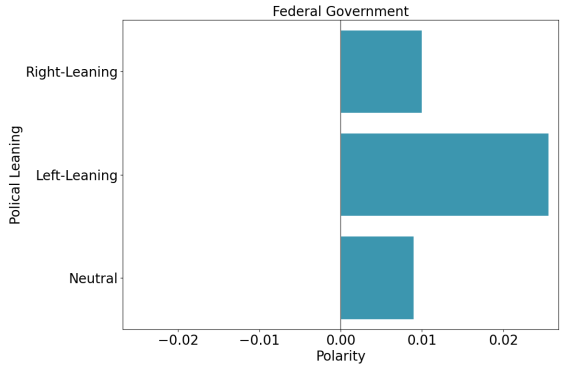


Fig 7.5: Federal Government - polarity

On the topic of "Social Issues," we observe more distinct patterns. The right-leaning subreddits display a notably negative sentiment (-0.1435), while the left-leaning subreddits express a positive sentiment (0.0568). These opposing sentiments indicate a potential echo chamber dynamic within each subreddit regarding discussions on social issues. Users are more likely to interact in the communities that reinforce their own beliefs. The neutral subreddits, with its close-to-neutral sentiment, provides a more balanced space for discussions on social issues, serving as a counterbalance to the echo chambers observed in the politically aligned subreddits.

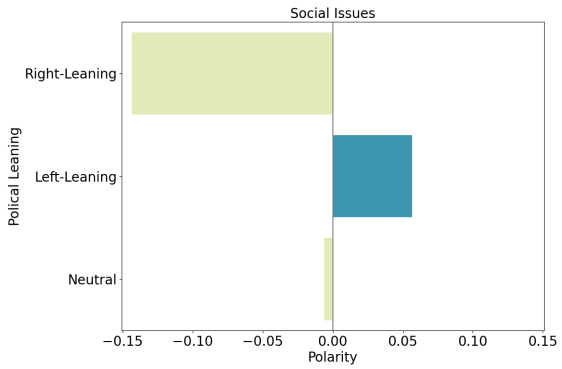


Fig 7.6: Social Issues - polarity

In general, while some topics have a more balanced sentiment across the political leaning groups, such as international affairs, others reveal stronger patterns of echo chamber dynamics, as exemplified in the topic of social issues. The neutral subreddits consistently demonstrate a more neutral sentiment, a product of the community’s guidelines that seek to foster diverse perspectives and mitigate echo chamber tendencies.

8 Submissions Distribution

Sentiment analysis revealed that partisan subreddits exhibit higher polarity scores on contentious topics. To add more context, the next step was to delve deeper into the authors of the submissions. Examining the distribution of submissions by authors can shed light into who is actively contributing to the formation of echo chambers. This involved finding the top Redditors by number of posts in each community.

In the partisan subreddits r/Republican and r/democrats, an analysis of the submissions revealed interesting patterns. In r/Republican, the top 20 authors were responsible for 18.6%. The distribution of contributions among these top users was relatively balanced, with each user contributing between 0.6% and 1.6% of the

submissions. This suggests a small cohort of highly active users. On the other hand, the distribution of submissions in r/democrats was notably different. While the top 20 authors are responsible for 23.5% of the community's submissions, just three users accounted for 16% of the total, while the remaining users in the top 20 posted less than 0.2% of the submissions in the subreddit. A look into these three users' history shows that they primarily post links to news articles. This observation raises the possibility that these are automated accounts or bots created to disseminate news content.

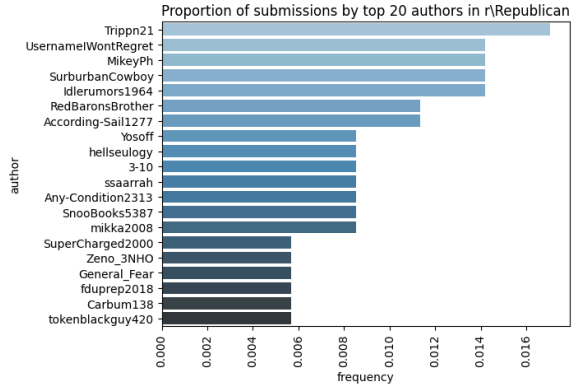


Fig 8.1: Top authors - r/Republican

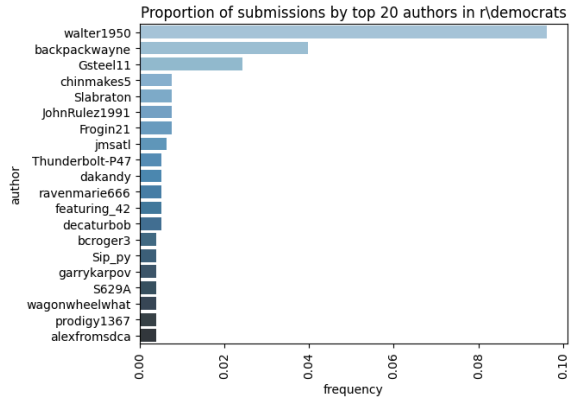


Fig 8.2: Top authors - r/democrats

In the neutral-leaning subreddits r/PoliticalDiscussion and r/NeutralPolitics, the top 20 authors contribute with 16% and 33.5% of the submissions respectively. The distribution of submissions followed a Poisson distribution, with each successive user contributing a progressively smaller percentage of the total submissions. This suggests a higher level of participation from a smaller number of active users in these neutral-leaning subreddits.

The concentration of posts among top contributors in r/NeutralPolitics (at twice the rate of r/PoliticalDiscussion) can be explained by the active participation of mods that constantly curate the content in the subreddit to maintain neutrality. This is part of their community guidelines and one of their main tenets.

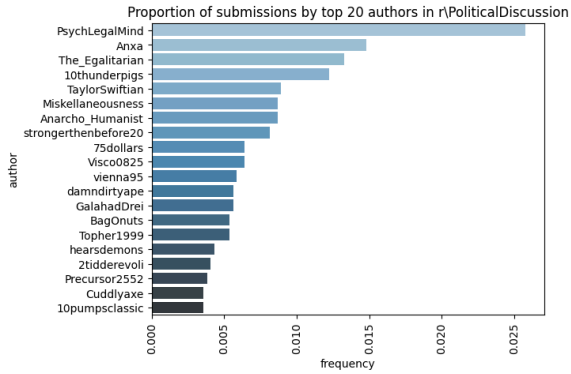


Fig 8.3: Top authors - r/PoliticalDiscussion

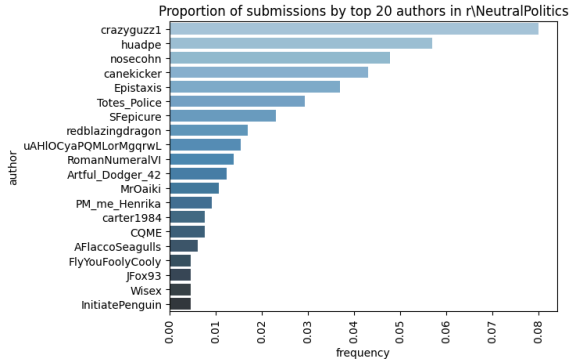


Fig 8.4: Top authors - r/NeutralPolitics

The findings on the submission rates and the concentration of contributions highlights some key features of echo chambers. The prevalence of highly active users in partisan subreddits may be a contributor to the reinforcement of echo chambers dynamics, while a tightly

moderated environment seems to foster the creation of content from different authors.

9 Further work

Exploring the changes of topics over time within subreddits can provide a deeper understanding into the factors that shape these communities. Examining how key events, such as elections or significant news, influence which topics are discussed could be valuable to discern how users' reactions differ when they take place within echo chambers.

Furthermore, sentiment analysis can be enhanced by employing more advanced sentiment modeling techniques. These techniques allow for a more nuanced understanding of the diverse range of emotions expressed by community members that goes beyond positive/negative polarity.

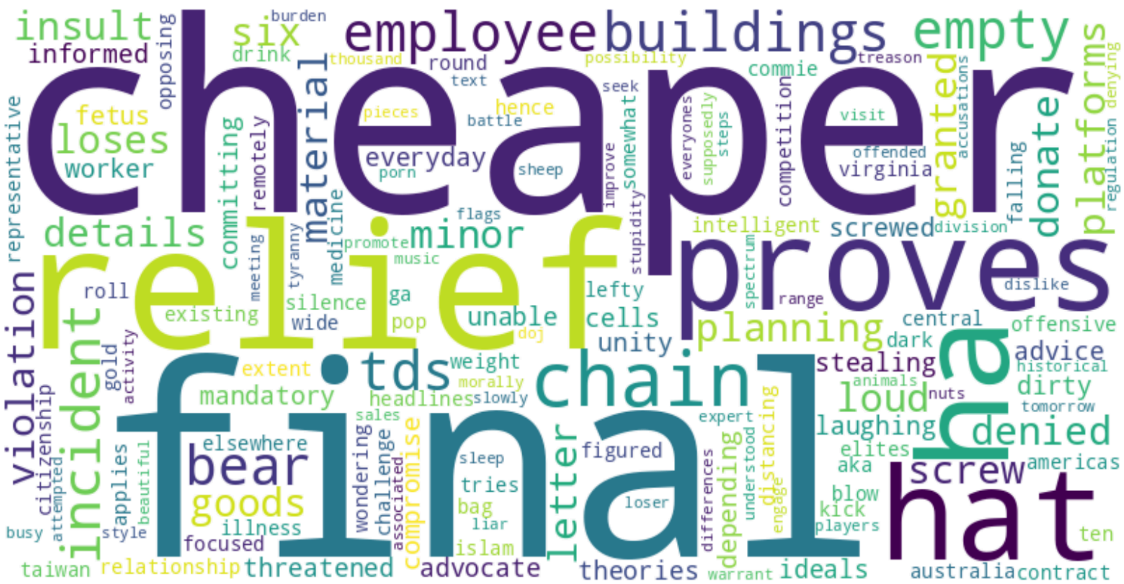
Finally, expanding the analysis beyond individual comments by incorporating user interactions can provide a holistic understanding of echo chamber dynamics. A comment with a high number of replies may have the power to influence users' sentiment towards the topic of the submission. This analysis of user interactions can enrich our understanding of sentiment alignment and the reinforcement of opinions that happens in echo chambers.

10 References

- Blei, D. M., Jordan, M. I., Griffiths, T. L., & Tenenbaum, J. B. (2003). Hierarchical Topic Models and the Nested Chinese Restaurant Process. Neural Information Processing Systems.
- The PRAW Development Team. (n.d.). PRAW: The Python Reddit API Wrapper. Retrieved from https://reddit-api.readthedocs.io/en/latest/
- Řehůřek, R., & Sojka, P. (2010). Gensim: Python framework for vector space modeling. Retrieved from https://radimrehurek.com/gensim/index.html
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "Reilly Media, Inc."
- Crowdsourcing a Word-Emotion Association Lexicon (2013), Saif Mohammad and Peter Turney, Computational Intelligence, 29 (3), 436-465.

11 Figures

Fig 4.2: r/Conservative and r/progressive



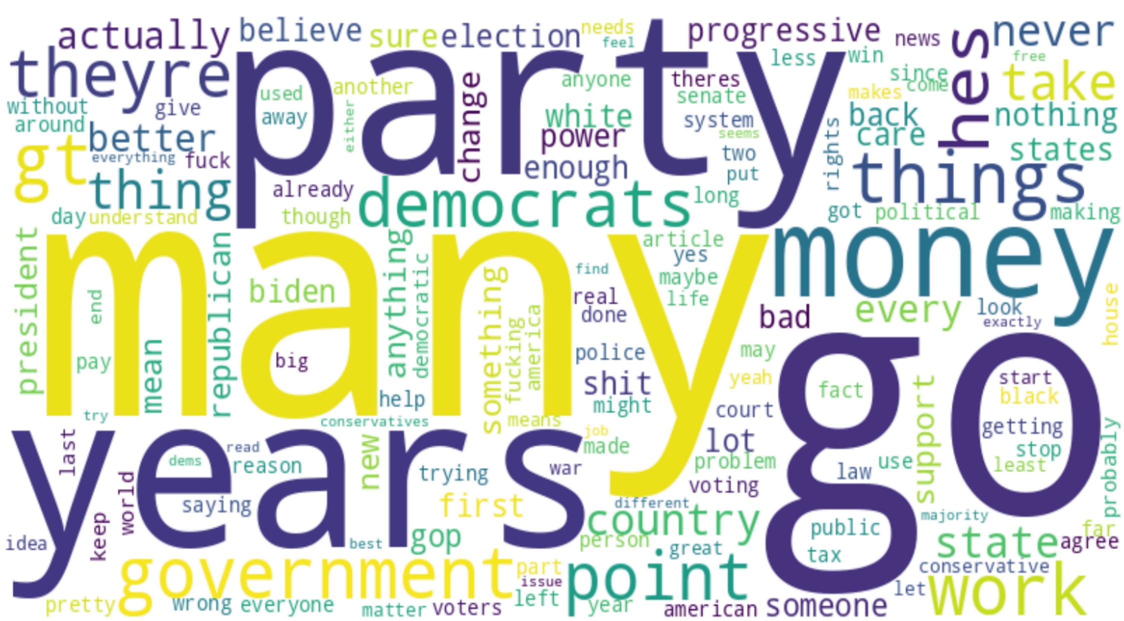


Fig 4.3: r/Republican and r/democrats

Fig 4.4: r/NeutralPolitics and r/PoliticalDiscussion

Fig 4.5: bigrams frequency

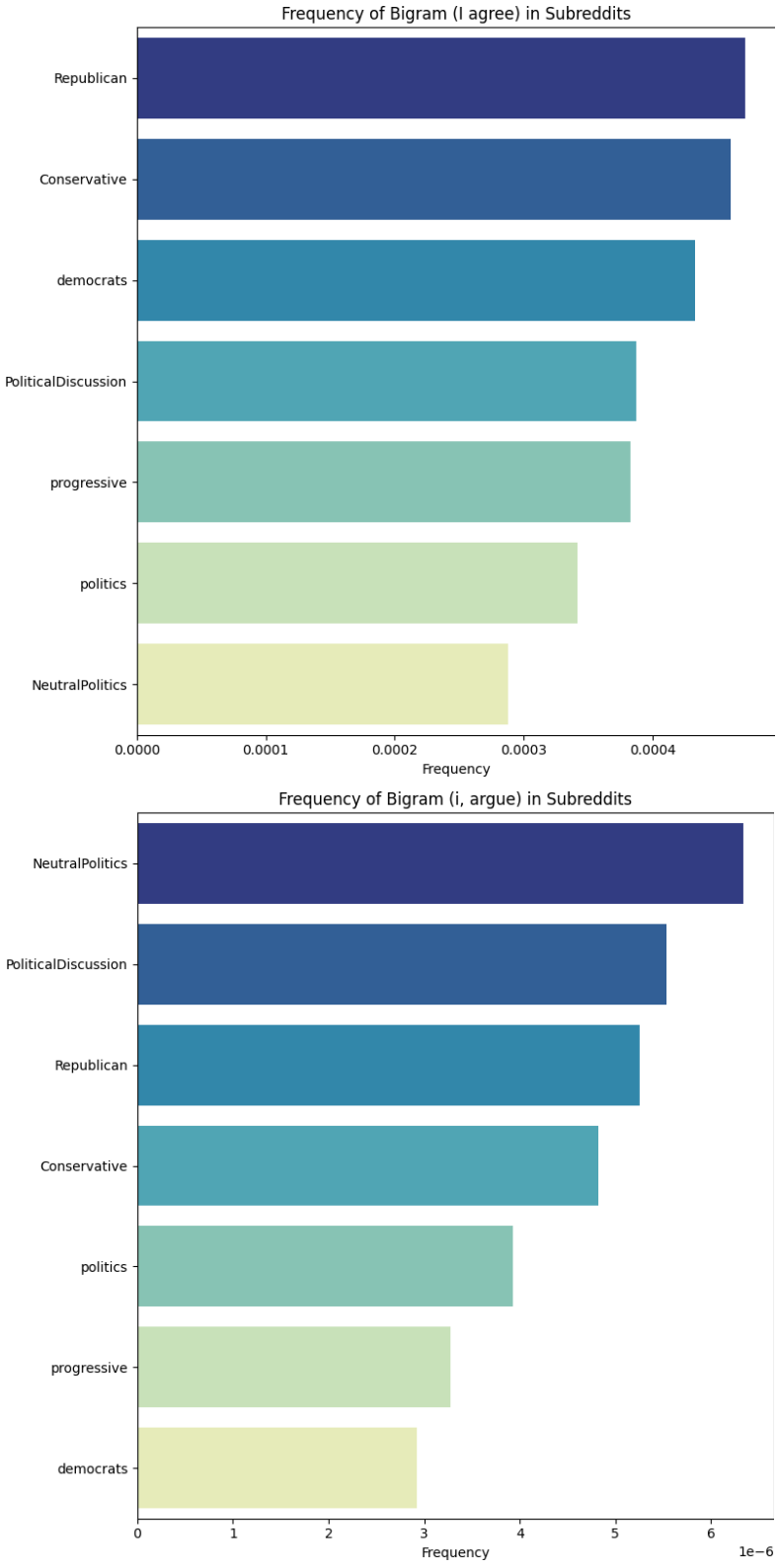


Fig 5.1: Similarity between Subreddits

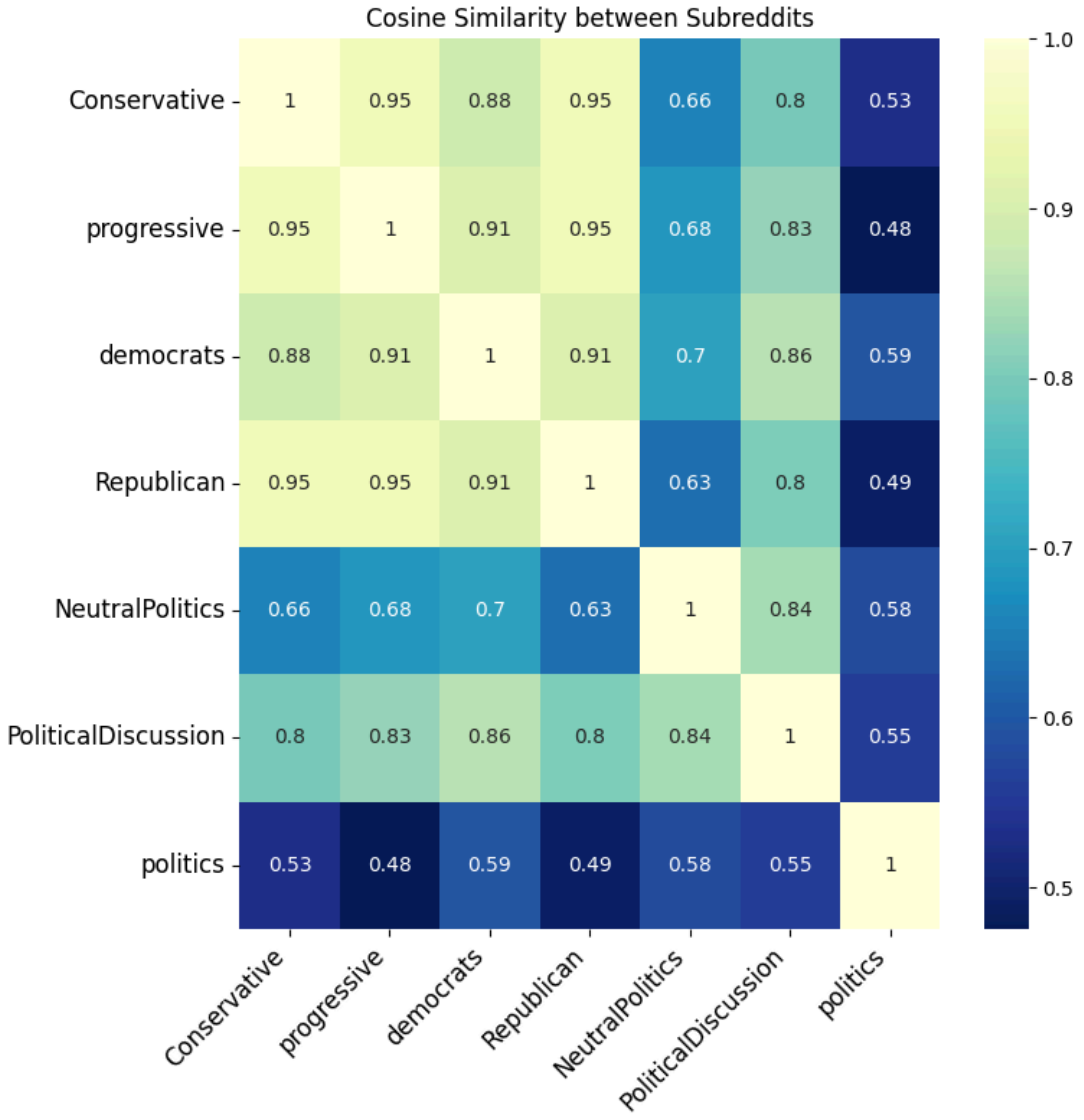


Fig 7.1:Judicial System - polarity

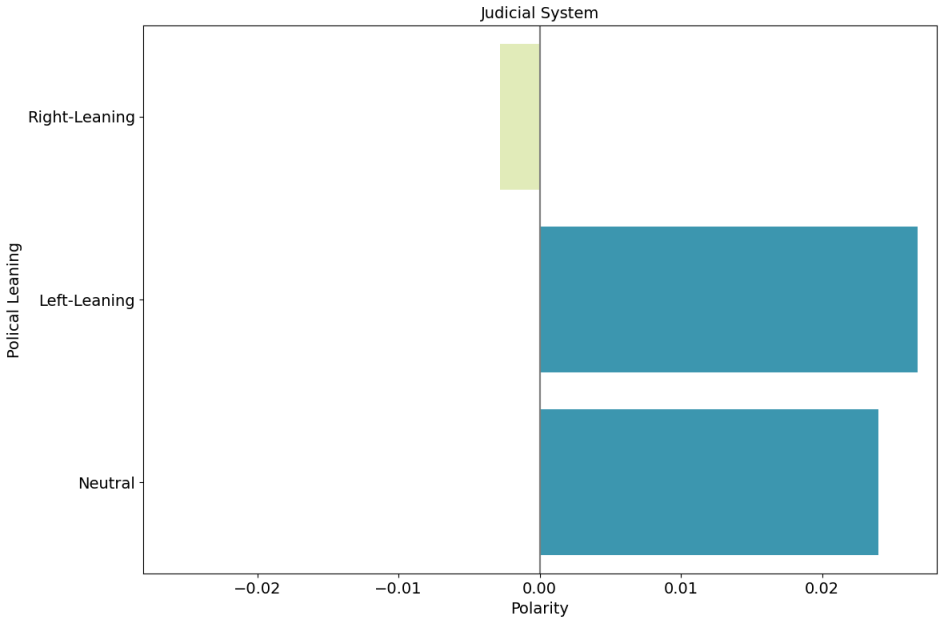


Fig 7.2: Political Figures & Investigations - polarity

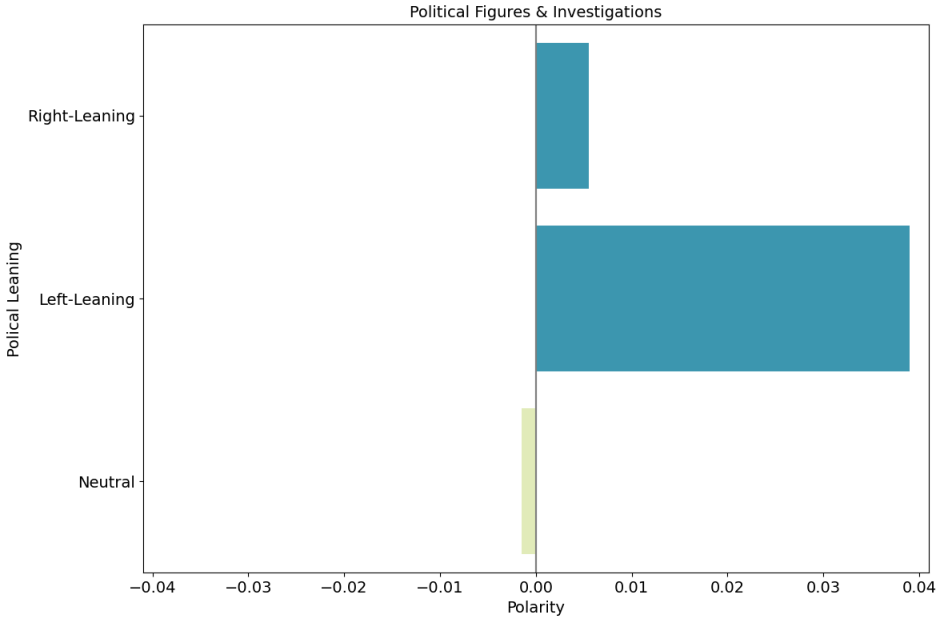


Fig 7.3: International Affairs - polarity

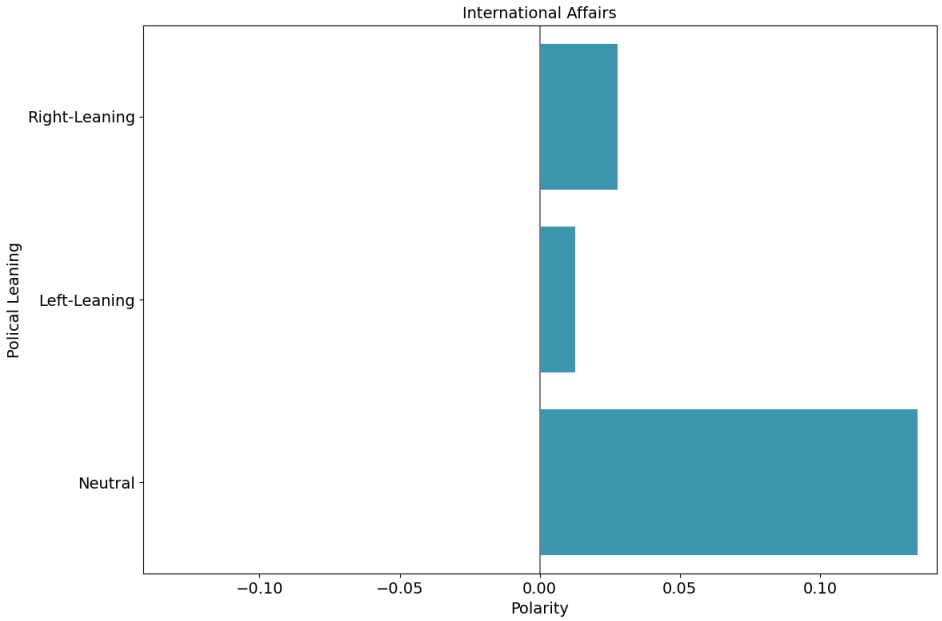


Fig 7.4: State Government - polarity

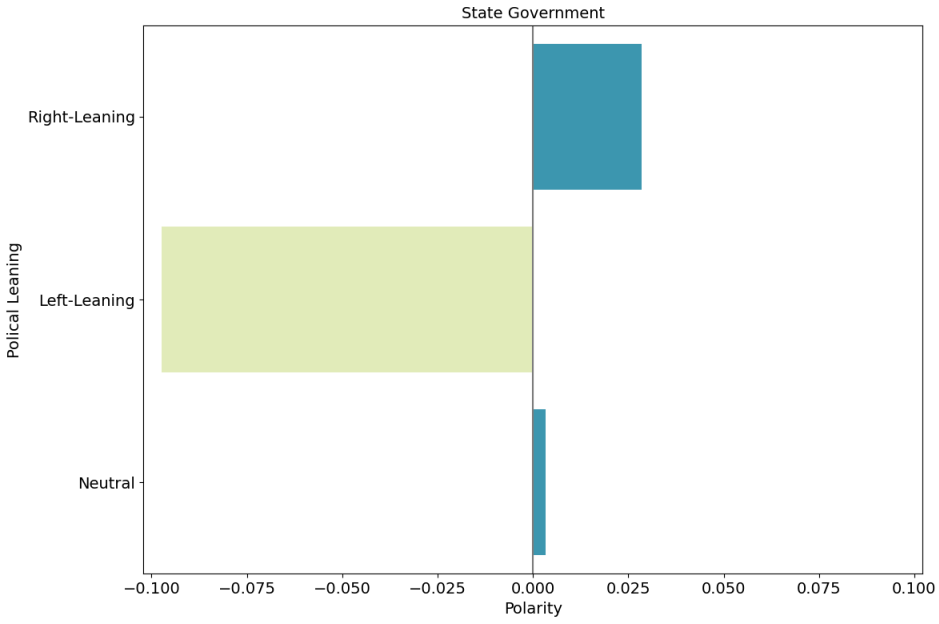


Fig 7.5: Federal Government - polarity

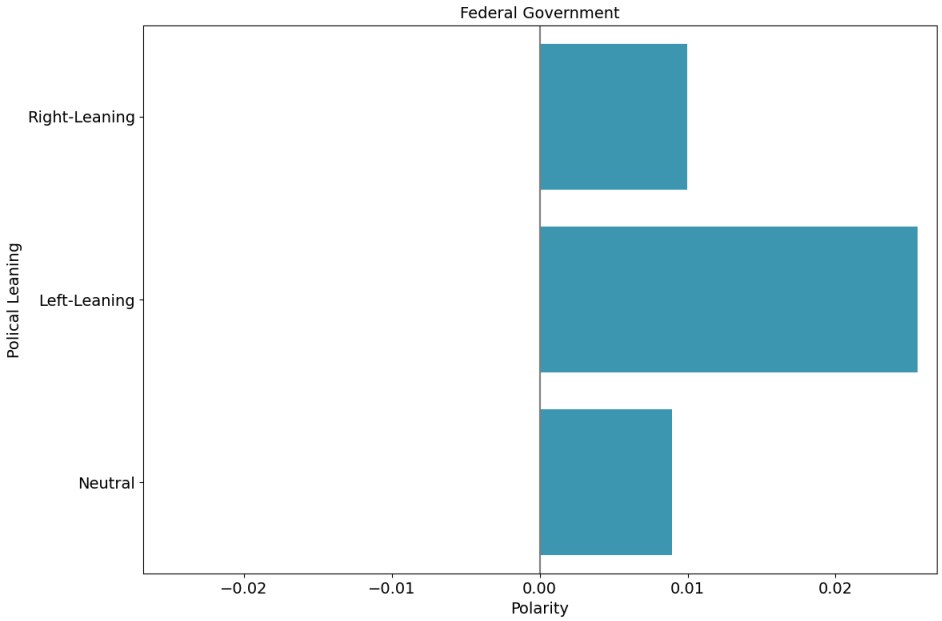
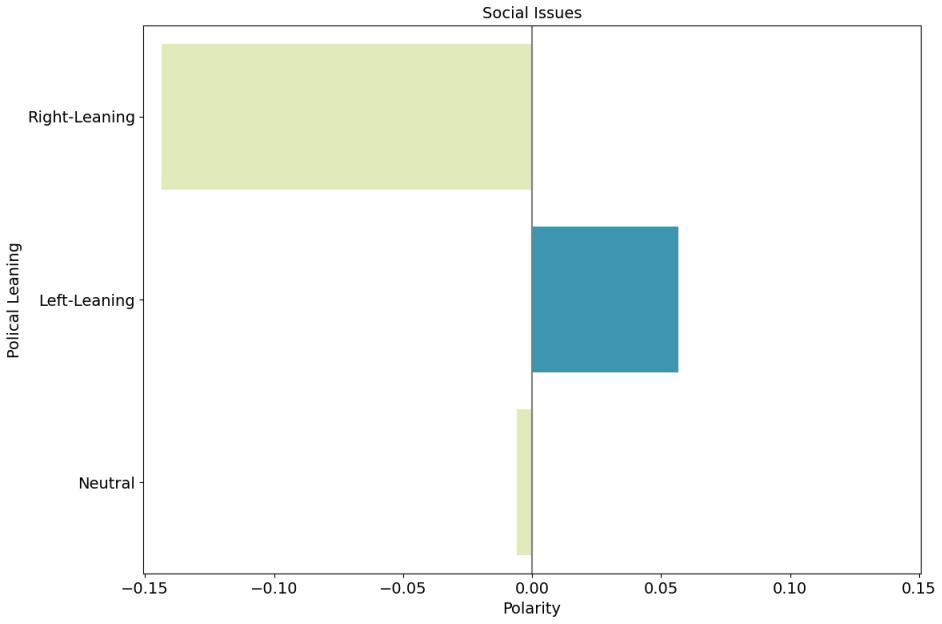
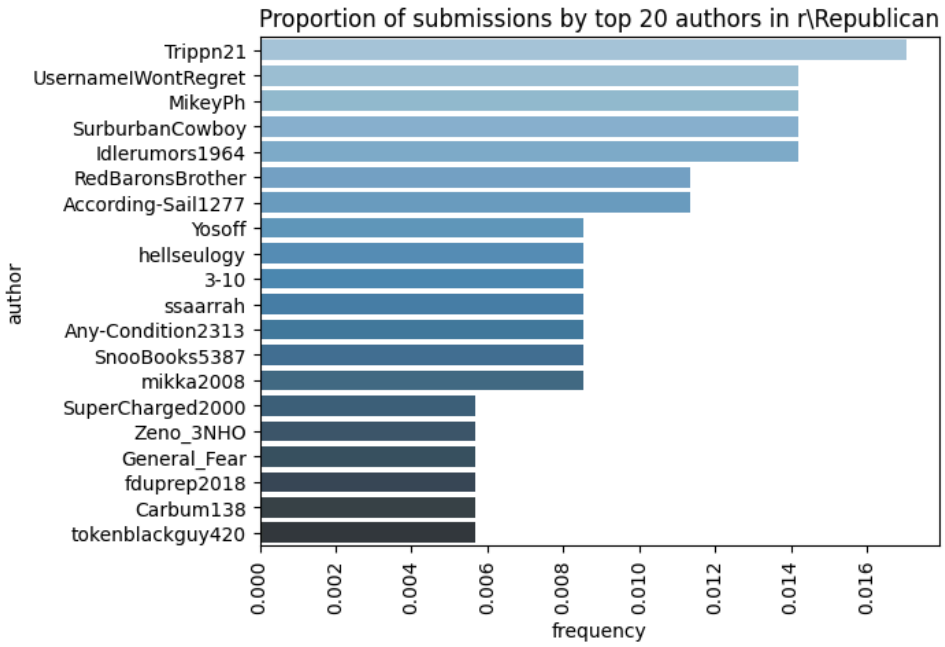
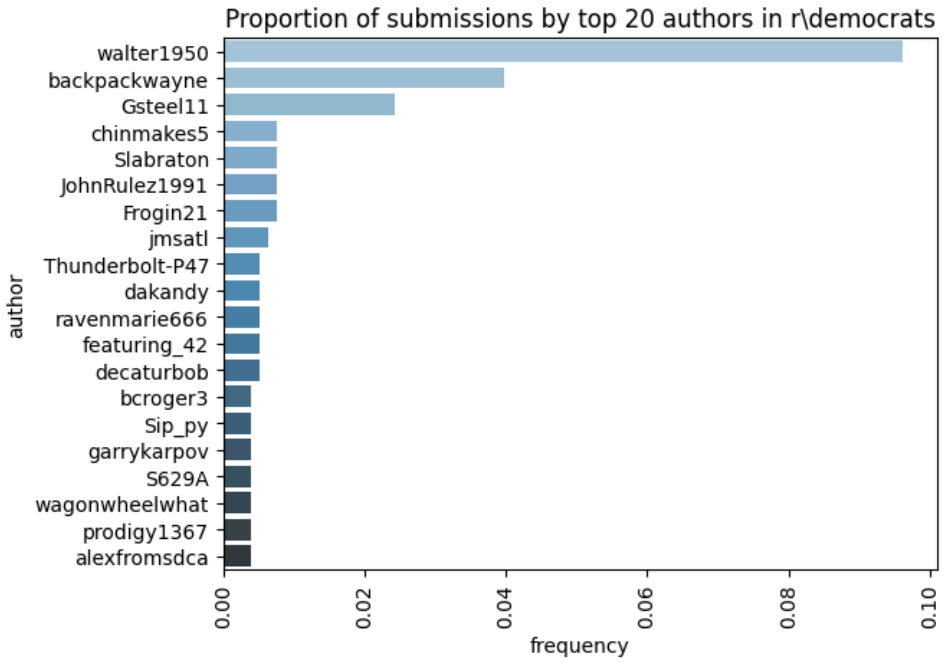
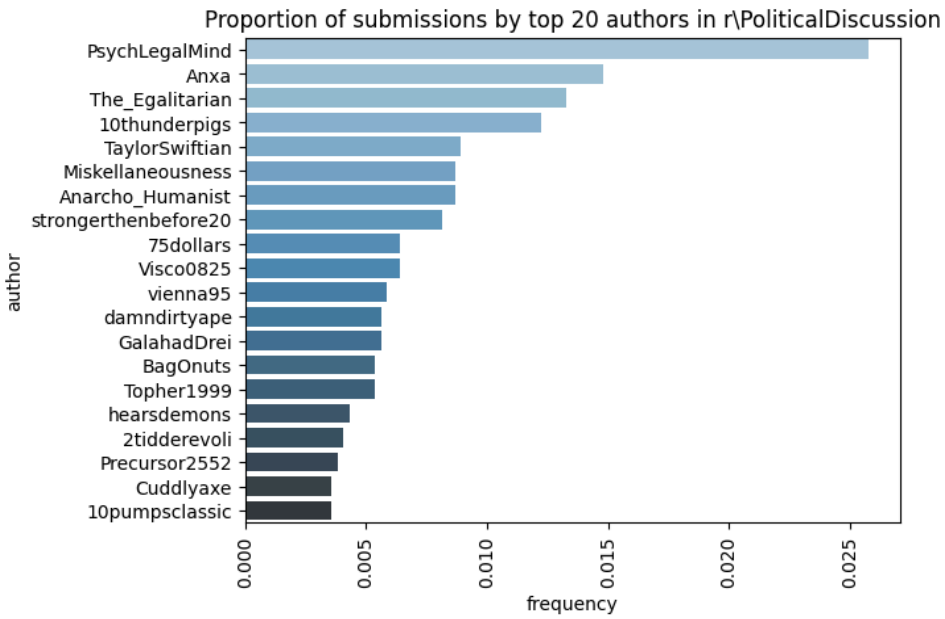


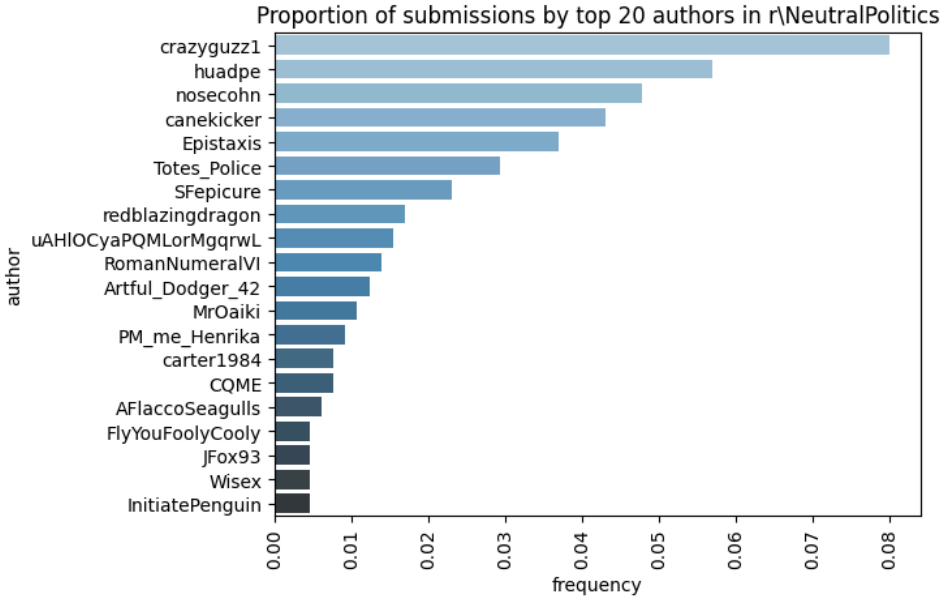
Fig 7.6: Social Issues - polarity




Proportion of submissions by top 20 authors in r\Republican
Trippn21
UsernameIWontRegret
MikeyPh
SurburbanCowboy
Idlerumors1964
RedBaronsBrother
According-Sail1277
Yosoff
hellseulogy
3-10
ssaarrah
Any-Condition2313
SnooBooks5387
mikka2008
SuperCharged2000
Zeno_3NHO
General_Fear
fduprep2018
Carbum138
tokenblackguy420
author
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
frequency


Proportion of submissions by top 20 authors in r\democrats
walter1950
backpackwayne
Gsteel11
chinmakes5
Slabraton
JohnRulez1991
Frogin21
jmsatl
Thunderbolt-P47
dakandy
ravenmarie666
featuring_42
decaturbob
bcroger3
Sip_py
garrykarpov
S629A
wagonwheelwhat
prodigy1367
alexfromsdca
author
0.00
0.02
0.04
0.06
0.08
0.10
frequency


Proportion of submissions by top 20 authors in r\PoliticalDiscussion
PsychLegalMind
Anxa
The_Egalitarian
10thunderpigs
TaylorSwiftian
Miskellaneousness
Anarcho_Humanist
strongerthenbefore20
75dollars
Visco0825
vienna95
damndirtyape
GalahadDrei
BagOnuts
Topher1999
hearsdemons
2tidderevoli
Precursor2552
Cuddlyaxe
10pumpsclassic
author
0.000
0.005
0.010
0.015
0.020
0.025
frequency


Proportion of submissions by top 20 authors in r\NeutralPolitics
crazyguzz1
huadpe
nosecohn
canekicker
Epistaxis
Totes_Police
SFepicure
redblazingdragon
uAHlOCyaPQMLorMgqrwL
RomanNumeralVI
Artful_Dodger_42
MrOaiki
PM_me_Henrika
carter1984
CQME
AFlaccoSeagulls
FlyYouFoolyCooly
JFox93
Wisex
InitiatePenguin
author
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
frequency