What's Everyone* Talking About? Exploratory Topic Modeling in Two Political Subreddits
Phil Cork

Executive Summary

The ongoing evolution of social media's role in political discourse makes analyzing conversations on these platforms a meaningful research area. This project evaluates the most popular posts in two partisan Reddit communities during the month leading up to the 2022 midterm elections. Using topic modeling, it seeks to answer a pair of pertinent research questions. First, are these partisan communities talking about the same things? Second, are they talking about them in different ways? The resulting topics and frequent terms suggest that these communities indeed discuss similar topics, but are also similar in how much the topics overlap, representing the inherently intertwined nature of politics. The findings also suggest the partisan communities discuss these topics in different ways. Some of these differences can be attributed to the structure of Reddit itself, highlighting how the platform helps shapes the conversation itself.

Introduction

While other social media platforms like Facebook and Twitter have been studied more frequently in analyzing political commentary, Reddit's forum structure provides ample opportunities for unique studies. By the platform's design, users must explicitly choose which communities (subreddits) to participate in. Thus, it is possible to study posts and comments with more precise categorization than other channels provide at scale.

Text analysis conducted using Reddit's platform often focusses on the user level to better understand and predict how individuals communicate with one another in varying political subreddits. This approach can limit the impact of findings due to users not exclusively belonging to a single group or participating in good faith. By instead analyzing differences among distinct groups at the community level, this project seeks to explore how topic modeling can illuminate the different conversations taking place among political user groups.

Analyzing Reddit Users

The forum structure of Reddit allows for a range of analytical approaches at the individual user level of analysis. Researchers have reported a comparative lack of an 'echo chamber,' (Morales et al., 2021) given that users often expand beyond their political communities, yet still engage in political conversations, even when discussing reality TV (Chen & McCabe, 2022), for instance. Further, in these non-political subreddits, conversations can be less partisan and less toxic (Rajadesingan et al., 2021). Some studies go further, exploring the comparative use of expletives across different political subreddits and the degree to which they consume and discuss news that aligns with their particular views (Ahmed et al., 2019). Even so, others highlight the challenges faced when treating users within particular communities as too similar given the multiple means of self-identification and ability to jump between multiple communities, particularly when contributing bad-faith commentary to the conversations (Alkiek, 2022).

Sentiment Analysis on Reddit

It is this particular limitation that leads to considering how higher-level, aggregated analysis may provide insights that while more general, may also be more precise by conducting research at the subreddit unit of analysis. In this way, headlines across Reddit have been studied using sentiment analysis to determine the overall positivity or negativity of the news being discussed (Khemani & Adgaonkar, 2021). This work is similar to other analysis of news headlines which uses a combination of natural language processing and support vector machines to predict the sentiment of a given headline (Chaudhary & Paulose, 2019). These approaches can be informative, but ultimately lack the contextual considerations for not just whether a headline is positive or negative, but how the language used around it compares across different communities. It also overlooks the implications of the headline itself in context of other posts, all vying for the same finite amount of the user's attention.

Text Mining & Topic Modeling on Reddit

Beyond sentiment analysis, topic modeling has also been performed across subreddits. The LDA algorithm has been used to analyze subreddit content in a number of studies, including exploring topics regarding e-cigarette use (Chen et al., 2015), eating disorders (Moessner et al., 2018), and the evolving discourse regarding same-sex marriage prior to its legalization (Hemmatian et al., 2019). In each of these cases, the authors highlight the use of machine learning to evaluate social attitudes and better understand the discourse taking place in Reddit's communities.

Data

To draw potential distinctions between the topics discussed among partisan political users, this project gathers shared through the Reddit API. With a single query, it is possible to request the 1,000 most popular posts in a given timeframe from a given subreddit. The most popular posts in the last month before the 2022 midterm elections provide a focussed point of comparison across the two subreddits of interest: r/conservative and r/democrats. Thus, the corpus includes up to 2,000 reddit post headlines, though not perfectly evenly divided by source among the two subreddits due to their respective sizes and engagement.

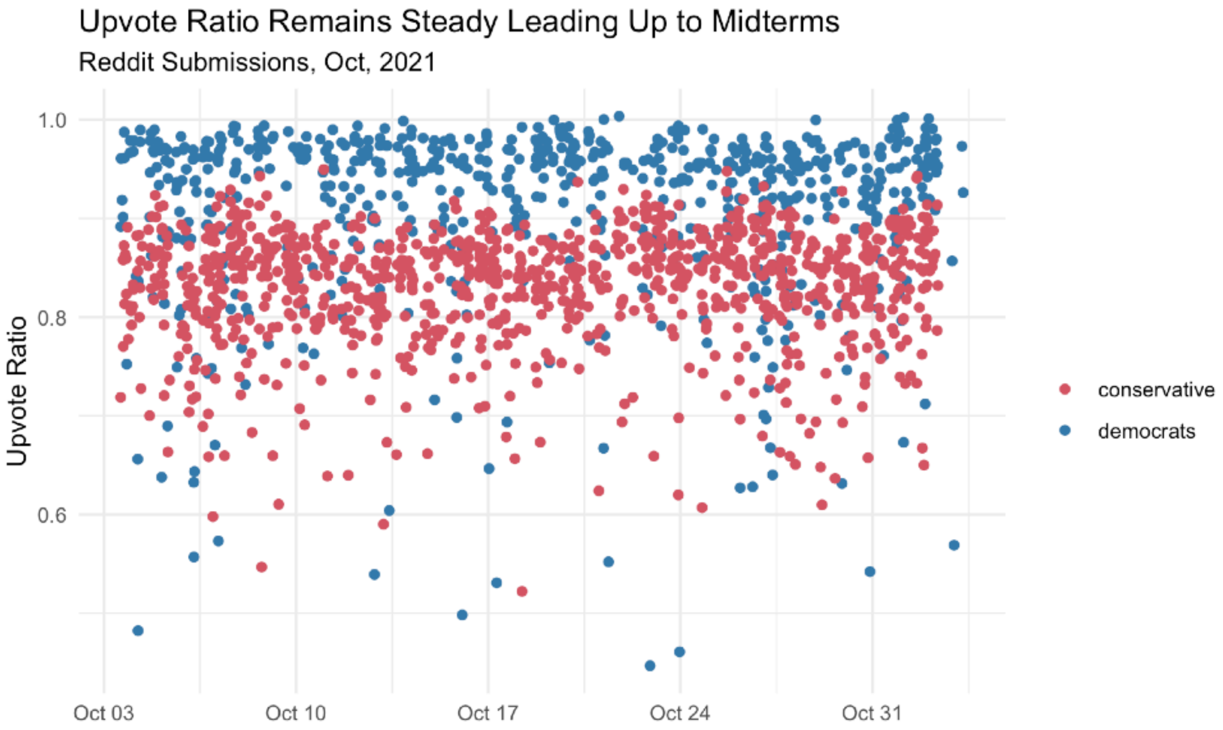
Subreddit	**Posts**	**Subscribers**	**Avg. Comments**	**Avg. Upvotes**
r/conservative	**910**	**1,009,987**	**33.3**	208.3
r/democrats	713	429,609	26.3	**215.8**

There is a clear discrepancy between the two communities both in size and engagement. Because Reddit has a left-leaning user base generally, the conservative communities tend to be more insulated and concentrated, whereas liberal communities are often more diffused across multiple subreddits. (Stocking et al., 2020) As such, it is not surprising to observe r/conservative to be twice the size of r/democrats. The latter was chosen to represent the liberal set of submissions as it appears to be the broadest, most relevant political community (see Appendix C).

Despite being more than twice as large, r/conservative only averages an additional seven comments per post than r/democrats and even sees seven fewer upvotes per post. Upvotes are a valuable metric for both engagement and general sentiment, as they are a form of endorsement either of the content, relevancy, or both, of a given submission within the subreddit.

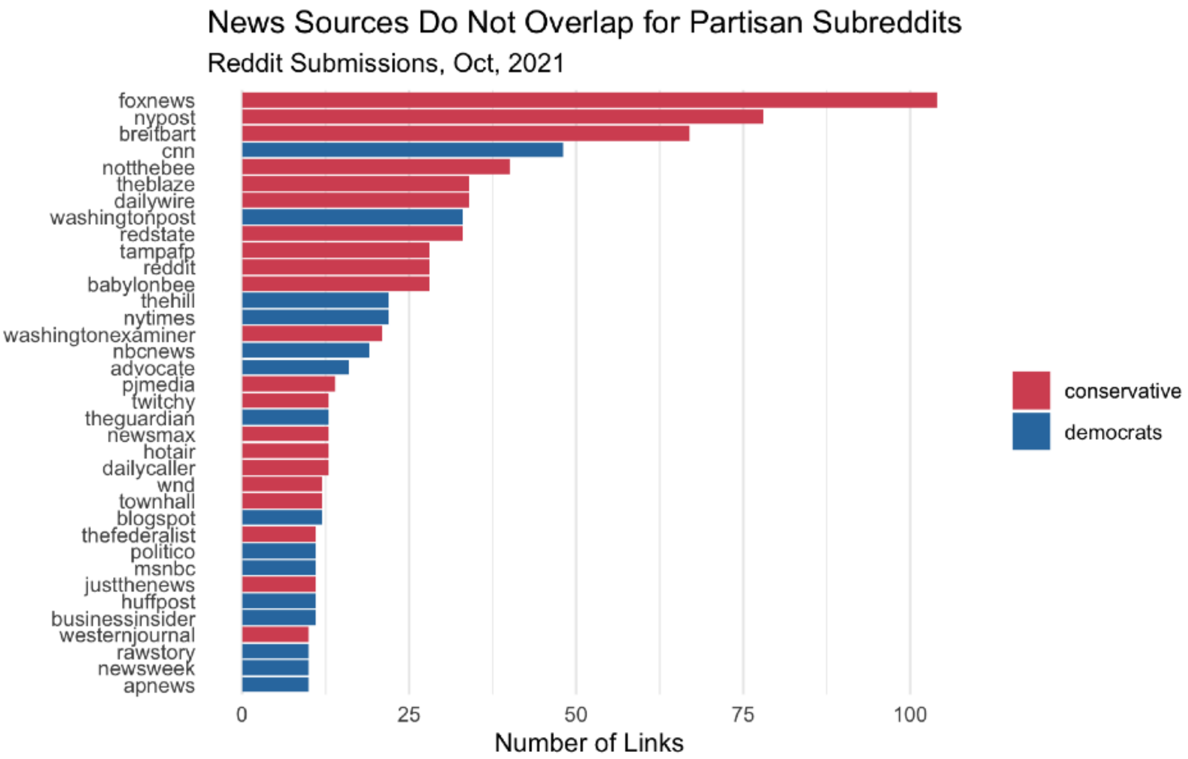
Subreddit	Upvote Ratio, Mean	Upvote Ratio, Median
r/conservative	0.78	0.8
r/democrats	**0.91**	**0.95**

When considering the upvote ratio, measuring the number of upvotes versus the number of downvotes, we observe a clear distinction between the partisan subreddits. On the whole, members of r/democrats tend to upvote content at a markedly higher rate than members of r/conservative, or conversely downvote submissions at a lesser rate. This trend could represent different levels of solidarity, enthusiasm, or consensus among members. While further analyze of this particular trend is beyond the scope of this report, it is interesting to note also the consistency with which the upvote ratio remains steady over the month leading up to an important election.



This visualization not only highlights the aforementioned distinction between communities, but the lack of change in this metric across both subreddits, which suggests that this month's worth of submissions are relatively representative and consistent. Thus, the topics that result from analysis are more likely to have meaningful implications and not be a result of noisy or dramatically atypical content being included in the dataset.

Before diving into text analysis, we also consider the website domains that each post references across both communities. By extracting the domain name from URLs included in each submission, we can compare the sources of articles shared in each community.



Notably, there is no overlap between the two communities when generic image sharing platforms are removed. This reality suggests that while topics found by the model may broadly overlap, it will not be surprising if the themes within those clusters differ, as different outlets will cover the same event through different framing techniques and with opposing sentiment.

Methodology

While the meta-data features above are helpful in illuminating differences and trends among the two communities, this project focuses on topic modeling with the collection of submission titles serving as the primary variable of interest. Each collection of posts serves as a corpus upon which the data processing methodologies and text analysis techniques are performed.

Data Gathering

The first step is described above, which is scraping the necessary text data by utilizing Reddit's API. This work is made more efficient by the PRAW package in Python, which allows for accessing the subreddits and individual submissions through a uniform process that can easily be done iteratively through a single API connection (Boe 2022). With the raw data stored locally, we then deploy straightforward data transformations on the metadata features, such as calculating summary statistics, identifying images, and categorizing posts by domain name, so as to perform the above exploratory data analysis.

Data Transformation

The second necessary technique is suite of data transformations to prepare the text. To make the content of each submission more easily interpretable to the topic modeling algorithm, it is crucial to transform the raw text data into a more standard, machine-friendly structure. The first step is to create a Bag of Words, in which each individual word is separated into a distinct term. While this removes the contextual information of the order of words, it allows us to further transform the data into a numerical structure that makes deeper analysis possible. Once we have our universe of terms, we reduce the list by removing stop words that are so common they will only be noisy data without predictive power. We further reduce words to their stems, dropping tense and plural suffixes to further shorten our list. By the same logic, bigrams were chosen to be included in the matrix of terms based on the corpus including so many political figures and proper names.

From this simplified document-term matrix, we create a TF-IDF (term frequency-inverse document frequency) matrix which essentially captures the frequency of terms in the corpus compared to the frequency in each document. This weighted metric will be highest when words are highly concentrated within a few documents and lower when words are frequent across many documents or infrequent across the entire corpus. Considering recent research evaluating text mining on short documents, these techniques are particularly important in this project since each term will carry more relative weight in determining their clustering (Albalawi et al., 2020). Given that each submission title is typically either a news article headline or brief comment on a meme, it is important to simplify and standardize the terms as much as reasonably possible, so the model can best compare each term within the documents.

Data Modeling

With our text data processed, we then implement the Latent Dirichlet Allocation (LDA) algorithm to perform topic modeling. The intuition behind this model is that each document contains some latent topics - themes users can potentially interpret and identify manually, but that are not explicitly communicated in a machine-friendly feature (Blei et al., 2003). We first deploy the above text transformations to provide the model with preprocessed input data. Then , through analyzing the comparative importance and frequency of terms within documents, the algorithm provides clusters of terms that create corpus-specific topics that users can study to better understand themes and trends among the documents as a whole. The model also assumes that multiple topics can be contained in each document, calculating a range of probabilistic outcomes for each document containing each topic.

In a similar example of LDA's implementation, topic modeling was used to better understand the various challenges parents face (Westrupp et al., 2022). This study identifies and considers forty-five unique categories of parental situations based on the results from deploying LDA across two different communities, representing mothers and fathers, respectively. This cross-community analysis allowed them to compare and contrast across similar, but distinct, self-assigned cohorts of reddit users. This process further motivates the design of this project. Given the general structure of LDA, it has the potential to be a relevant means of understanding broader context of how the two disparate communities are communicating within their respective spaces. By allowing the model to find patterns within the terms used and cluster them into topics, we can then analyze the two subreddits by the comparison between their respective clusters.

Parameter Tuning

During the data modeling stage, we also explored and selected parameters for the model. In the hyper-parameter tuning step, we eventually arrived at nine topics for r/democrats and fifteen topics for r/conservative. Because LDA does not have an inherent validation metrics in the same way that supervised learning does, this decision was left up largely to domain knowledge and framing. While the perplexity of the models were tested for different numbers of topics, results suggested that perplexity increased linearly with the number of topics regardless of how many were selected (see Appendix B). As such, it came down to examining the topics and resulting clusters of terms to determine the best way forward.

Ultimately, the driving consideration for choosing the final number of components and the thresholds for document frequency came from how well clearly distinct topics could be separated by the model. For instance, many of the submission titles in both communities touched on the Pennsylvania senate race between John Fetterman and Dr. Oz. When there were too few topics, the "Pennsylvania Senate Race" topic would include both words associated with this discussion, such as the candidates names and their talking points, but also a majority of other words that seemed to be unrelated. When too many topics were selected, however, this topic would seemingly disappear, with the key terms dispersed across other groupings. In considering a range of values for both the frequency thresholds and number of topics, we arrived at a balancing point that provides a handful of sufficiently discernible topics for further analysis.

Findings

Before addressing our primary research questions, we first present a subset of noteworthy topics and denote the most frequently assigned terms for each group (see Appendix A for full list of topics and frequent terms). In an initial study of these topics, we begin to see the rough outline of their interpretable meaning and the structure of the subreddit's conversations as a whole.

R/Democrats Topics

LDA Results, r/democrats	
Topic	**Terms**
D1	vote, 2022, walker, say, abortion, herschel walker, herschel, man, think, year
D2	obama, fetterman, oz, michigan, pete, woman, florida, debate, georgia, million
D3	trump, judge, fraud, jan, donald, donald trump, order, video, men, violence
D5	pelosi, attack, early, news, security, social, cut, home, social security, nancy

For the Democrats subreddit, Topic D1 clearly denotes terms associated with the Georgia midterm elections and the recent news regarding Herschel Walker and his stance on abortions, with other, more general terms pulled from the headlines. Topic D2 is a notably less precise, including both candidates in the Pennsylvania senate race, but also two other states. Perhaps these represent headlines discussing all prominent senate races together. It's also possible the inclusion of "woman" in this instance is related to the headlines regarding Dr. Oz's stance on abortion discussed during the "debate", another term that appears. We can consider Topic D3 the "Trump" topic for the democrats subreddit. In it we find references to the January 6th Committee ("jan"), "fraud", "violence", and the potential consequences related to these concerns in the term “judge.” Given the partisan divide over former President Trump, neither this topic nor its frequent terms is particularly surprisingly. If anything, the organization of this topic reinforces the validity of this model, even amidst other topics that are more broad in scope. Finally, Topic D5 seems to cover the news regarding the intrusion and attack on Paul Pelosi in the week leading up to the midterms. Interestingly, the topic also includes references to “social security.” This overlap is perhaps due to “attack” being used by liberal news outlets in both the literal sense to report on recent events and a more figurative sense to describe the Republican plan to reduce social security in the near future.

R/Conservative Topics

LDA Results, r/conservative	
Topic	**Terms**
C1	fetterman, senate, race, debate, senate race, oz, win, calls, john, john fetterman
C7	white, joe, house, biden, joe biden, inflation, fight, tweet, white house, getting
C8	pelosi, paul, paul pelosi, attack, cnn, kanye, west, woke, kanye west, nancy
C13	twitter, musk, elon, elon musk, fired, daily, daily wire, wire, report, takeover

In the model results for r/conservative, we observe several other noteworthy trends. Here, the Pennsylvania senate race can be comfortably distinguished from other topics in Topic C1, even more so than in the former example. In Topic C7, there is a number of references to Biden, the white house, and inflation, tying current economic conditions to the administration’s performance. Topic C8 includes many similar terms regarding the attack made against Paul Pelosi. Curiously, however, it also includes a number of terms related to Kanye West and his presence in the spotlight regarding recent anti-Semitic comments made. It seems both Kanye and the Pelosi’s, for whatever reason, are also associated with the terms “cnn” and “woke.” Finally, Topic C13 focusses exclusively on the $44 billion acquisition of Twitter by Elon Musk, his subsequent takeover of the CEO’s role, and the layoffs and firings that have taken place since. Presumably, the Daily Wire is the source of these articles, rounding out the topic’s frequent terms.

Analysis

Similarities Between Communities

With these preliminary examples in mind, we revisit our first research question - are these communities talking about the same topics?

As this subset of topics present, several public figures appear prominently in both. These overlaps include the most prominent midterm elections and the individuals associated with the most dramatic news of the week. The LDA model creating topics out of these terms in both datasets suggests that in both cases, these submissions and the terms used in their titles were in some ways distinct from other posts. It is also noteworthy that in both cases, terms related to the Pelosi family are connected to seemingly disparate terms, whether it be Kanye West or Social Security. One interpretation of this trend may be that the dramatic language surrounding these three topics are interpreted as more similar by the LDA model. Alternatively, perhaps neither subset of words has enough support to form distinct topics based on the chosen parameters and the necessary balancing of precision and interpretability.

Finally, one broad similarity between the two communities is the general lack of precise topics that divide neatly, whether it be in these examples or the full list of topics. The linearly increasing perplexity observation noted above further supports this observation. The most likely cause of these overlapping topics may be the nature of the data itself — a mix of news headlines, political commentary, and meme captions. These results highlight, regardless of partisan support, the often interconnected, messy nature of political discourse and social media dynamics.

Differences Between Communities

Having considered the similarity in topics between the two subreddits, we next address our second research question - are they talking about topics in different ways?

The first apparent contrast is the terms topically associated with each opposing party's president. In r/democrats, Trump is mentioned in the same topic as terms that potentially highlight his role in the January 6th insurrection and investigations. In r/conservative, the topic regarding Biden appears strongly connected to inflation and the economic struggles many have experienced of late. This contrast is not to say that Biden does not appear in r/democrat topics or that Trump isn't present in r/conservative topics. Both occur, but tellingly, each figure's name appears most closely associated with more comparatively neutral terms, including "vote," "democrats," "republicans," or "poll" (Topics D6 and C15, respectively, see Appendix A).

A second difference observed in these topics speaks to the broader nature of how these communities differ. The democrats community is, by design, much more focussed on elections and politics. Only a few particularly relevant news pieces are strongly represented. Meanwhile, the conservative community appears to cover a wider breadth of news and current events in addition to the political and election themes. This trend can be seen in the inclusion of public figures like Elon Musk and Kayne West in the r/conservative topics while they are entirely absent in the r/democrats topics, despite their frequent appearance in political news outlet headlines during this time.

This divergence in scope is likely due to the Reddit ecosystem as a whole. Given the left-leaning tendency of Reddit's user base, it makes sense that r/conservative would serve as more of a bastion of right-leaning individuals who can discuss politics and other topics with other likeminded individuals. In contrast, the more politically liberal individuals have more options for their political and current events discourse, with a smaller, more specific portion of it occurring in r/democrats. So while these communities do discuss similar topics with different terms and sentiment, it is the difference in the scope of topics which is most noteworthy in this instance.

Limitations & Future Considerations

With the above in mind, it is worth considering the limitations within this study. First, this project only examines two subreddits. A more comprehensive project might bundle together collections of subreddits that fall along the political spectrum and be able to draw conclusions between the left and right, but also between different subsets of each ideology's supporters. Second, this project only considers one month's worth of posts. While this limitation is due to the temporal variance inherent in political discussions, a similar project that could be more far-reaching might bring to light trends missed in this exploratory investigation.

Conclusion

The analysis of online conversations as a means to evaluate public opinion remains an ongoing, important area of research. Reddit, though less ubiquitous than other social media platforms, provides unique structural advantages to analyzing sentiment and topics across a variety of self-selected communities. By comparing the content shared across two partisan subreddits, this project aims to use topic modeling to characterize the similarities and differences between them.

The findings suggest that regardless of partisan position, the discussion of politics is often a tangled, interconnected collection of themes. These trends mirror reality in the ways that we interact with the public sector in all sorts of different ways on a daily basis. One of the major differences between what the two communities talked about highlights the influence a platform can have on the conversation itself. These comparisons, if pursued further, could lead to deeper insights regarding the role of social media in politics and the nature of online political discourse.

Bibliography

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence, 3*. https://doi.org/10.3389/frai.2020.00042

Ahmed S., Hafer J., & Lemmerich, F. (2019). A Characterization of Political Communities on Reddit. *In Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19).* Association for Computing Machinery. New York, NY, *USA.* 259–263. https://doi.org/10.1145/3342220.3343662

Alkiek, Kenan. "Political Users on Reddit." *Medium*, 31 Aug. 2022, https://medium.com/@kenan.r.alkiek/political-users-on-reddit-83926d7354c6.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. JMLR, 993–1022.

Boe, Bryce. (2022). PRAW: The Python Reddit API Wrapper (Version 7.6.0). [Source Code]. https://praw.readthedocs.io/en/stable/index.html

Chaudhary, J. & Paulose, J. (2019). Opinion mining on newspaper headlines using SVM and NLP. *International Journal of Electrical and Computer Engineering,* 9, 2152-2163. https://ijece.iaescore.com/index.php/IJECE/article/view/10797

Chen, A., & McCabe, K. T. (2022). Roses and thorns: Political talk in reality TV subreddits. *New Media & Society*. https://doi.org/10.1177/14614448221099180

Chen A, Zhu S, Conway M. (2015) What Online Communities Can Tell Us About Electronic Cigarettes and Hookah Use: A Study Using Text Mining and Visualization Techniques. *J Med Internet Res*, 17(9):e220. https://www.jmir.org/2015/9/e220

De Francisci Morales, G., Monti, C. & Starnini, M. (2021). No echo in the chambers of political interactions on Reddit. *Sci Rep,* 11, 2818. https://doi.org/10.1038/s41598-021-81531-x

Hemmatian, B., Sloman, S.J., Cohen Priva, U. et al. (2019). Think of the consequences: A decade of discourse about same-sex marriage. *Behav Res,* 51, 1565–1585. https://doi.org/10.3758/s13428-019-01215-3

Khemani, B. & Adgaonkar, A. (2021). A Review on Reddit News Headlines with NLTK tool. *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021.* http://dx.doi.org/10.2139/ssrn.3834240

Moessner, M., Feldhege, J., Wolf, M., & Bauer, S. (2018). Analyzing big data in social media: Text and network analyses of an eating disorder forum. *Int J Eat Disord,* 51, 656– 667. https://doi.org/10.1002/eat.22878

R/Conservative. (2008, Jan 25). Reddit. Retrieved October 5, 2022, from https://www.reddit.com/r/Conservative/

R/Democrats. (2008, October 4). Reddit. Retrieved October 5, 2022, from https://www.reddit.com/r/Democrats/

Rajadesingan, A., Budak, C., & Resnick, P. (2021). Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 525-536. https://ojs.aaai.org/index.php/ICWSM/article/view/18081

Reddit API. Reddit. (2006, Jan 17). Retrieved October 5, 2022 from *https://www.reddit.com/*wiki/api/

Stocking, G., Holcomb, J., & Mitchell, A. (2020, August 27). *1. Reddit news users more likely to be male, young and digital in their news preferences*. Pew Research Center's Journalism Project. https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Westrupp E.M., Greenwood C.J., Fuller-Tyszkiewicz M., Berkowitz T.S., Hagg L., et al. (2022) Text mining of Reddit posts: Using latent Dirichlet allocation to identify common parenting issues. *PLOS ONE,* 17(2): e0262529. https://doi.org/10.1371/journal.pone.0262529

Implementation Appendix

Included are additional supplemental materials developed or considered throughout the project.

Appendix A. Complete Topic Modeling Results

In addition to the example topics selected for discussion in the final report, below are the topics for each subreddit and the top terms included in each.

Topic Model Results - r/democrats

LDA Results, r/democrats	
Topic	**Terms**
D1	vote, 2022, walker, say, abortion, herschel walker, herschel, man, think, year
D2	obama, fetterman, oz, michigan, pete, woman, florida, debate, georgia, million
D3	trump, judge, fraud, jan, donald, donald trump, order, video, men, violence
D4	people, republican, gop, just, midterm, new, vote, texas, lgbtq, plan
D5	pelosi, attack, early, news, security, social, cut, home, social security, nancy
D6	biden, poll, senate, state, dems, run, vance, day, democrat, like
D7	republicans, election, campaign, right, house, crime, maga, comment, truth, box
D8	democrats, race, republicans, senate, governor, ballot, nevada, supreme, supreme court, candidate
D9	cnn, family, president, politics, ad, cnn politics, big, school, california, twitter

Similar to the subset of topics presented in the report, the additional topics provide opportunity for a couple of noteworthy observations. Topic D7 curiously includes “maga”, but it does not appear in the aforementioned “Trump” topic of D3. This divergence could point to articles that specifically mention “maga” being more tied to broad, far-right political trends, hence why the term appears with common midterm elections keywords like “election” and “campaign” and partisan terms like “republican” and the “right.” Topic D9 seems to be a “CNN” topic, which makes sense given that it is the most popular linked domain among the r/democrats posts. It is interesting to note the frequent terms of this topic do not match Topic C8 in which “cnn” appears, further highlighting how these communities talk about similar terms in different ways.

Topic Model Results - r/conservative

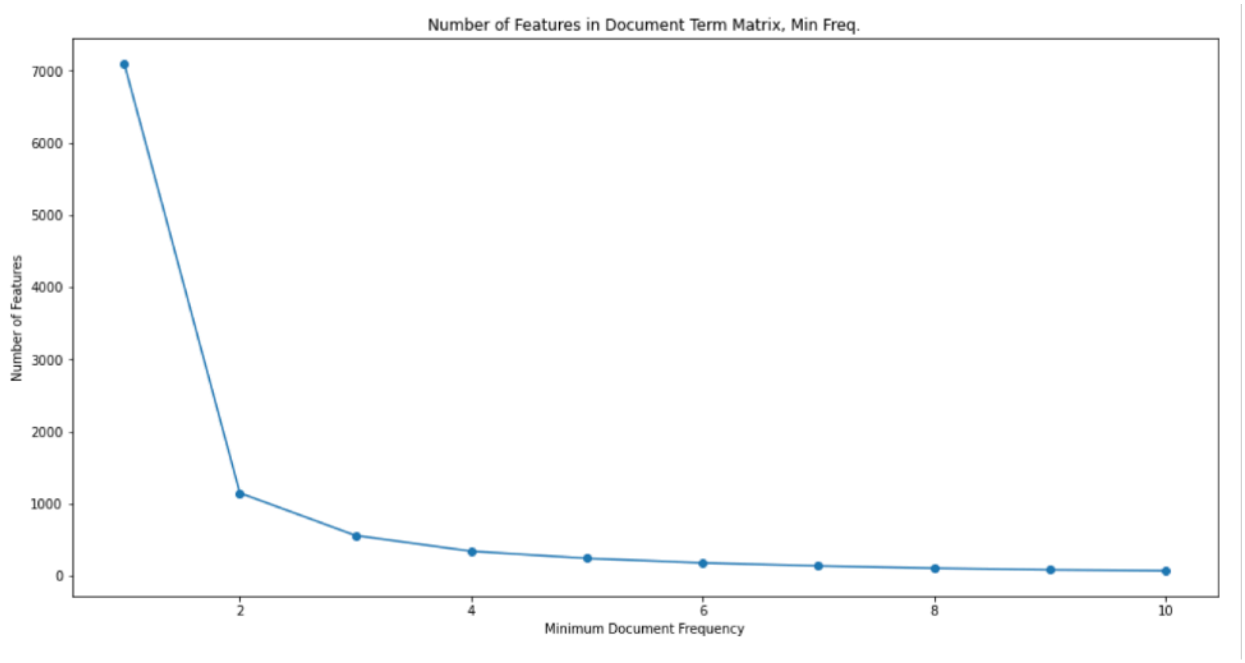
Full LDA Results, r/conservative	
Topic	**Terms**
C1	fetterman, senate, race, debate, senate race, oz, win, calls, john, john fetterman
C2	paypal, want, video, time, misinformation, like, biden, said, president, jill
C3	desantis, florida, board, school, ron desantis, ron, dem, employees, cdc, wing
C4	democrat, wa, medium, candidate, year, lake, party, kari lake, kari, arizona
C5	new, says, america, york, new york, ukraine, gop, child, say, americans
C6	biden, warns, economy, threat, oil, gabbard, war, tulsi, tulsi gabbard, american
C7	white, joe, house, biden, joe biden, inflation, fight, tweet, white house, getting
C8	pelosi, paul, paul pelosi, attack, cnn, kanye, west, woke, kanye west, nancy
C9	covid, make, claim, climate, people, say, kid, amnesty, pandemic, clinton
C10	court, election, aoc, supreme court, supreme, office, mail, jan, pennsylvania, got
C11	fbi, life, pro, hunter, account, evidence, administration, obama, sex, biden
C12	left, judge, live, ha, jail, michigan, free, law, kamala, affirmative
C13	twitter, musk, elon, elon musk, fired, daily, daily wire, wire, report, takeover
C14	border, abrams, student, know, federal, stacey abrams, stacey, loan, student loan, group
C15	democrats, republicans, vote, democracy, trump, just, children, voting, woman, win

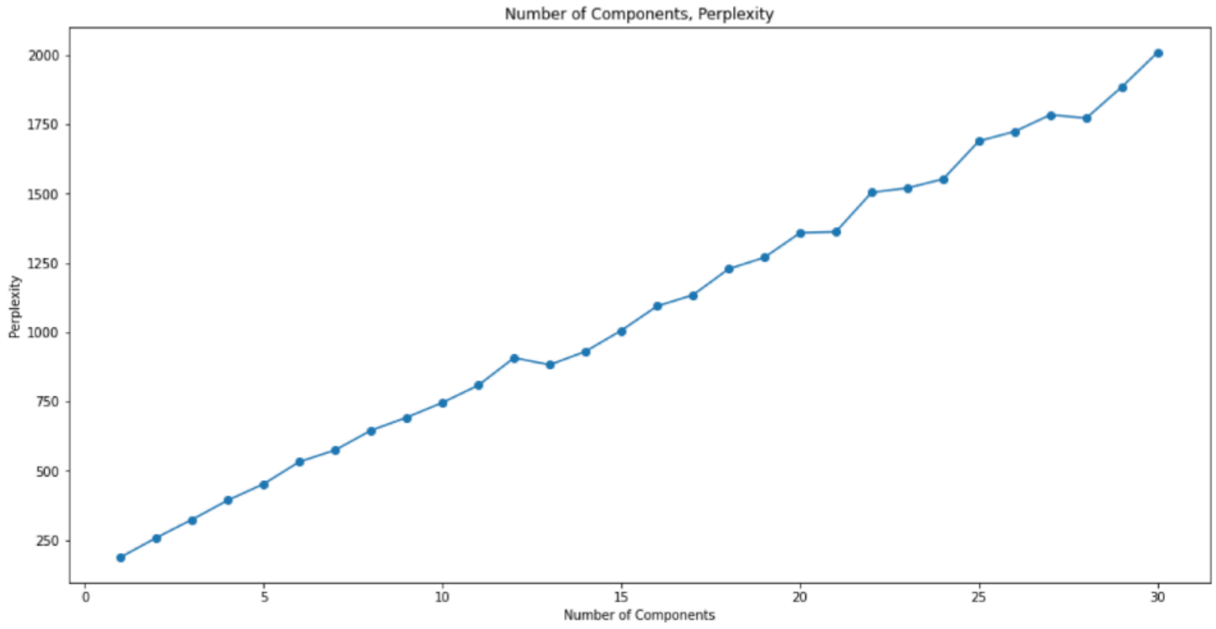
In the r/conservative topics, we observe more details regarding how this community is comparatively broader in the scope of its topics and how it differs from r/democrats when discussing similar topics. Topic C3 is clearly the "Ron Desantis" topic and connects the governor to school boards and the CDC. Topic C6, similar to C7 presented above, connects Biden to negative sentiment terms like "warns" and "threat" and economic factors like "economy" and "oil." Interestingly, it also overlaps with the timely discourse regarding Tulsi Gabbard as well. We also observe a number of public figures across these topics that do not appear in r/democrats topics, including Kari Lake, Hunter Biden, AOC, and Stacey Abrams. Broadly, these topics still present the general themes of intertwined political discourse across both communities and the wider scope of discussions in r/conservative in comparison to r/democrats.

Appendix B. Selecting LDA Hyper-parameters

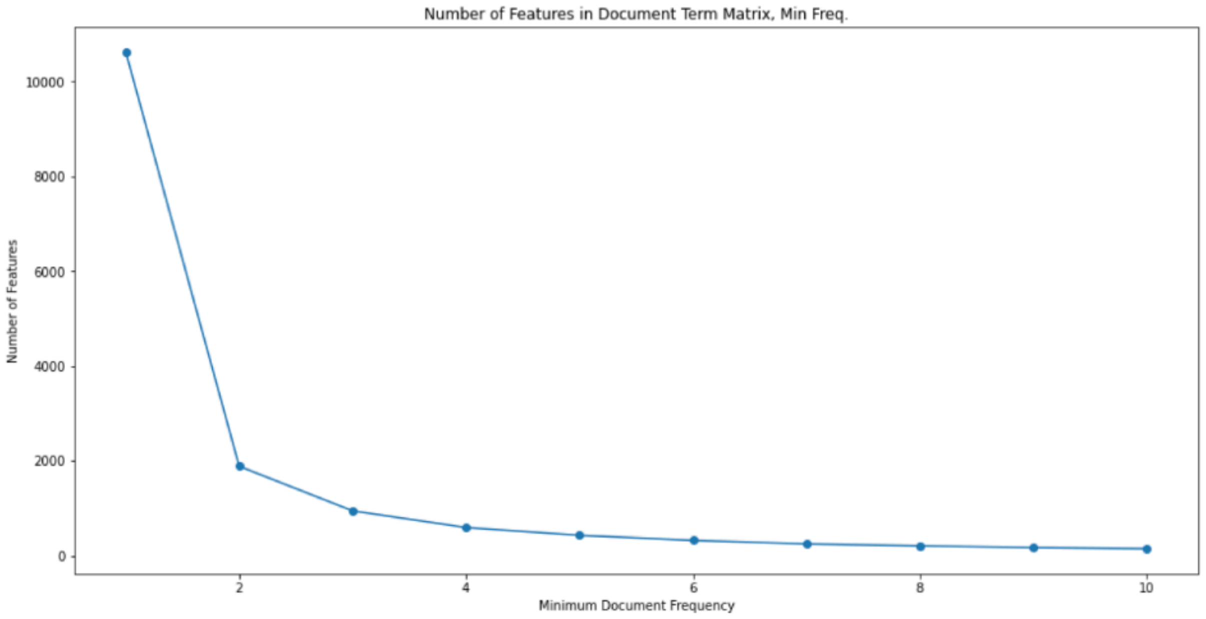
In the process of model tuning and selecting hyper-parameters, analysis was conducted for determining an optimal minimum document frequency as well as the number of components. Below are the resulting plots that led to the decision to use a minimum document frequency of six for both subreddits, as it provided an effective removal of infrequent terms, but maintained enough of the corpus to create the relatively discrete topics discussed above. Given that perplexity increased with the inclusion of each additional topic, domain knowledge was the primary factor for choosing the number of components.

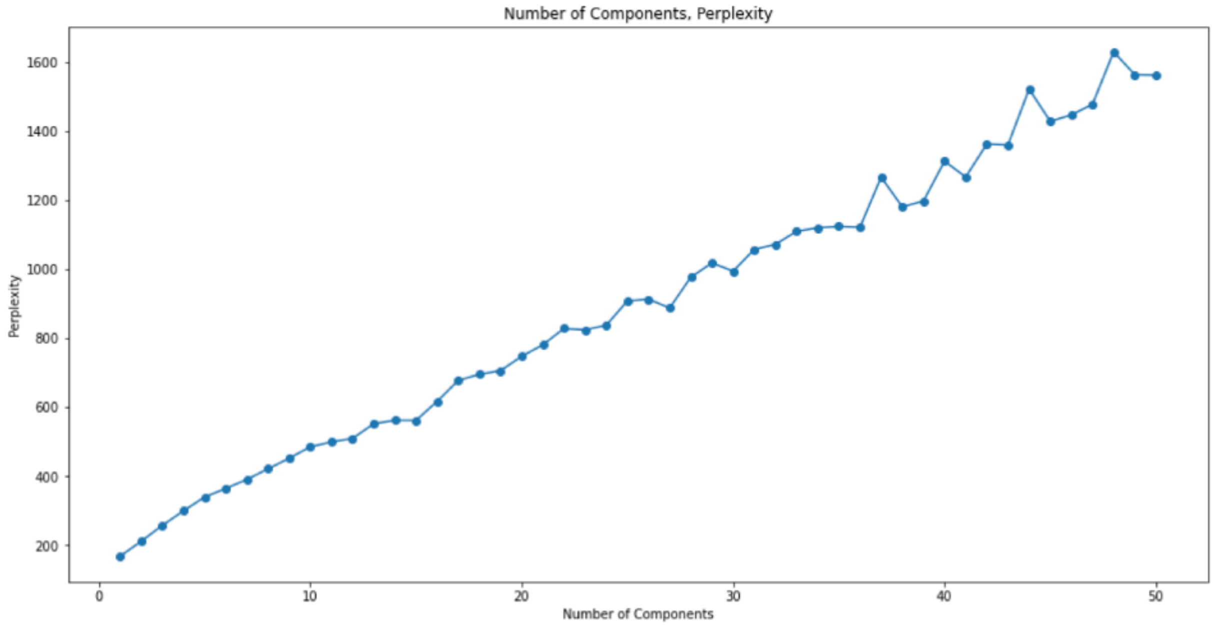
Model Tuning Parameters - r/democrats





Model Tuning Parameters - r/conservative





Appendix C. Other Subreddits Considered

In presenting this project, a valuable question was posed as to why r/democrats was chosen to represent left-leaning political discourse on Reddit when several other subreddits include similar topics and are larger and more active, making them potentially a better counterpart to the more all-encompassing r/conservative community. In the initial stages of this project, many of these subreddits, including r/news, r/uspolitics, r/progressives, and r/liberal, were considered.

Several, such as the news and politics subreddits, are moderated in such a way as to remain relatively neutral in their submission titles, even if the comments and conversation sways towards the left. As such, even if the community is more broadly representative, the documents of the corpus would not be, reducing the effectiveness of the comparisons in this project. Other specifically left-leaning subreddits are either smaller, less active, or both than r/democrats and are more specific to a particular ideology or subset of democratic discussions, meaning each submission including in the corpus might be too focussed for effective comparisons to the r/conservative conversations. As such, while comparing the more narrow and smaller r/democrats to the larger and more diverse r/conservative is not quite an equivalent representations, it proves to be the most logical point of comparison for the purpose of this endeavor.