

Reddit Reactions to the Overturn of Roe v. Wade: A Sentiment Analysis and Topic Modelling Study

By Minh Quach

Introduction:

On June 24, 2022, the Supreme Court of the United States (SCOTUS) overturned Roe v. Wade in its monumental decision in Dobbs v. Jackson Women's Health Organization. This judicial action ended the federal constitutional right to abortion that had been established nearly 50 years prior in Roe v. Wade, returning the legality of abortion prior to viability back to individual states. The Dobbs decision elicited a major reaction across the United States, especially on popular social media platforms where useful insights into population-level public opinion can be gleaned.

Given the significant personal, medical, and political implications of the overturning of Roe v. Wade, and considering Reddit's prominence as a top social networking site, this study aims to investigate the evolving discourse of the digital American public on Reddit in light of this historic judicial milestone. By analyzing changes in the sentiment and themes of discourse on Reddit in the months immediately before and after the Dobbs decision, this research seeks to shed light on how this event impacted public opinion as expressed in online social spaces.

For a brief review, we noticed a growing body of research which has begun to leverage social media data to examine public discourse and attitudes related to abortion. Mane et al. (2022), Ujah et al. (2023), and Valdez et al. (2023), for example, used Twitter data to examine temporal, geographical, ethno-racial, and sentiment patterns in the public's reaction post-Dobbs. Valdez et al. (2024) also looked at reproductive rights discourse on Reddit, but only before our key events in 2022, and thus in absence of the political effect of interest. More broadly, we have papers such as Doan et al. (2022) which looked at the online reactions to another event of similar nature that was Georgia's House Bill 481 (LIFE Act) in 2019, or Molek-Kozakowska and Wanke (2019), in the context of abortion law debates on Polish social media.

This study builds off of the reviewed literature in several key ways. Firstly, it focuses specifically on Reddit, which compared to Twitter, remains understudied as a data source for examining public sentiments and specifically abortion attitudes. While lagging behind Twitter in terms of instantaneity and viral short-form dissemination of information, Reddit's unique features come from its pseudonymity, its longer, more diverse formats, and its focus on forums and in-depth exchanges of information (Priya et al., 2019). Secondly, the study combines a mix of methodological recommendations from the literature, namely topic modelling sentiment analysis, and extends on the applications of these tools in this field. Thirdly, the study seeks to confirm, or challenge, the common findings by existing literature on the diverse thematic characteristics of these forums, on the impacts of the key events, and on the moods and concerns of these forums over time.

Based on our literature review, we have 4 research hypotheses as stated below:

Hypothesis #1:

Following the leak of the Supreme Court opinion and the overturning of Roe v Wade would cause an increase in activities overall on the most prominent reproductive rights forums on Reddit, and especially among posts of political nature.

Hypothesis #2:

Politics would be a minor topic within these forums, before the Overturn increased its thematic prominence.

Hypothesis #3:

Compared to the pre-period, sentiment of posts and comments about reproductive rights on Reddit would be more negative overall following the overturning of Roe v Wade.

Hypothesis #4:

The negative effect of the Overturn on the sentiment will decrease and dissipate eventually.

Data and Tools:

The data comes from the Pushshift Archive which has collected Reddit data and made it available to researchers since 2015. Previously, the data used to be publicly available until Reddit made changes to its API terms of service in 2023. The data is stored on a private Amazon Web Services (AWS) cloud storage space, and available thanks to the courtesy of Professor Amit Arora from the course PPOL-5206.

There are two component datasets, concerning submissions and comments on the media platform. Each of the datasets contains information regarding the subreddit the submission/comment belongs to, its textual content, author name, timestamp, engagement metrics, controversiality, and more.

For this project, we picked the subreddits “r/abortion” and “r/birthcontrol”, both are the de-facto “support group” on Reddit for the respective areas of reproductive rights. The rules of both forum demand “respect”, prioritize “factuality” and “accuracy”. The data spans March 2022 to August 2022, 3 months before to 3 months after the overturning. This timeframe is particularly significant as it encompasses two pivotal events: (1) the unprecedented leak by Politico of the draft majority opinion in early May, which ignited a firestorm of speculation and debate, and (2) the official ruling by SCOTUS (Supreme Court of United States) on June 24th that upended the reproductive rights that has existed for nearly five decades in America. Centering our data around these key inflection points, with a balanced buffer of two months before the leak and two months following the decision, facilitates our research plan to capture a comprehensive picture of how the conversation unfolded.

After retrieval, our dataset contains 24,653 posts and 139,822 comments.

For the programmatic tools, the project was written in Python, using a serverless Spark cluster on AWS for the majority of the data retrieving, wrangling, and modelling.

For the preprocessing steps, we took advantage of the built-in functions of the Jonsnowlabs' sparkNLP library, including document assembler (transforming input data into Spark NLP annotated documents), tokenizer (splitting the text into individual words, also known as tokens), normalizer (performing various text normalization techniques, such as converting text to lowercase, removing special characters, and replacing abbreviations with their full forms), lemmatizer (reducing each word to its base form - lemma, which can improve the consistency of analysis by grouping together related words), and stopword cleaner (removing common stop words from the text, such as "the", "and", "a", etc., in addition to a custom list of more topic general words, which are generally not helpful in some parts of our analysis).

Analysis Methods:

To address hypothesis #1, we use simple techniques of exploratory analysis such as frequency count, TF-IDF, and plotting engagement metrics across time to identify and measure the supposed increase in user activities on the selected subreddits after the key events.

To address hypothesis #2, we proceed to use two different unsupervised topic modelling methods: Latent Dirichlet Allocation (LDA) which uses probability distributions and Gibbs sampling, and BERTopic (a neural network approach that leverages pretrained embeddings from the Bidirectional Encoder Representations from Transformers (BERT) and subsequent clustering and weight filtering to improve topic interpretation). This multiple-model approach is recommended by several of the reviewed literature.

Specifically, Mane et al. (2022) used LDA to examine temporal, geographic and sentiment patterns in tweets related to Roe v. Wade and abortion bans collected before and after the Supreme Court opinion leak and official ruling. They found LDA useful for identifying prevalent conversation topics over time and across states. Similarly, Valdez et al. (2024) employed both LDA and BERTopic models to compare semantic and content differences in Reddit posts from two subreddits - one focused on abortion social support and the other on abortion debate. Here, the authors argued that BERTopic generated more nuanced and contextualized topics compared to LDA. More broadly, LDA and BERTopic are arguably the most prominent topic modelling methods currently used by social researchers, thanks to their robustness and wide applications.

To address hypotheses #3 and #4 that concern with mood approximation, we use two different pretrained models for sentiment analysis: SentimentDL, a multi-class model inherent to the SparkNLP library which built upon millions of IMDB reviews and their corresponding GloVe embeddings, and Valence Aware Dictionary and sEntiment Reasoner (VADER) which is an increasingly common "lexicon and rule-based feeling analysis instrument that is explicitly sensitive to suppositions communicated in web-based media". While the former is more easily implemented within our Spark environment, the latter approach produces a more nuanced, and precise numerical approximation of emotional affect (continuous values between -1 and 1 instead of binary

classification of ‘positive’ and ‘negative’). This makes VADER more appropriate for further sentiment analysis across more granular subsets of data and more specific time periods.

To specify the linguistic dictionaries that comprehensively cover the different aspects of the discourse (timeline, legality, accessibility, political, etc.), we qualitatively built several lists of words which drew from Mane et al. (2022)’s keyword lists, Valdez et al. (2024)’s identified key terms, and Sable and Galambos (2006)’s key concepts. We also made the necessary modifications to account for modern slang and more recent political concepts and identities that are pertinent to the main discussion.

Results:

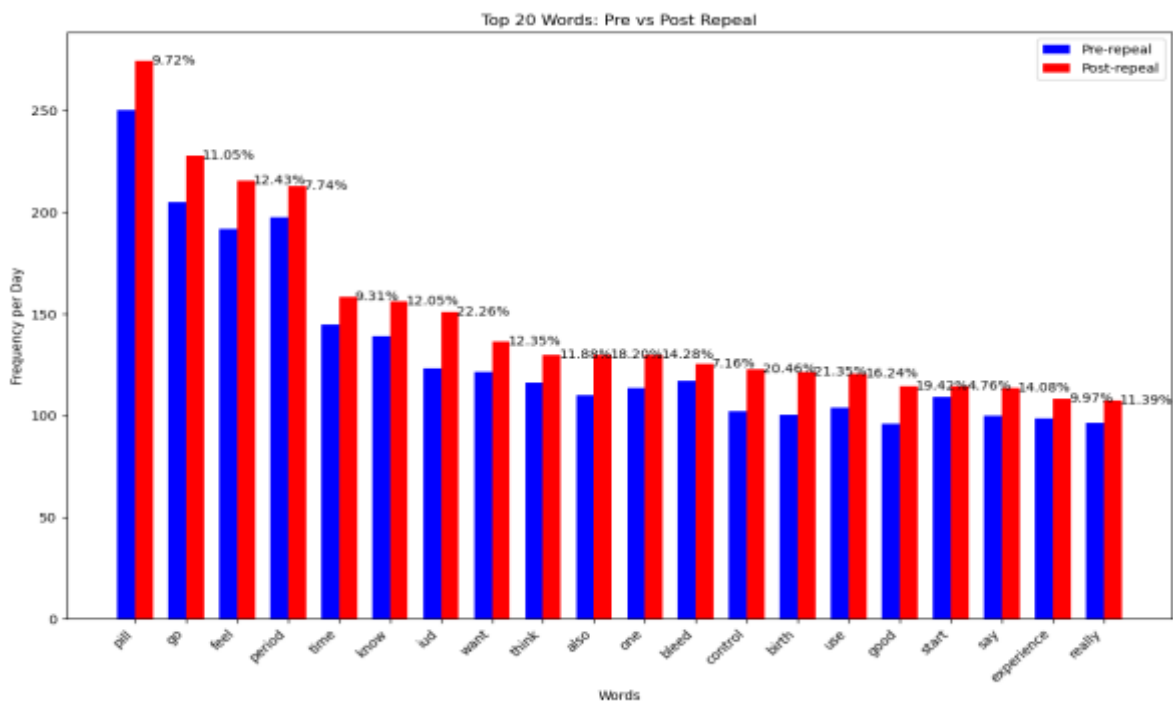


Figure 1

Looking at the top common words and bigrams in these forums (Figure 1 and 2), we noticed a universal pattern: Daily activities post-Dobbs increased, ranging from 10% to 20%. Furthermore, if we qualitatively examine the most common words and bigrams, we notice many common domain-specific terms, many topic-general words, but none of them are political. This is consistent with our initial hypotheses.

Looking at the engagement metric of post and comment scores (Figure 3), and especially the lines of best fit where the trends are consolidated, we observe a notable increase in average scores right after the two main events. Such increase then dissipated as time went on, which is consistent with our initial arguments of shifting news cycles and incidental news engagement. Looking specifically at the political items found in the corpus (Figure 4), we observe a very similar pattern. Right after the corpus leak by Politico, mentions of the most popular political terms (Judge, Court, politic, SCOTUS, etc.) all significantly increased. These effects waned down for the few weeks that followed, before resurging to another height of mentions after the official ruling in late June 2022. This effect also eventually dissipated, further supporting our hypothesis #1.

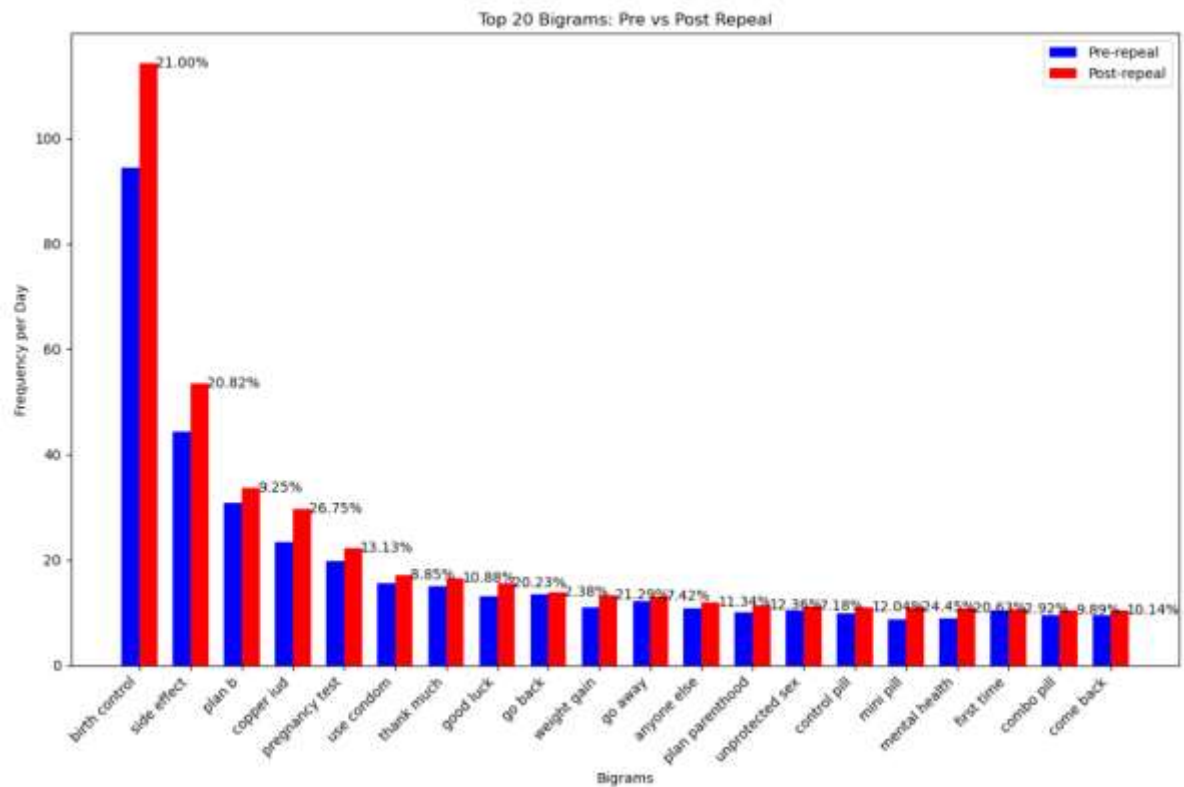


Figure 2

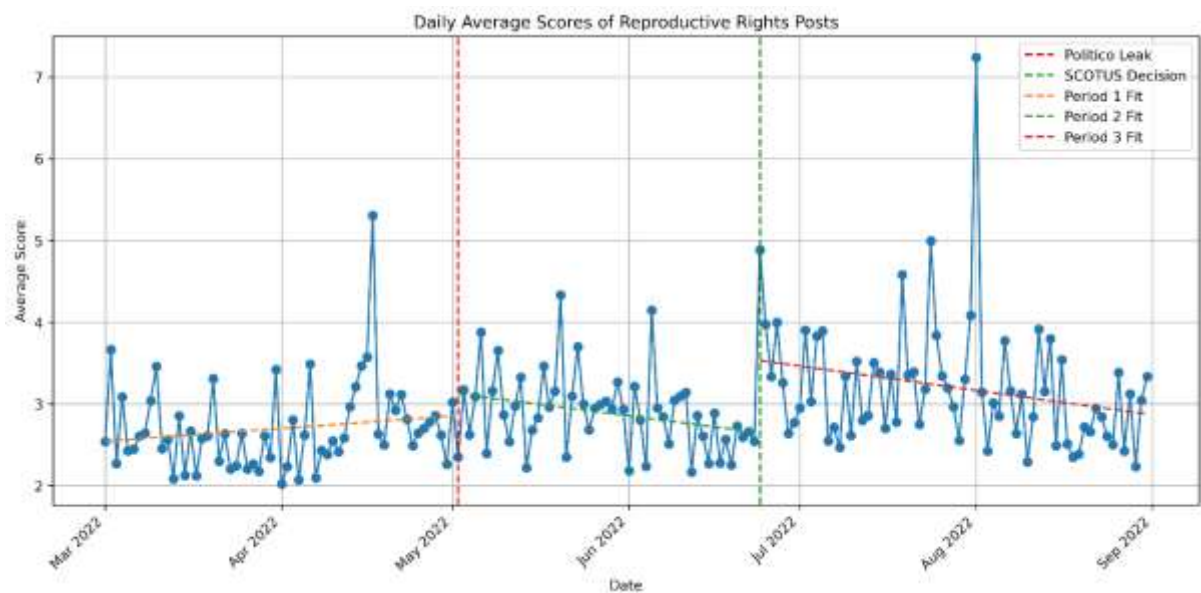


Figure 3

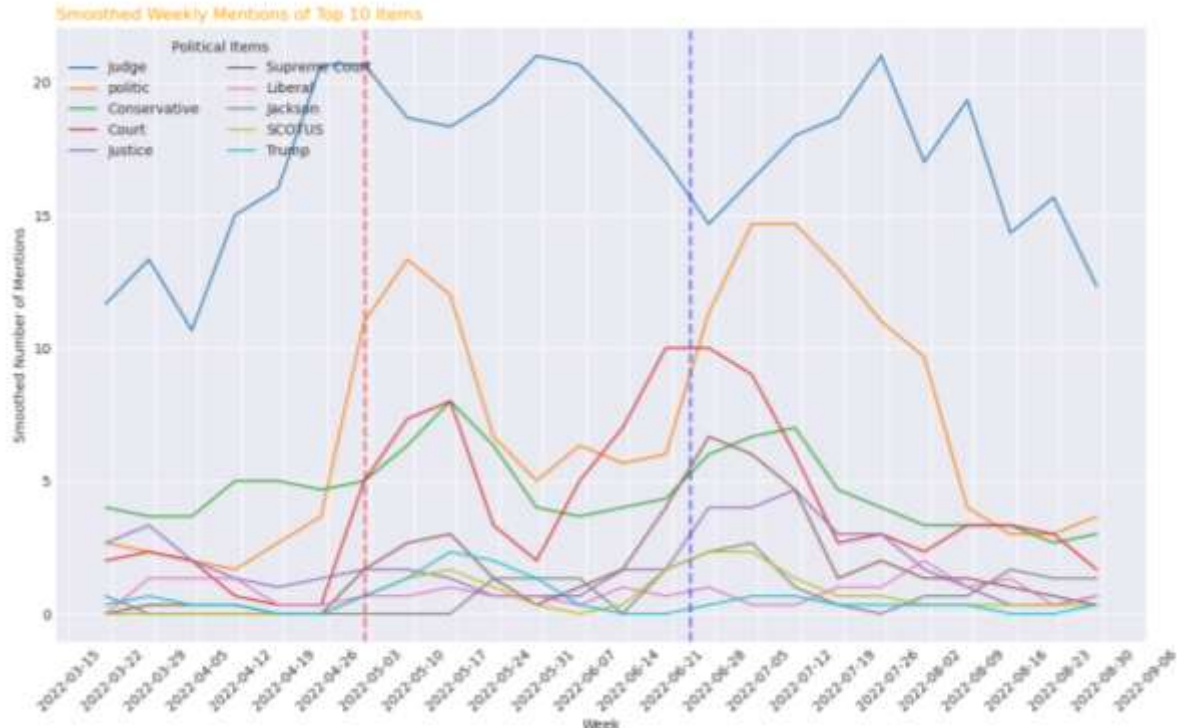


Figure 4

Moving on to topic modelling, our LDA approach produced the results below (Table 1). Parameter tuning was carried out and the topic coherence scores were used as the main evaluation metric. Eventually, we arrived at 10 topics and an average coherence score of 0.6 across iterations. These identified themes are consistent with what Valdez et al. (2024) has found in r/abortion. The rest of the topics below also do not contain any political shade, which is consistent with our initial thinking in hypothesis #2, that politics took up a very small parts of the discourse.

topic number	topic words	Interpretations
1	effect, side, iud, pill, plan, weight, b, bc, gain, hormonal	Birth Control Side Effects
2	feel, want, know, go, abortion, think, time, baby, life, right	Abortion Decision
3	pill, use, control, birth, condom, sex, time, period, start, pregnancy	Birth Control - General
4	test, pregnancy, go, body, one, say, hormone, period, negative, positive	Prenancy and Test
5	iud, doctor, copper, birth, control, want, one, say, use, ask	IUD and Doctor Interaction
6	control, go, birth, pill, period, start, feel, experience, back, effect	Birth Control Pill
7	abortion, go, clinic, pregnancy, call, know, appointment, state, see, procedure	Abortion Procedure and Access
8	thank, good, much, go, help, know, well, hope, need, really	General Support Language
9	period, bleed, start, pill, spot, first, last, normal, stop, time	Menstruation
10	cramp, pain, feel, period, go, bad, bleed, experience, iud, insertion	IUD Experience and Issues

Table 1

Using the BERTopic approach, we come up with a very similar coherence scores of the newly identified topics (0.60), showing little improvement from the LDA results in this regard. Examining the topics qualitatively, we also see major overlaps between the results produced by the two approaches. More specifically, abortion decision is again found to be one of the most important and fundamental themes of the subreddits in both sets of results. Furthermore, abortion procedure, legality, and access is also among the top 5 topics identified, indicated by the red circle (which is visually engulfed in the broader theme of topic 1) (Figure 5).

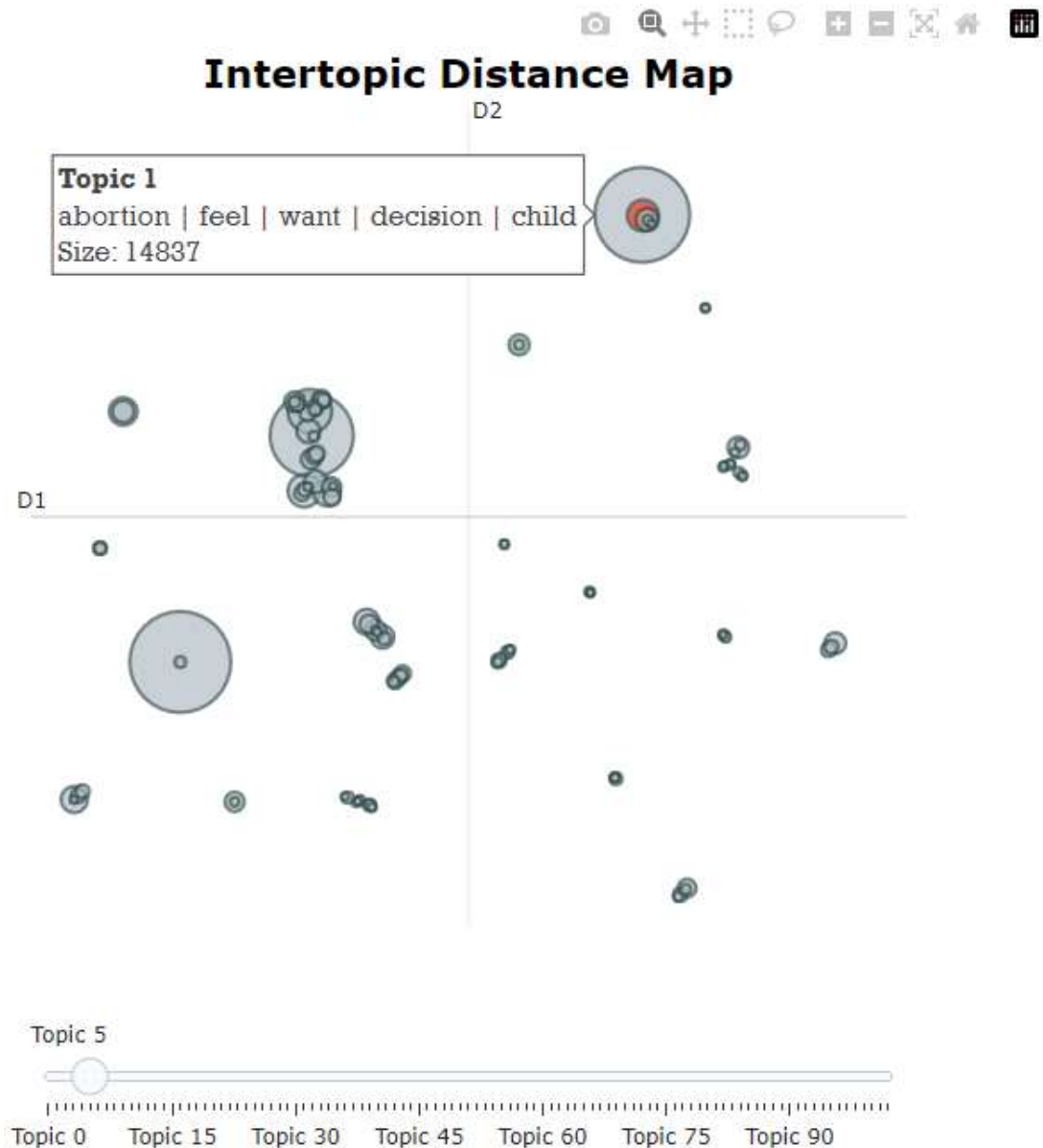


Figure 5

Employing the same BERTopic pipeline for the 3 subset of our datasets divided by the two main events, we still did not notice any politically charged cluster of posts broke into the top 10, even after the corpus leak, and after the overturn. "Abortion decision" however remained the top cluster throughout, and its size increased by 54% after the overturn, indicating significant increase in concern and/or inquiry over this major aspect. This

challenges our hypothesis #2, that politics would rise in terms of thematic prominence post-Dobbs.

Looking at the sentiment across time of the Reddit forums of interest (Figure 6), we notice two drops in positivity after the two main events. However, given these variations are small relative to the full range of the sentiment scores (a 0.03 drop in sentiment score accounts for 1.5% of the full range from -1 to 1). This is possible due to the inherent nature of the subreddits which shun political discussions and prioritize fact-based communications with a small dose of positive encouragement. Together with our previous finding about the thematic importance of politics over time on these forums, we may argue that the Overturn did affect the discourse overall, though such effect did not channel into any conspicuous political venues, which are not encouraged on these forums.

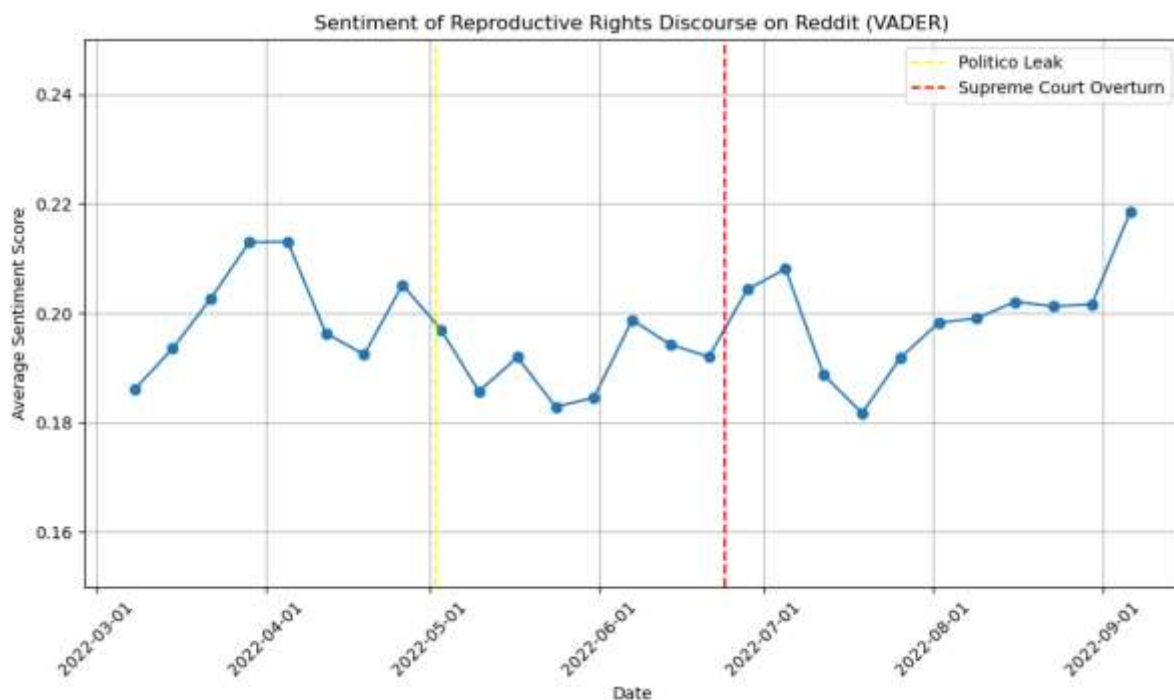


Figure 6

Using the same approach but segregating the texts into different concerns (Figure 7), we observe a much clearer picture of the political effect of interest. After the two main events, most concerns did not observe any significant drop in sentiment despite our initial hypothesis. However, posts of political concerns exhibited drastic drops in morale, especially after the corpus was leaked to the public and the prospect of abortion bans started looming over a palpably near horizon. Given how political posts had traditionally taken up a very small corner of these forums (mainly due to community rules and norms), we may attribute the previously observed drop in overall sentiment to these small number of items, which did exhibit the political effect of inquiry. This speaks to the validity of our hypothesis #3. Furthermore, even among the political posts, the souring of the mood also lasted for approximately two weeks before the sentiment started to pick up again subsequently. This corroborates our hypothesis #4, arguing that the observable negative effects of these political events on Reddit diminished over time.

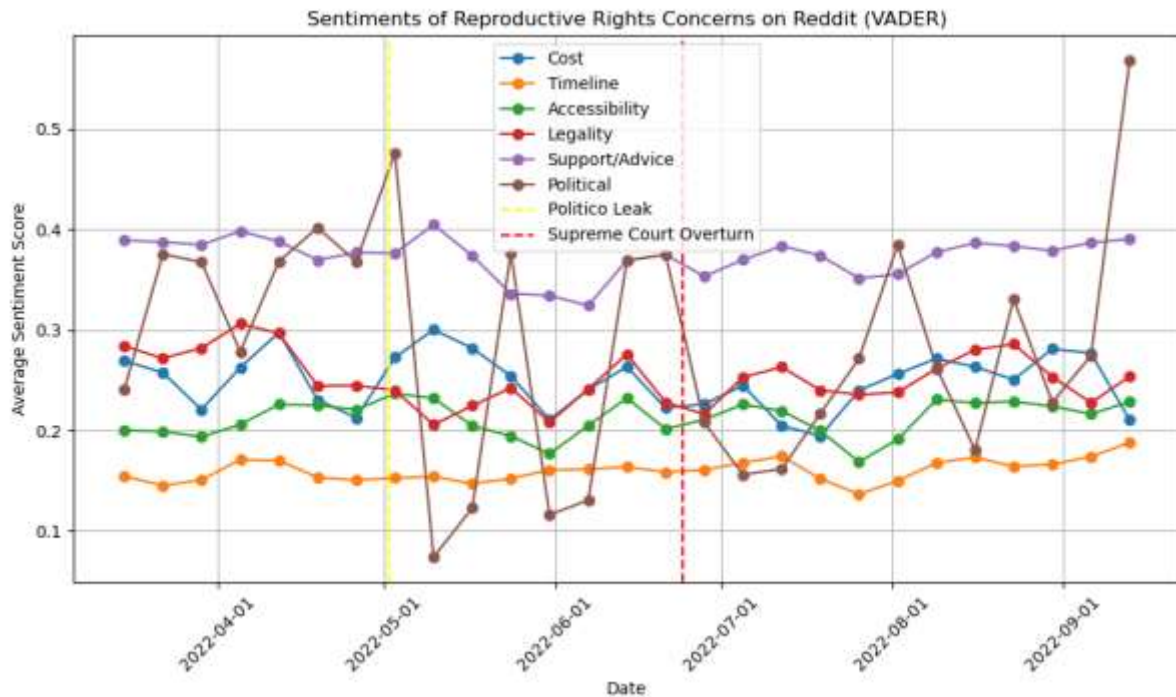


Figure 7

Discussion:

The main findings of this study provide valuable insights into the evolving discourse surrounding reproductive rights on Reddit in the context of the historic *Dobbs v. Jackson Women's Health Organization* decision. The results support most of our initial hypotheses, revealing increased user activity and engagement, particularly in posts of political nature, following the Supreme Court opinion leak and the official overturning of *Roe v. Wade*. Topic modeling using LDA and BERTopic identified key themes consistent with the reviewed literature, such as abortion decision-making, procedure, legality, and access. Notably, political discussions were found to be a minor topic within these forums before the *Dobbs* decision, which did not subsequently increase their thematic prominence despite our initial thinking. This can be mainly attributed to the disinclination to engage in politics, that is particular the selected forums.

Sentiment analysis using VADER demonstrated an overall drop in positivity after the two main events, albeit small relative to the full range of sentiment scores. This modest decline can be attributed to the nature of the subreddits, which prioritize fact-based discussions and positive support while discouraging political debates. However, a more nuanced analysis segregating texts into different concerns revealed a significant drop in sentiment specifically among posts with political content, corroborating our hypothesis that the overturning of *Roe v. Wade* would lead to more negative sentiment in related Reddit discussions. Furthermore, the observed negative effect on sentiment dissipated approximately two weeks after each event, supporting our argument that the impact of these political developments on Reddit sentiment would diminish over time.

The contributions of this study are threefold. First, it focuses on Reddit, an understudied platform compared to Twitter, leveraging its unique features such as pseudonymity, longer and more diverse post formats, and dedicated forums for in-depth information exchange. Second, the study combines methodological recommendations from the literature, employing topic modeling and sentiment analysis to provide a comprehensive understanding of the thematic and emotional landscape of reproductive rights discourse on Reddit. Third, the findings confirm and extend the conclusions of existing literature regarding the diverse thematic characteristics of these forums, the impact of key events on discussion topics and sentiment, and the temporal dynamics of public opinion in response to major political developments.

To further advance this research, several directions can be explored. Expanding the dataset to include a longer time frame and additional relevant subreddits could provide a more comprehensive view of the evolving discourse on reproductive rights. Incorporating demographic information about Reddit users, if available, could enable an analysis of how opinions and sentiment vary across different subpopulations, and across states with varied jurisdictions and legal outlooks over reproductive rights. Applying more advanced natural language processing techniques, such as aspect-based sentiment analysis or emotion detection, could offer a more granular understanding of the specific emotions and concerns expressed in these discussions. Additionally, a comparative analysis of the discourse on Reddit with that on other social media platforms, such as Twitter or Facebook, could shed light on how platform-specific characteristics shape public opinion and engagement with sensitive political topics.

References:

- Mane, H., Yue, X., Yu, W., Doig, A. C., Wei, H., Delcid, N., Harris, A. G., Nguyen, T. T., & Nguyen, Q. C. (2022). Examination of the Public's Reaction on Twitter to the Over-Turning of Roe v Wade and Abortion Bans. *Healthcare (Basel, Switzerland)*, 10(12), 2390. <https://doi.org/10.3390/healthcare10122390>
- Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media*, 9, 17–29. <https://doi.org/10.1016/j.osnem.2018.11.001>
- Sable, M. R., & Galambos, C. M. (2006). Prevention and Women's Reproductive Health: A Matter of Concern. *Health & Social Work*, 31(3), 163–166. <https://doi.org/10.1093/hsw/31.3.163>
- Ujah, O. I., Olaore, P., Nnorom, O. C., Ogbu, C. E., & Kirby, R. S. (2023). Examining ethno-racial attitudes of the public in Twitter discourses related to the United States Supreme Court Dobbs vs. Jackson Women's Health Organization ruling: A machine learning approach. *Frontiers in Global Women's Health*, 4, 1149441–1149441. <https://doi.org/10.3389/fgwh.2023.1149441>
- Valdez, D., Mena-Meléndez, L., Crawford, B. L., & Jozkowski, K. N. (2024). Analyzing Reddit Forums Specific to Abortion That Yield Diverse Dialogues Pertaining to Medical Information Seeking and Personal Worldviews: Data Mining and Natural Language Processing Comparative Study. *Journal of Medical Internet Research*, 26(1), e47408–e47408. <https://doi.org/10.2196/47408>
- Valdez, D., Mena-Meléndez, L., Crawford, B. L., Arvind, A., & Jozkowski, K. N. (2023). Online Social Media Reactions to the Overturn of Roe v. Wade: Public Health Implications and Policy Insights. *Sexuality Research & Social Policy*. <https://doi.org/10.1007/s13178-023-00892-2>