Medini Chopra
Professor Debayan Gupta
The New Geography of the Information Age
13th May, 2023

Baby, Can't You See I'm Commenting?
Comparing **Toxic**ity in Gendered Reddit Communities

Abstract

Reddit, a popular social media platform, has built its reputation as an 'unsafe' space due to its unique interface involving user anonymity and community moderators; and has led to the emergence of movements such as the Manosphere. Most of this hate is directed towards women, making it a *toxic* and unwelcome place for discourse, discussion, and finding like-minded people. In this paper, I attempt to examine the toxicity of comments on submissions from male-dominated and female-dominate subreddits which fall under the overarching theme of STEM. I scraped relevant subreddits including r/askWomen, r/askMen, r/womenEngineers, and r/askEngineers. My methods for content analysis include measuring toxicity using the Perspective API and topic modeling using Latent Dirichlet Allocation (LDA).

Keywords: gendered subreddits, toxicity, topic modeling, online communities, inclusivity, Reddit analysis, women in stem

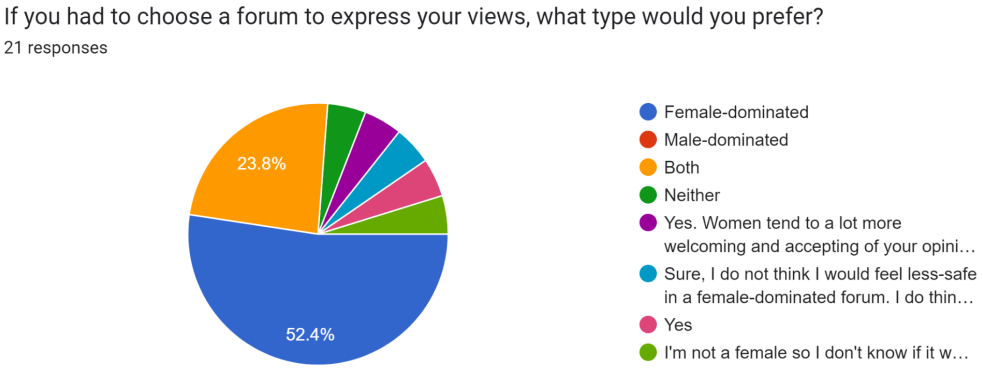
1. Introduction

Toxicity is the presence of harmful, aggressive, offensive, or abusive language in online communication, and it manifests in the form of insults, threats, and sexual harassment [5]. Reddit especially, is infamous for being a hub of such behavior and toxicity, which has lasting effects on the viewers in the form of self-censoring, degraded mental health, and discouragement to stay on the application. Reddit, the self-proclaimed 'front page of the internet', has the unique functionality of upvoting and downvoting submissions, where each submission gives birth to whole new sub-conversations. Its biggest downfall, however, is the structure of anonymity and moderation for subreddits. Due to Reddit's unique interface, it was chosen as the platform to study the relationship between toxicity, gendered language, and gendered forums.

Building on toxicity in male-dominated spheres, I was interested in examining whether gendered communities (which exist within larger discussion forums); are toxic toward the community they are intended for. Furthermore, I'm interested in analyzing what these conversations constitute. The practice of asking questions and getting guidance at the *start* of your career is crucial to making any professional development, and the environment you grow in can determine your overall success. Due to the prevalence of women in STEM, who constitute minorities, it is important to analyze how these conversations differ in male and female-dominated forums, mainly how users react and reply to questions, comments, and discussions.

This study is important because it can shed light on the disparities in the treatment, representation, and overall experiences of individuals based on their gender within these online communities. Further, it can give insight into the effectiveness of community moderation strategies and policies. Lastly, the most obvious would be a case study of user experience on Reddit and how to improve the same.

On conducting a user survey on the usage of online forums for expression, and how safe the users would feel on certain types of forums, I found that people would prefer voicing their opinions in female-dominated forums rather than male.



The follow-up question to the previous chart asked why they would choose such a platform, and then general consensus was that it is “safer”, “rational” and “welcoming”. This forms the basis of my inquiry. I identified relevant forums on Reddit: r/askEngineers, r/WomenEngineers, r/askWomen, and r/askMen and I address the following questions:

RQ1: What is the relationship between the toxicity of comments in women-oriented and men-oriented subreddits?
RQ2: What are the topics discussed in these comments?

To analyze the first question, I employed Google’s Perspective API to measure the toxicity of individual comments. For the second question, I sampled the highest scored comments and ran them through an LDA topic model to map out the dominant topics in these discussions.

2. Related Work

Harassment, hate speech, and cyberbullying have been widely detected on the internet using various natural language processing and machine learning techniques. These are focused on social media sites like Twitter, Meta, and Reddit, and analyze the structure of toxic language [1], identify triggers of toxicity, and examine the effects of moderation on propagating such toxicity [6].

Toxicity on Reddit is also attributed to the Manosphere movement, which is a combination of many sub-communities that are interested in the crisis of masculinity. These include “Men’s Rights Activists (MRAs), Men Going Their Own Way (MGTOW), and Involuntary Celibates (Incels)”, and studies have shown the toxicity of these subreddits has been increasing in the past few years [9][10]. The increasing fear around toxic or oppressive masculinity, misogynistic content, and violent user behavior

manifests toward women and the feminist movement [8][11]. Many men's rights movements actually seek to deprive women of resources and bar them from male-dominated domains, all while masquerading as advocates for men's liberation [13].

When it comes to women's perspective on these subreddits, studies have analyzed the dominant topics that are discussed in fields such as STEM [12][14]. This is because of how underrepresented women are and how important it is for them to find a supportive community when they are in the initial stages of breaking into the field. These specific subreddits for women in STEM/engineering branch out from the larger male-dominated communities due to how unwelcoming and toxic they are toward women as shown above. While there has been research on toxicity in male-dominated spheres toward women, there has not been any comparative analysis of the toxicity of male-dominated versus female-dominated forums. To extend this previous work, my study contributes as a novel use case of exploring and comparing the toxicity of Reddit in the context of gendered communities.

3. Data and Methodology

a. Data

i. A preliminary glance over https://www.reddit.com/subreddits/ gave a list of subreddits that matched the relevant keywords of "stem", "engineers", "tech", "women", and "men". It followed a quick search on how many subscribers and comments on submissions existed in those subreddits that were available. I had chosen subreddits with the highest number of subscribers coupled with the highest keyword relevance. This is because those who have subscribed are personally invested in that content, and are more likely to view, upvote, and comment on submissions.

ii. I chose comments instead of submissions because the submissions themselves hardly have any text. The real conversations happen in the comments because it's the most direct way of interacting with others, and Reddit allows for nested comments which have the scope for many side conversations that stem from a single submission. I sampled comments such that there were almost equal numbers of comments on both the male and female-dominated communities.

iii. The Reddit API, although free, only allows for certain types of queries like the top/trendy submissions, and has a rate limit of 1000 entities. Due to this severe limitation, I made use of Pushshift, "a social media data collection, analysis, and archiving platform available to researchers" [2]. Many recent studies specifically in the field of NLP which focus on social media sites, have made use of this API for their analysis [7].

iv. However, because the Reddit API is cutting off archival data to Pushshift, I queried from other relevant subreddits such as r/LadiesofScience, r/xxstem, r/womenintech, r/TwoXChromosomes, r/TrollYChromosome, and r/OneY to ensure a more comprehensive analysis. Wherever the gendered-forums were more generalized, I employed search terms like 'stem' and 'engineer'. I was able to scrape approximately 1,20,000 comments for women and 40,000 comments for men subreddits.

b. Content Analysis
    i. Preprocessing: After concatenating the subreddits into two separate DataFrames, I removed the [deleted] comments, stripped the text of new line characters, punctuation, unidentified characters, extra whitespaces and converted it to lowercase. From this, I generated some preliminary graphs including a bar graph of the most frequent words, a word cloud, and employed sentiment analysis (polarization and subjectivity) using TextBlob. The graph of frequent words allowed me to identify custom stopwords. These included “women”, “men, “tech”, and “stem” since those were the keywords used to query the comments in the first place. It also included words like “one”, “would” and “like”, “work”, “people”.
    ii. Toxicity: I used Google’s Perspective API, an open-source toxicity measurement. The API works such that given a textual input, it calculates a score from 0-1 for categories such as threats, general toxicity, profanity, etc, and so I ran this over every comment. It also requires heavy pre-processing for punctuation, special, and unidentified characters to give proper results. However, due to processing power constraints, I could only analyze a subset of the dataset I collected. To do so, I sorted the comments based on the Reddit ‘score’ it had, which is the:

        `no. of upvotes - no. of downvotes`

        I sampled the top 4000 in each of the two categories of subreddits, and ran the toxicity analysis on them.
    iii. Topic modeling - Due to the larger size of the women’s dataset, I sampled the top 40,000 comments from it so the two categories (men and women) had equal number of comments. I employed LDA using Gensim because it is a great method to extract the most probable topics in the documents, and it helps with dimensionality reduction as well. For the purpose of this study, I wanted to analyze the most toxic comments from the two sets, where I take the first 50% after sorting for descending toxicity. However, due to the aforementioned limitations on the toxicity analysis, I ran the topic model on the entire datasets (40k aside). This would provide insight into what conversations were taking place around tech in women’s versus men’s forums. I tested out with a number of topics ranging from 5 to 10, landing on 6 topics as an optimal number.

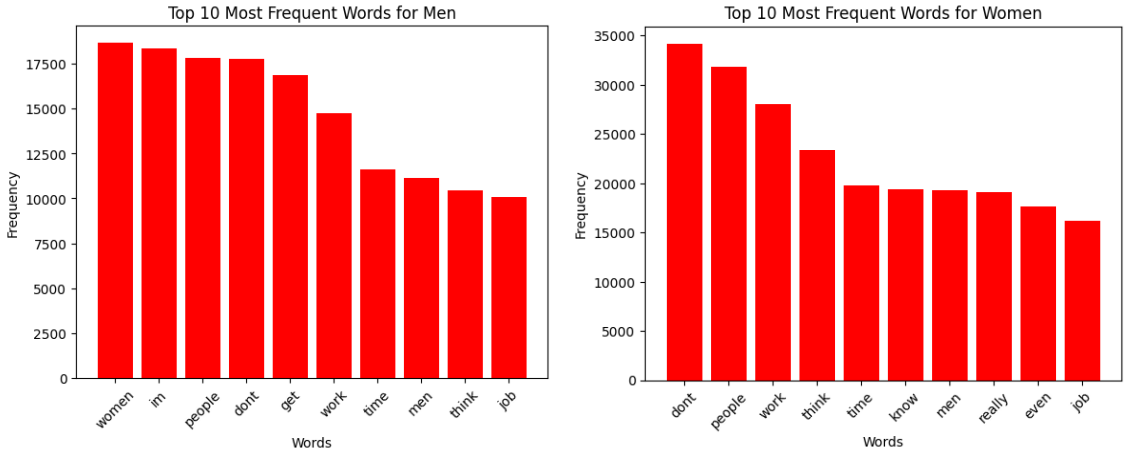
c. Limitations
    i. As discussed above, it is difficult to access consistent data from Reddit using both the API and Pushshift. Inconsistencies in data may lead to certain inconsistent results.
    ii. Perspective API, while being state-of-the-art, also has its limitations. The smaller models that it is built on can be more specifically fine-tuned for more accurate results in newer data. The API currently generalizes a lot [3]. It is also subject to other attacks that can manipulate the results [4].
    iii. Topic modeling is subjective because the processing of inferring topics from words can differ from person to person. Furthermore, this study can benefit from

alternate NLP techniques that can provide more robust analysis, including topic modeling in tandem with models like CNN's.

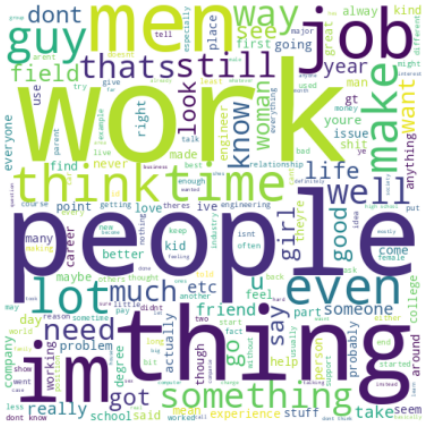
iv. This study only represent users on Reddit and are not extremely representative of online conversations in tech as a whole. Future work can take forums like StackOverflow into consideration.

v. The results of the user study are very limited since only certain people choose to fill out forms and in this case, it tends to be women. Having said that, with more information, the fact that certain people decided to fill has the potential for some interesting insight itself.

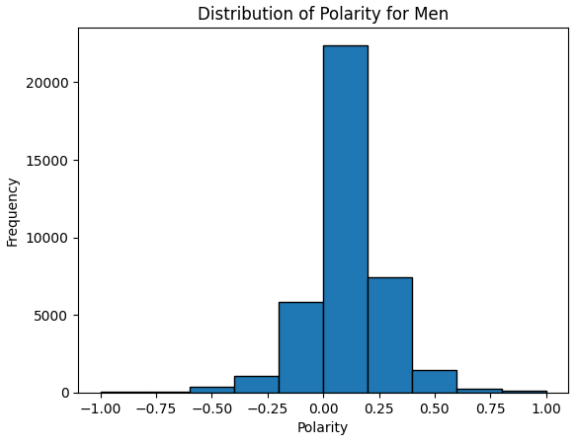
4. Analysis



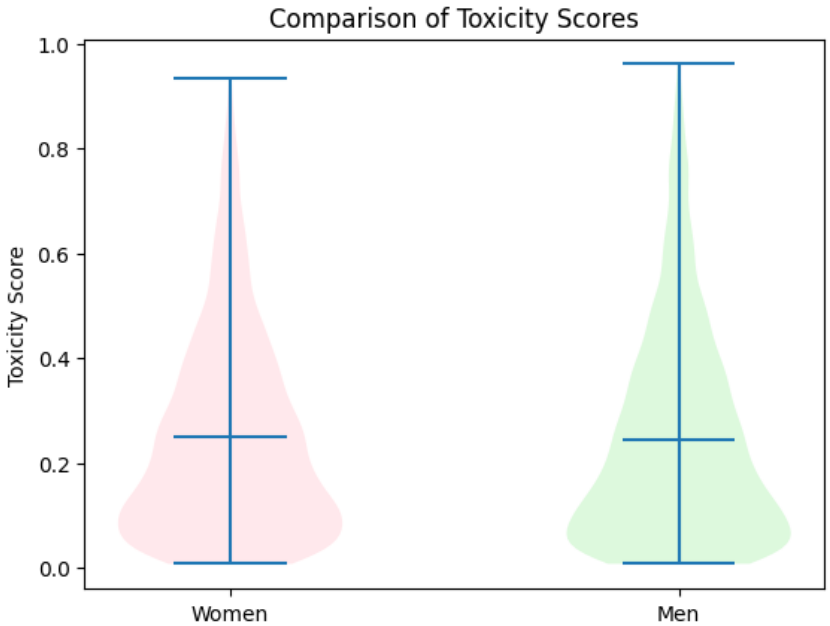
This graph allowed me to remove the custom stopwords. Through the graphs we see that the overall words with the highest counts are similar, including work, think, time, job, where the only difference is the word "don't" is the most common in womens as opposed to that of mens. While not as significant, it calls for an interesting starting point for further analysis.



This was followed by sentiment analysis using TextBlob where I measured the polarity of texts. Again, both male and female gave similar results: in both of them, the comments were more positively polar than negative.



The first research question was to analyze toxicity for men and womens' forums, for which I generated a score and appended it to a new column of the DataFrame. I created a violin chart to visualise the comparison:



The width of the violin at a given value shows how many comments have that particular score, and while it may seem that around 0.25 there is higher distribution of women toxic scores than men, overall it is pretty much equal, which has been consistent throughout the cursory linguistic analysis.

When I take the top 5 toxic comments for men and women, we see that the average toxicity of men's comments is slightly higher than womens, but there is higher negative polarity for women than men, which calls for an interesting comparison.

	comment_text	score	Unnamed: 0	comment_id	post	polarity	subjectivity	toxicity
19952	try being a clean cut businesstech oriented bl...	32	19952.0	ckl6sqv	t3_2goja2	-0.075000	0.558333	0.960691
15964	i hate that commercial because its classic app...	608	15964.0	dt2sc5q	t3_7s7sop	-0.233333	0.638095	0.933832
7863	well fuck you buddy youre just bitter you majo...	10	7863.0	dfpvwvz	t3_62w4io	0.016964	0.439881	0.933832
7863	i felt the same way for a long time in my life...	21	7863.0	hupwm48	t3_sfevkg	0.172222	0.513448	0.933832
11125	dude roses have fucking thorns unmanly you say...	10	11125.0	c5ppfzt	t3_xtqjq	0.000000	0.875000	0.933832

Men

	comment_text	score	Unnamed: 0	comment_id	post	polarity	subjectivity	toxicity
18926	fuck all the way off lmfao	32	18926.0	fhfrk8i	t3_f305ca	-0.400000	0.600000	0.933832
7152	this is fucking creepy and i know hvac guys mo...	51	7152.0	eusd90r	t3_chd6tu	0.047917	0.629167	0.933832
8251	dude she was a fucking teenager the pharm tech...	62	8251.0	e5uvawa	t3_9f8t52	-0.600000	0.800000	0.920998
17318	i dont know anybody that would chose the tech ...	88	17318.0	c3gok5l	t3_oekl4	-0.400000	0.600000	0.920998
3781	fuck that shit i hear my mother say she doesnt...	701	3781.0	gjqce9q	t3_kzvp5u	-0.110011	0.435083	0.916254

Women

While both metrics were calculated using different methods, it would be assumed that negative sentiments are more closely related to toxicity than relatively positive sentiment. That is, unless the language is complicated in a way that is sarcastic, or makes use of certain words that may be toxic, but the context is positive. When I did a qualitative analysis of the given 10 comments, I found that:

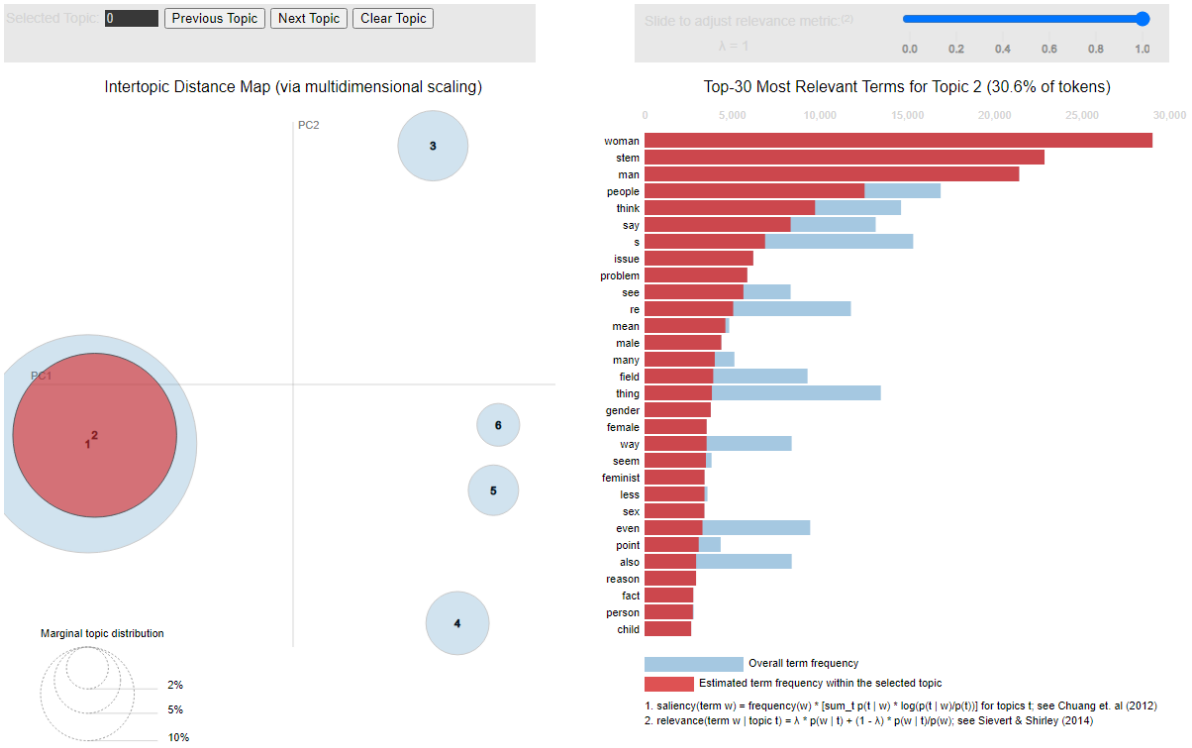
1. The one thing in common was the curse word used, which seemed like a very rudimentary way of classifying the comments as toxic.
2. The content of the men's side was leaning towards technology in general, for example,

   *"i hate that commercial because its classic apple trying to disassociate themselves from the rest of the tech world like they arent just wrapping up an inferior product in shiny casing and calling it revolutionary i hate that its so perfectly designed to push my buttons and i hate that it works f*ck you apple and your elitist mentality you suck"*

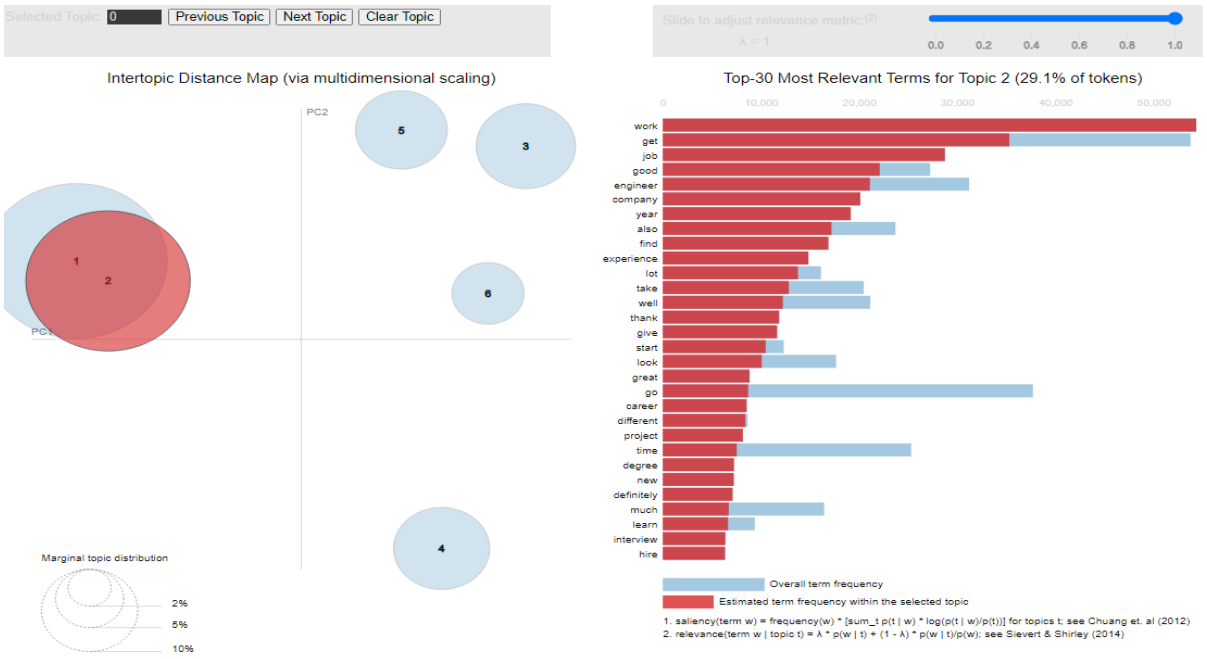
3. The women's comments were more oriented towards women in STEM, issues in the workplace, and choosing tech as a career, for example,

   *"f*ck that shit i hear my mother say she doesnt know squat about computers and repeatedly point out that she was neither taught nor permitted to learn it because it was a boys thing and she damn well can learn ampx200b this is why i mentor stem and steer every single female intern into other women in the company to go have a conversation or used to need new job."*

When I took on the task of topic modeling, the second highest topic in the men's comments were aimed towards gender, problems in STEM, a lot about men and a little about feminism. Other inferred topics include masculinity and its politics with respect to women, and working and time management.



For women, the topic seemed to be having a successful career in STEM, and has a lot more uplifting each other in their professional development. Other inferred topics include being treated as an inferior in familial settings, women being considered less than men, normalized culture of pain and patience of women.



To conclude the analysis seems to suggest little to no correlation in the toxicity of text (measured by TextBlob) and actual sentiment across women-oriented and men-oriented subreddits. This is possible for the following reasons:

1. As mentioned in the limitations, these comments were sampled from Reddit, which houses a unique demographic of users, typically urban, educated, young and tech-oriented. This reduces the expected variance across the sampled demographic in terms of worldview, expression and lived experiences.
2. We cannot extract the gender of the users because that information isn't public, and so we cannot ascertain with confidence that a women-oriented subreddit is indeed populated by female-identifying users. This is where Reddit user anonymity acts as a barrier. This also does not account for gender fluidity of the user, and is therefore not inclusive that way.
3. Finally, as with all forms of content analysis, the scale of the data we are dealing with (i.e 8000 comments) makes it difficult to study a minority population (i.e the spreaders of hate and toxicity) from aggregated statistics - which is by definition, expected to hide the true story.
4. The qualitative study is able to point out the differences, be it in the topics preferred to be discussed, or in terms of the top most toxic comments across gendered-subreddits.

5. Discussion and Future Work

While this study gave limited results, I believe there is scope for better findings with a smaller dataset size and higher computation power. The comments can be chosen in a qualitative manner. Studying the minority is challenging when using big data because crucial insights can get hidden in the majority. The proposed pipeline can be executed for any use case, including a broader study of simply examining men versus women forums, and can include more genders for inclusivity. Along with that, it can be recreated for any other subtopics within the Reddits like r/AskWomen and r/AskMen.

There is scope for diving deeper into the issue of women in tech, tech-adjacent, and business fields to examine the fine line between abusing them and sexualizing them. There are, unfortunately, *numerous* subreddits where the discussion takes place about ”hot women in suits”.

It will be interesting to examine alternative methods to take the most toxic conversation threads (taking groups of comments on certain submissions instead of randomly sampled comments) and run topic modeling on them to get better results for the overall coherent discussions happening.

References

1. Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA 12 Pages. https://doi.org/10.1145/3442381.3449861
2. Baumgartner, Jason, Zannettou, Savvas, Keegan, Brian, Squire, Megan, and Jeremy Blackburn. "The Pushshift Reddit Dataset." ArXiv, (2020). Accessed May 14, 2023. /abs/2001.08435.
3. Kumar, Deepak, et al. "Designing Toxic Content Classification for a Diversity of Perspectives." *ArXiv*, 2021, /abs/2106.04511. Accessed 15 May 2023.

4. Hosseini, Hossein, et al. "Deceiving Google'S Perspective API Built for Detecting Toxic Comments." *ArXiv*, 2017, /abs/1702.08138. Accessed 15 May 2023.
5. Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In Proceedings of the ACM Web Conference 2023 (WWW '23). Association for Computing Machinery, New York, NY, USA, 2797–2807. https://doi.org/10.1145/3543507.3583522
6. Hind Almerekhi, Supervised by Bernard J. Jansen, and co-supervised by Haewoon Kwak. 2020. Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators. In Companion Proceedings of the Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 294–298. https://doi.org/10.1145/3366424.3382091
7. Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. Proc. ACM Hum.-Comput. Interact. 4, CSCW2, Article 108 (October 2020), 23 pages. https://doi.org/10.1145/3415179
8. Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In Proceedings of the 10th ACM Conference on Web Science (WebSci '19). Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/3292522.3326045
9. Horta Ribeiro, M., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., & Zannettou, S. (2021). The Evolution of the Manosphere across the Web. Proceedings of the International AAAI Conference on Web and Social Media, 15(1), 196-207. https://doi.org/10.1609/icwsm.v15i1.18053
10. Pelzer, B., Kaati, L., Cohen, K. et al. Toxic language in online incel communities. SN Soc Sci 1, 213 (2021). https://doi.org/10.1007/s43545-021-00220-8
11. Maxwell, D., Robinson, S.R., Williams, J.R. et al. “A Short Story of a Lonely Guy”: A Qualitative Thematic Analysis of Involuntary Celibacy Using Reddit. Sexuality & Culture 24, 1852–1874 (2020). https://doi.org/10.1007/s12119-020-09724-6
12. Jacobs, Allison, Shivangi Chopra, and Lukasz Golab. "Reddit Mining to Understand Women's Issues in STEM." EDBT/ICDT Workshops. 2020.
13. Deligianni A, Horne Z. Analysing hate speech towards women on Reddit. PsyArXiv; 2023. DOI: 10.31234/osf.io/fsrq7.
14. Khan, Abeer, and Lukasz Golab. "Reddit Mining to Understand Gendered Movements." EDBT/ICDT Workshops. 2020.