Analyzing the Impact of Reddit Quarantine on r/TheRedPill

Joseph Halada

Jacob Roy

Kunal Narula

Hemant Sathian

Nicholas Yannacci

School of Communication & Information, Rutgers University

ITI 221: Data Curation & Management

Professor Shagun Jhaver

April 27, 2022

Introduction, Background, and Significance

On Reddit, communities referred to as subreddits form around shared topics. Sometimes, subreddits contain behavior that is against Reddit's policies, such as intense toxicity, inciting violence, or harassment. Reddit then intervenes using various means, such as banning a subreddit outright or "quarantining," where the subreddit is removed from search results and the front page in order to make the subreddit harder to find. In this analysis, the problem we are attempting to solve is whether or not quarantine strategies are effective in reducing toxicity and user engagement. The results of our research can help to better inform digital policy-makers as to what moderation practices are appropriate for different types of communities, or, in this case, subreddits.

Online social platforms sometimes serve as places of misinformation, toxicity, conspiracy theories, and harassment. These ailments lead to real-world consequences. The platforms, responsibly, must create moderation policies (Copland, 2020). However, there needs to be quantitative evidence to help determine the correct intervention method given the situation.

Specifically, we want to acquire and transform data via the Pushshift Reddit API such that we will be able to analyze the impact that quarantine moderation interventions have on a community. In this instance, the community is r/TheRedPill. These are the overall research questions that guided our project:

RQ1: What were the effects of the quarantine, in terms of activity on r/TheRedPill?

RQ2: What were the effects of the quarantine, in terms of toxicity, in r/TheRedPill?

While many digital outlets have attempted to moderate their content to prevent hate speech, it is still very prevalent in many places across the internet. We hope that our research will uncover whether the practice of quarantining subreddits is successful in reducing toxicity levels on Reddit. The implications of finding out whether toxicity-containment strategies such as quarantining are far-reaching. This research may inform Reddit on the value of a more active strategy to combat hate speech online.

Related Work

Previous scholars have done work in order to measure activity levels within various subreddits, the influx of new users, and levels of misogyny and racism in communities before and after their

quarantine. Two common concerns raised by these works include if the active intervention by Reddit staff leads to reduced hate speech as well as if it is feasible to engage in intervention in similar communities (Habib et al., 2021; Shen & Rose, 2019).

Two specific papers that greatly shaped our own project include *Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit* from Chandrasekharan, et al., as well as Trujillo & Cresci's *Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald*. Both of these studies examined the effects of quarantining on r/The_Donald. Chandrasekharan, et al. also analyzed r/TheRedPill (2021). Thus, we seek to partially replicate and validate these prior works. We will be analyzing r/TheRedPill, rather than r/The_Donald. We seek to validate Trujillo & Cresci's work by analyzing r/TheRedPill, a subreddit often compared to r/The_Donald. However, we will not be conducting their second research question pertaining to political bias and factual reporting. Due to Media Bias Fact Check lacking an API, we would need to create a web scraping tool to gather data. This is currently beyond the scope of the course and this project. We will not be conducting their third research question in relation to the core user groups of r/TheRedPill either. Due to difficulty in obtaining data as outlined below, there is not sufficient time to study the second order impacts of the quarantine on the core users of the subreddit. As such, we will be addressing research question one as outlined in Trujillo & Cresci's study with a focus on r/TheRedPill.

Our work expands on the work done in these previous studies, hoping to analyze the effect that quarantine had on r/TheRedPill on activity and toxicity. Given that the topics are relatively new, academic work studying moderation interventions has lagged behind the demand for such studies. We hope to help solve this issue by creating quality research that can join the larger literature regarding the effects that moderation policies can have on subreddits and other online communities. It is our hope that our analysis will serve as a jumping off point for future research in the field.

Objectives, Goals and Outcomes

Our initial set of goals included answering research questions regarding the activity, toxicity, and core user impact of a quarantine on r/TheRedPill over a period of 60 weeks. Our project narrowed significantly in scope with feedback from the instructors. After presenting our initial project proposal, the instructors recommended dropping our third research question in regard to analyzing the activity of core users of r/TheRedPill outside of the subreddit following the moderation intervention. This was due to time constraints, as answering this research question would have been a slow and time consuming process.

Another way in which the scope decreased is the timeframe of data collection. Following the milestone report presentation, our timeframe reduced from 60 weeks in total to 120 days. This was due to data collection concerns as well as time constraints in completing the project. The final way in which our initial goals changed has to do with our measure of toxicity. While the initial plan was to get the toxicity scores from Perspective API, the size of our dataset meant that we would not have a sufficient amount of time to collect all the scores. Instead, using the lexicon dictionary described in Farrell et al.'s *Exploring Misogyny across the Manosphere in Reddit*, we determined the frequency with which comments containing words found in their dictionaries appeared before and after the quarantine intervention. These dictionaries contain keywords for many types of online hate speech, including violence, racism, and homophobia, among others (Farrell et al., 2019).

With that being said, our objectives and findings are as follows:

The first objective of our project consisted of answering the impact of the quarantine on r/TheRedPill in regards to activity. This involved looking at the number of comments pre- and post-quarantine. This data, shown in Figure 1, indicates a clear spike and downward trend following the quarantine intervention. Additionally, we looked at the daily active users pre-quarantine and post-quarantine of the subreddit in Figure 2. Similar to Figure 1, the quarantine did have an impact on the number of unique daily users in the subreddit.

Our second major objective was to examine the impact of the quarantine on toxicity within the subreddit. To do so, we examined the frequency of hate speech lexicon dictionary keywords in comments

from the subreddit before and after the quarantine. In Figure 3, there is an immediate impact of the quarantine on the frequency of the keywords for nearly all of the dictionaries. Finally, Figure 4 shows the counts of the toxicity keywords for each dictionary in our sample dataset before and after the quarantine. In all significant cases, the keywords appeared less frequently following the intervention.

These findings align with prior literature. Our findings and outcomes match with our expected results. If there had been a deviation between expectations and outcome, the first place we would have looked at more closely would be the initial dataset. Looking at a larger period of time as we initially set out to do could have provided more insight and serves as a point for future analysis.

Description of Work Accomplished: Data

To begin our data gathering, we elected to use the Pushshift API in order to conduct this project. Trujillo and Cresci made use of it in their analysis of moderation on r/The_Donald (2022). In their paper, they analyze the submissions and comments of r/The_Donald over a large period of time, which inspired our replication within r/TheRedPill.

Through the group's research process around Pushshift API, we discovered the limit of 500 submissions or comments per query as dictated by the size parameter. Due to this, the group looked for workarounds in the form of libraries built on top of Pushshift that enable consecutive queries to Pushshift without needing to write overly complex code to collect our dataset of hundreds of thousands of comments. The two libraries that are commonly used include PMAW and PSAW. The PMAW library was simpler to write with and comprehend, becoming the project's preference over PSAW. PMAW's thoroughly documented API, multi-threaded wrapper, and rate limits made it easy to collect the data.

Using the PMAW library from mattpodolak, we wrote a Jupyter Notebook ("trp_data_grabber.ipynb"), gathering submissions and comments data over the initial specified timeframe (03/02/2018 - 04/26/2019), inputted as Unix time. This date range includes 30 weeks before and after r/TheRedpill's quarantine on September 28, 2018. In a span of time between one and two hours, 7,118 submissions and 200,139 comments were obtained in this manner.

This data was then preprocessed. In each dataset, the number of columns was reduced to those in the scope of the project. For the comments dataset, the retained columns included “id”, “created_utc”, “user_removed”, “author”, and “author_fullname”, and “body”. For the submissions dataset, the retained columns included “id”, “created_utc”, “author”, and “selftext”.

The “id” key is a randomly generated string of numbers and letters that can be used to identify individual comments. The “created_utc” key is the date and time that the comment was created in seconds from the Unix epoch (01/01/1970, 00:00:00 UTC). The “user_removed” key is a boolean field that represents whether or not a user’s comment was removed. The “author” key contains the user names of the users who made individual comments. The “author_fullname” key seems to be a randomly generated string of numbers and letters used to identify users; not all comments/users contain this field, however. The “body” key contains the text of the comment. The “selftext” key is the body text of the submission, independent of the submission title.

The datetime function converted the UNIX time of “created_utc” to a more human readable form. In anticipation of using Perspective API, we counted the number of comments in the collected date range that were removed and no longer had text to analyze. This was around 5% of all comments in the dataset. Finally, comments that were not removed and had the “user_removed” value as np.nan were set to 0. We also counted the number of submissions in the collected date range that were removed and no longer had text to analyze. This was around 51% of all submissions, a significant portion of the data.

The final measure we took during the data collection process was running a data sanity check to see if we were actually capturing data for our entire timeframe. Below is the count for each month of comments and submissions in our initial time frame:

```
comments.groupby([comments['year'], comments['month']]).size()
```

```
year  month
2018  4        26678
      5        30311
      6        30677
      7        28616
      8        22791
      9        23387
      10       10862
      11        9774
      12       10935
2019  4         6108
dtype: int64
```

```
submissions.groupby([submissions['year'], submissions['month']]).size()
```

```
year  month
2018  3          833
      4          951
      5         1067
      6          976
      7         1071
      8          930
      9          870
2019  2          231
      4          189
dtype: int64
```

Our project as proposed was centered around 9/28/2018, observing 30 weeks before and 30 weeks after this date, from around March 2018 to April 2019. As of the data sanity check, comments data from March 2018, January 2019, February 2019, and March 2019 are missing. In addition, submissions data from October 2018, November 2018, December 2018, January 2019, and March 2019 are missing. We speculate the missing data is the result of one of two possibilities. Either the data that we require is currently on an inactive storage shard, and thus is inaccessible, or there were significant moderation activities that resulted in the lack of archiving of this data, meaning that the data was removed before it was archived.

As a result of this challenge in the collection and preparation of the data, we adjusted the time frame of the project down to 120 days. Moving forward, we only used the comments dataset, dropping the submissions from the scope of the project due to the lack of usable data. After filtering the comments

dataframe to the new range of 7/30/2018-11/27/2018, our data was ready to start analysis and visualization for our first research question.

The project, including scripts and links to the dataset, can be found at the following github page: https://github.com/jroy12345/reddit-TheRedPill. This data preprocessing can be found in the "clean_data_initial.ipynb" file.

Description of Work Accomplished: Approach

The initial data collection and preprocessing, along with the data sanity check and subsequent narrowing of scope, allowed us to address our first goal in determining the activity of the subreddit over time. The visualizations in order to answer this goal involve the number of comments and daily active users before and after the quarantine. These metrics were used by Trujillo and Cresci in their analysis of moderation on r/The_Donald (2022), and the visualizations based on their own were easy to create in the "analysis_RQ1.ipynb" file.

The second question in regards to tracking toxicity in the subreddit required a lot of discussion and different approaches. We needed to integrate an outside method of measuring toxicity. While we initially looked at Perspective API as the Habib et al. paper and other prior works used it as the toxicity measure, this was rejected due to the time constraint of the semester (2021). Following instructor feedback, we moved in favor of tracking the frequency of keywords for each of the lexicon dictionaries in Farrell et al.'s paper (2019). This serves as the proxy for toxicity. The "clean_data_RQ2.ipynb" file in the github walks through the process of merging the lexicon dictionaries with the collected subreddit comment data. The comments in the file document the steps taken in order to actualize this. This file returns a combined data frame that can be visualized, which is done in "analysis_RQ2.ipynb." These visualizations are inspired by those found in the Farrell et al., paper (2019).

Some limitations of the project have to do with the metric for toxicity. While the frequency of keywords serves as a proxy for toxicity, the words are taken out of context and tokenized. This misses the overall toxicity of the sentence or entire comment, which is something that the Perspective API would have been able to include. Second, some of the keywords found in the lexicon dictionaries include special

characters. The process by which the comments were tokenized into words means that some words that were present in the comments that should have been counted are not included in the final counts of the dictionaries. Our counts of toxicity may be lower than in actuality.

The final limitation of our approach is that the timeframe may be too restrictive to see trends play out over time. The 60 day period after the quarantine may be too short to see the full effect.

The main strength our data approach takes is that it is firmly founded in past research. We have data and visualizations that can be compared to other studies. These visualizations make the trends simple to understand. The other strength of our approach is that this process is easier to work with in the constraints of the semester. We did not have time to troubleshoot the data collection issues or work with Perspective API. However, we did have sufficient time to do a full analysis using the lexicon dictionaries.

Description of Work Accomplished: Results

Figure 1

Figure 1 is a line chart that shows daily comment totals in the subreddit, r/TheRedpill, 60 days before and 60 days after 09/28/2018, the day the quarantine was imposed. A dotted line in the middle of the stacked line chart represents this quarantine date. When viewing the line chart, it is clear that after the quarantine was enacted, the daily totals for comments in the subreddit saw a sharp decline. Nearly 1,200 comments were posted on the date the quarantine was put in place, with only around 400 comments being

posted in the subreddit a few days after the quarantine was introduced. In the 60 days after the quarantine was enacted in r/TheRedpill, totals never reached the 800 comment threshold. Based on Figure 1, the quarantine was impactful in reducing daily comment totals.

Figure 2

Figure 2, a line chart, depicts the number of unique authors in the subreddit, r/TheRedpill, 60 days before and 60 days after the quarantine date, with the dotted line representing this date. Similar to Figure 1, it is clear that after the quarantine was enacted the daily totals of unique authors in the subreddit

sharply declined. Nearly 500 unique authors were active on r/TheRedpill on the date the quarantine was enacted, with only around 200 unique authors remaining active on the subreddit a few days after the quarantine was introduced. Throughout the 60 days after the quarantine was enacted in r/TheRedpill, unique author totals never reached 350 daily authors. Through Figure 2, the quarantine was impactful in reducing daily unique author totals in the subreddit.

Figure 3

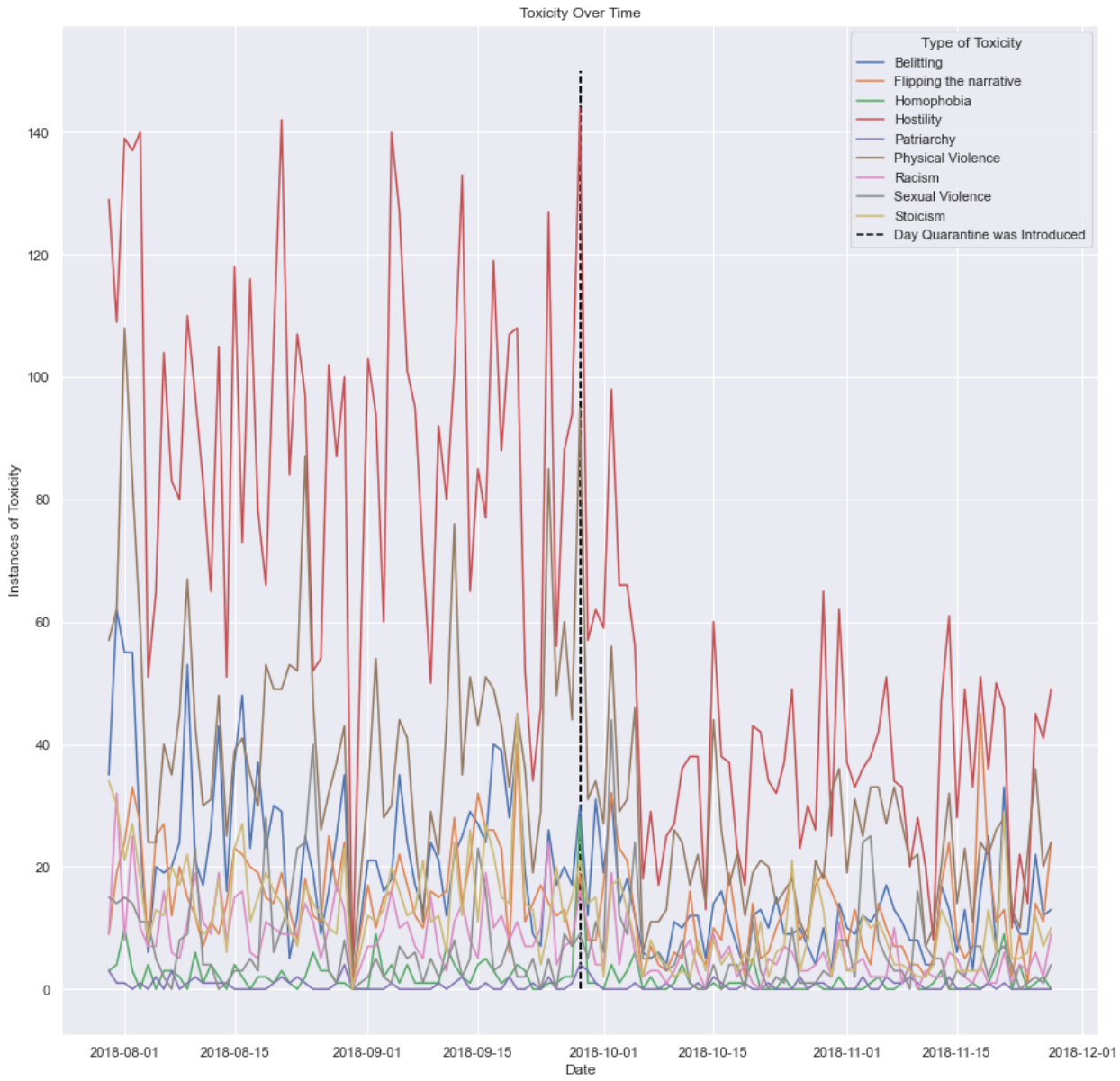

Figure 3 is a stacked line chart that represents the cumulative totals of keyword frequency of a specified type of toxicity in the comments dataset 60 days before and 60 days after the quarantining. Toxicity classifications mentioned in the stacked line chart derive from the lexicon dictionaries described by Farrell et al. and include belittling, flipping the narrative, homophobia, hositility, patriarchy, physical violence, racism, sexual violence, and stoicism (2019). A dotted line in the middle of the stacked line chart represents the quarantine date.

In the data from before the quarantine, the frequency of words classified as ‘hostile’ are much more prevalent over keywords classified by other types of toxicity. Comments including words classified as ‘belittling’ and promoting ‘physical violence’ share second place in terms of total occurrences with similar cumulative totals over the 60 day period before the quarantine. Comments including words classified as ‘stoic’, ‘flipping the narrative’, and promoting ‘racism’ share third place with similar cumulative totals, while comments including words classfied as ‘homophobic’ and promoting ‘patriarchy’ share fourth place.

After the quarantine moderation intervention went into effect, there is a drastic reduction in the frequency of keywords classified as ‘hostile’ in the subreddit. This reduction in ‘hostile’ keyword frequency is readily apparent in Figure 3. There is no longer a distinct, visible gap with any other frequency of dictionary keywords. Another interesting trend in Figure 3 after the quarantine is introduced is the lack of major change in the cumulative frequency of keywords from the non-‘hostile’ dictionaries classifications. This lack of significant change in comment totals not classified as ‘hostile’ is addressed further in Figure 4.

Figure 4

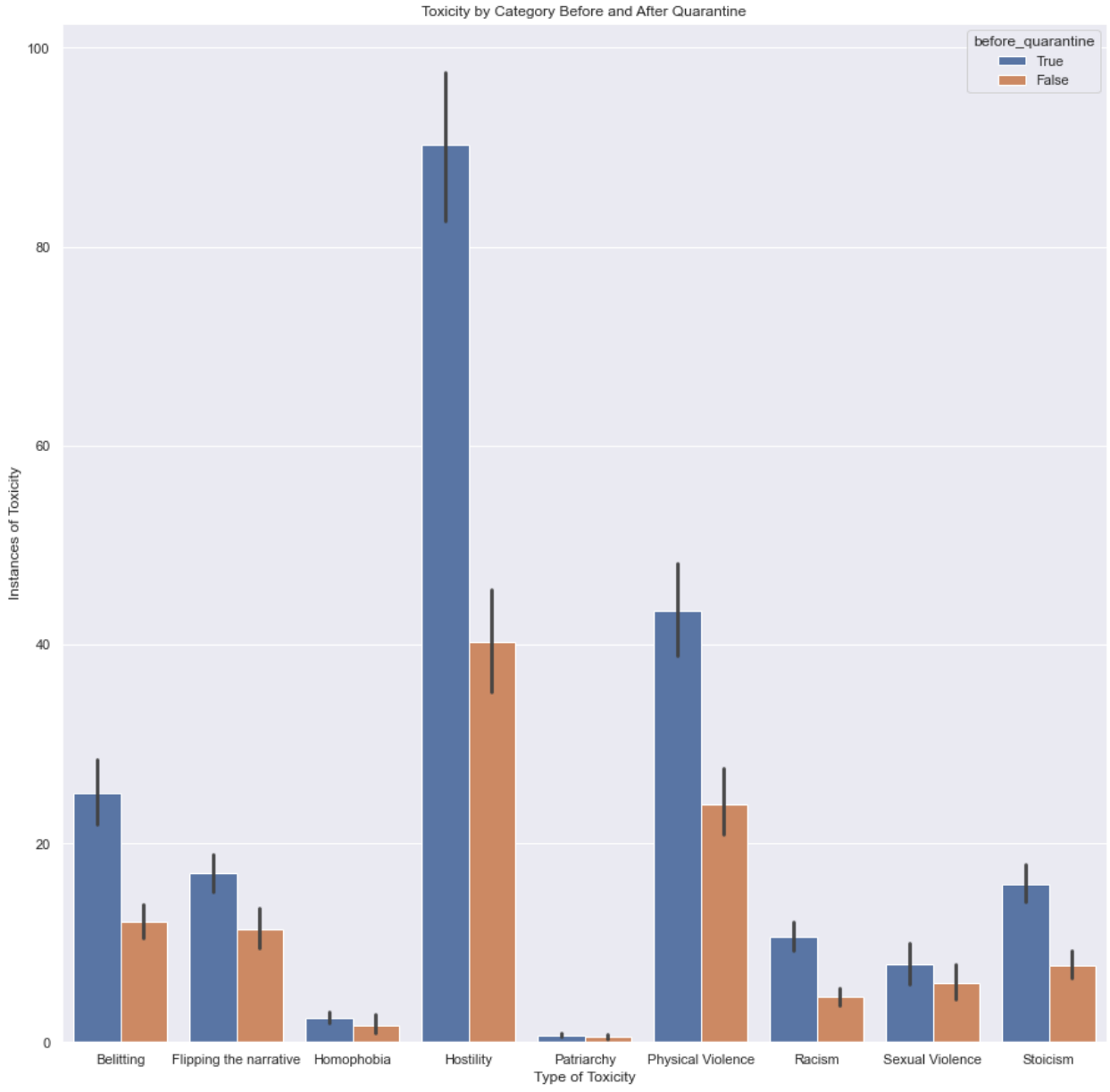

Figure 4, a grouped vertical bar chart, depicts keyword totals for each lexicon dictionary specified by the specific type of toxicity before and after the quarantine was enacted in r/TheRedpill. The blue bars represent keyword totals before the quarantine was enacted, while the orange bars represent keyword totals after the quarantine was enacted. As mentioned previously in the analysis of Figure 3, comments including words classified as 'hostile' saw a large decrease in frequency after the quarantine was enacted. To a lesser extent, comments including words classified as 'physical violence', 'racism', 'stoicism', and

‘belittling’ saw drops in frequency following the quarantine. Interestingly, comments including words classified as ‘flipping the narrative’, ‘homophobia’, ‘sexual violence’, and ‘patriarchy’ did not see much difference in their comment totals after the quarantine was enacted, with comments including words classified as ‘patriarchy’ with the least amount of change in totals before and after the quarantine was introduced. While the quarantine imposed on r/TheRedpill was effective in substantially reducing comments including words classified as ‘hostile’, it was not unilaterally effective in reducing comments that include words of different toxicity classifications.

Discussion of Outcomes, Implications and Conclusion

Upon the introduction of the quarantine, there was a severe reduction in both unique users per day, as well as comments per day. Similarly, the prevalence of toxicity as measured by the frequency of different lexicon dictionary words in the comment database experienced a severe reduction. Based on the prior examined literature, these outcomes were expected. This lends additional support and validation to prior works that have examined similar communities and the impacts of moderation interventions like quarantining. The potentially most surprising thing in our data was the severity of the drop in activity and toxicity within the subreddit, as we witnessed a more severe drop compared to similar studies. However, this may be as a result of our project limitations, as discussed below.

There are few new insights gained from this project. Instead, it provides validation to prior insights gained via similar papers. One insight is that moderation interventions have a significant effect on user activity of subreddits, as well as the types of posts made. Quarantining seemed to have an extremely high impact on the amount of toxicity, especially relative to the unique user activity. In addition, it seems that from week-to-week and day-to-day, communities have wide variation in the number of posts and toxicity as a result. Similar findings are present in prior works.

There are some limitations associated with our work that have been previously outlined. First, the time frame of 60 days before and after quarantine may be too restrictive, especially compared to prior literature, which observed time frames of up to 2 years (Trujillo & Cresci, 2022). In addition, we only analyzed comments, rather than also analyzing submissions, due to missing data from Pushshift API. One

direction for future work would be to analyze both submissions and comments over a larger time frame, so long as the data is available. Another limitation of our approach is that by opting to analyze toxicity via a lexicon dictionary, we missed out on using Perspective API, which is more exhaustive. We opted to use this lexicon due to the Perspective API's rate limitations, which would result in the project taking too much time. Using the Perspective API could provide a more accurate and thorough analysis of toxicity, as well as align more closely with prior research. An additional place of potential future study is looking at the behavior of core users of the subreddit. While this was included in the original project proposal, it was dropped following instructor feedback. The second order impacts of quarantines and moderation interventions more generally are a rich area of potential study. Finally, our process can be replicated and adapted to any subreddit by changing the time range and target subreddit within the "trp_data_grabber.ipynb" notebook, then proceeding through the other notebooks.

List of Teammate Names

The group members in this project are Joseph Halada, Jacob Roy, Kunal Narula, Hemant Sathian, and Nicholas Yannacci.

References

Chandrasekharan, E., Jhaver, S., Bruckman, A., & Gilbert, E. (2021, January). *Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit*. Retrieved February 8, 2022, from https://shagunjhaver.com/research/articles/jhaver-2021-quarantining/jhaver-2021-quarantining.pdf

Copland, S. (2020, October 21). *Reddit quarantined: can changing platform affordances reduce hateful material online?* Internet Policy Review. Retrieved February 8, 2022, from https://policyreview.info/articles/analysis/reddit-quarantined-can-changing-platform-affordances-reduce-hateful-material

Farrell, T., Fernandez, M., Novotny, J., & Alani, H.: *Exploring Misogyny across the Manosphere in Reddit.* Proceedings of the 10th ACM Conference on Web Science - WebSci 2019, pp. 87–96 (2019). https://doi.org/10.1145/3292522.3326045

Habib, H., Musa, M. B., Zaffar, F., & Nithyanand, R. (2021, November 22). *Are Proactive Interventions for Reddit Communities Feasible?* arXiv.org. Retrieved February 8, 2022, from https://arxiv.org/abs/2111.11019

Mattpodolak. (n.d.). *Mattpodolak/PMAW: A multithread pushshift.io Api wrapper for reddit.com comment and submission searches.* GitHub. Retrieved March 30, 2022, from https://github.com/mattpodolak/pmaw

Shen, Q., & Rose, C. (2019, August). *The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy*. ACL Anthology. Retrieved February 8, 2022, from https://aclanthology.org/W19-3507/

Trujillo, A., & Cresci, S. (2022, January 17). *Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on R/the_donald*. arXiv.org. Retrieved February 8, 2022, from https://arxiv.org/abs/2201.06455