The Usage of Authoritarian Language in Reddit Communities

Leo Wang
leowang@terpmail.umd.edu

Kyle Lin
klin1215@terpmail.umd.edu

David Li
dl1@terpmail.umd.edu

Abstract

We analyze the use of authoritarian language in Reddit communities (subreddits), focusing on the distinction between users versus moderators and between drama, controversial, and/or political subreddits. In contrast to existing literature, which focuses on the use of such language by governments and influential figures [7], this analysis provides a first look at how such rhetoric may be reflected in the interactions between the general public, as measured through submissions and comments on Reddit. At its core, this work is a novel extension of the authoritarian language classifier developed by Mochtak to the Reddit platform [6].

We find that there is a distinct difference between the language used by users and moderators, with moderators having a lower mean and a more concentrated distribution of authoritarian scores. There is also little correlation between a posts score and the associated authoritarian score. Instead, authoritarian score is more closely related to the type of subreddit, with political subreddits having higher mean authoritarian scores. Finally, we examine the banning of extremist political subreddits to show a correlation between the authoritarian score and hate speech/extremism.

1 Introduction

Reddit is a popular internet forum consisting of many user created and moderated communities, or subreddits, within which users discuss a broad range of topics. Reddit moderators are users with moderation powers for a given subreddit. We distinguish moderators as users with moderation powers and users as those without such privileges. Notably, these moderators are neither selected nor paid by the Reddit staff; they are typically either the creators of a given subreddit or otherwise promoted by the existing subreddit moderation team and work for free.

Both moderators and users can post submissions and comments. A submission is a top-level entry that may appear on the front-page of a subreddit, while comments are posted underneath a submission or in response to other comments. We will refer to the combined submissions and comments as either "entries" or "posts."

Users and moderators can furthermore "vote" on a post, giving it either a positive "upvote" or negative "downvote." A post's score is the net upvotes minus downvotes, which we use as a proxy for popularity.

Outside of Reddit, authoritarian language, especially as it relates to governments or elites, is a well-researched topic [7]. But, less is known about how this language may be reflected within the emergent online communities that characterizes the 21st century. Our primary contribution is thus the novel application of Mochtak's authoritarian language classifier [6] onto the submissions and comments generated by Reddit contributors. As Mochtak's classifier sourced its training data from UN speeches, this analysis will center around how such language may be replicated by Reddit contributors, and if so, whether it mirrors authoritarian language. We interpret the output probability as an "authoritarian score," with 1 indicating "very authoritarian" and 0 indicating "very democratic."

Specifically, we analyze the difference in language used between users and moderators and between 20 subreddits split evenly between a political, controversial, drama and control group. The analysis relating to the users versus moderators will be referred to as the "Users" category while the analysis relating to the subreddits will be referred to as the "Subreddits" category. We measure the same values for users, moderators, and each subreddit:

- distribution of authoritarian scores,

- mean and standard deviation of authoritarian scores over time,

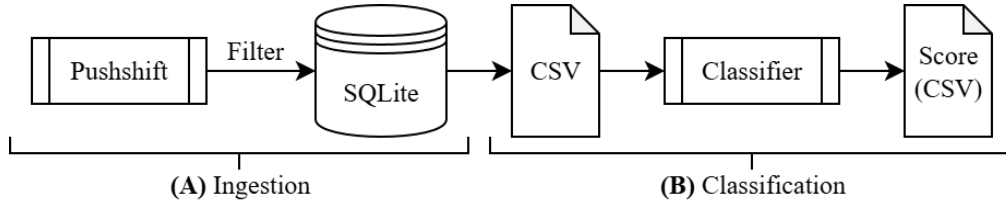- mean authoritarian score with respect to a comment's/submission's score.

Figure 1: The data pipeline, split into **A.** ingestion (filtering and preparing raw Pushshift data) and **B.** classification.

2   Related Work

There is a growing sphere of literature around Reddit's social dynamics, though none share this report's focus. Almerekhi developed a classifier to classify toxic Reddit submissions and comments and further investigated the divide between users, moderators, and subreddits. Their work is structured very similarly to our report, though with a different focus and an end goal of producing a well-tuned model for classifying toxic posts (as opposed to our goal to use an existing model to present novel data analysis) [1].

Research on authoritarianism in Reddit was also investigated by Hendricks, though was more limited in scope. His work centered on the discourse in r/The_Donald (an alt-right subreddit centered on Donald Trump) and how self-proclaimed "authoritarian communists" within r/socialism defined authoritarianism. His work places greater emphasis on examining how such rhetoric builds community or leads to splintered ideologies, while we focus on a more holistic analysis of authoritarianism as defined by Mochtak's classifier [3].

Lastly, Marrazzo compares Reddit's decentralized moderation practices, e.g. delegating subreddit moderation to community moderators, to a federalist government system. This resembles our goal of comparing moderator authoritarian language to global authoritarian discourse [5].

Our source data is derived from from the Pushshift Reddit dumps [2]. Pushshift is an API for retrieving raw Reddit data developed by Jason Baumgartner (Stuck_In_The_Matrix), et al., from which we used the dumped data maintained by Baumgartner, Watchful1, and RaiderBDev [8–11, 13]. We ingest comment and submission data for both the monthly and per-subreddit dumps, which provides us with extensive post metadata to analyze, including a post's score, body, poster status (user or moderator), and timestamp.

The ingested data is classified using the authoritarian language classifier developed by Mochtak [6]. It is a large language model that outputs a probability a text is authoritarian, with 0 representing "very democratic" and 1 representing "very authoritarian." The model was trained on United Nations speeches and the V-Dem Electoral Democracy Index, where text from well known authoritarian leaders is considered to define what authoritarian speech is (and vice versa). Due to the difference in mediums between UN speeches and

Reddit communities, it is necessary to note that absolute authoritarian scores should not be seen as authoritative. Instead, our analysis primarily focuses on the relative differences between users/moderators and subreddits. Other considerations, including filtering and incorrect classifications, are further elaborated on in the Methodology and Results sections.

3   Methodology

The data processing pipeline is described in Figure 1. Submission and comment data was derived from the Pushshift dumps [8–11, 13], then filtered and organized into an SQLite database. The classification step extracts the relevant entries (a submission or comment) from the database as a CSV file then feeds it into the classifier. This returns an authoritarian score for both the entire entry as well as per sentence, which forms the basis of our analysis.

Submissions are treated as top-level comments and thus are not differentiated in the classification step, though they do differ slightly in their respective filtering rules during the ingestion step. However, note that the title of a submission is stripped to maintain parity with comments, as comments do not have a title. In total, 351,368 entries were ingested and classified, 129,988 for the users category and 221,380 for the subreddits category.

3.1   Ingestion

Due to compute and storage limitations, an aggressive filter was applied to reduce the size of the input data. The applied rules were:

1. the body must have at least 5 sentences,

2. the body must contain only ASCII characters,

3. comments may not be from AutoModerator (but submissions can).

Common markdown formatting was also adjusted. This includes:

1. links in the form [display](url) are replaced with "display",

2. links in the form https:\\ or http:\\ are removed,

3. superscript (^), blockquote (>), and header (#) notation is removed.

Of note is the 5 sentence minimum and restriction to the ASCII alphabet. This is because the classifier was trained off UN speeches, and so expects well-formed, long-form text. The ASCII restriction also doubles as a proxy to filter out non-English entries.

The only difference between the submission and comment filter was that comments may not be from the AutoModerator bot. This is because AutoModerator comments are typically automated responses posted to most, if not all, submissions, while no such pattern exists for AutoModerator submissions (which are often manually posted by moderators).

3.1.1 Users

The submission and comment data for the user category was sourced from the Pushshift monthly dumps [8–11] for the 13 month period from 09/2024 to 09/2025. Each month has 355 million entries on average, with the entire dataset consisting of 4.6 billion entries. This was filtered down to 244 million points. For each month, 10,000 entries were randomly sampled, 5,000 from non-moderator users and 5,000 from moderators. In total, this section of the analysis consists of 129,988 entries (12 entries were dropped due to failing to be classified) evenly distributed across a year, with 64,997 moderator entries and 64,991 non-moderator entries. Table 1 describes the collected months.

3.1.2 Subreddits

The submission and comment data for the subreddit category was sourced from the Pushshift subreddit dumps [13], which covers the time range from 06/2005 to 12/2024. 20 subreddits were manually selected to be used in this analysis, which broadly fit into the sub-categories of drama, politics, controversial, and control (subreddits that are expected to have low authoritarian scores). Each sub-category has 5 constituent subreddits and were chosen to ensure an even distribution over the political spectrum and to have a medium to large user-base. For each subreddit, 1,000 entries were sampled per month for the most recent 12 months that had data (as some subreddits were banned). In total, this section of the analysis consists of 221,380 entries. Table 2 describes the collected subreddits.

A brief overview of each subreddit is as follows:

1. Drama

   (a) r/IAmA: Q&As about a notable event/profession regarding the poster.
   (b) r/AITAH: users post a story and seek advice on whether they were the "asshole" in the situation.
   (c) r/offmychest: users post about an emotional or otherwise impactful personal experience.

   (d) r/changemyview: users debate potentially controversial opinions.
   (e) r/unpopularopinion: users post unpopular opinions.

2. Politics

   (a) r/politics: the largest US politics subreddit.
   (b) r/Conservative: the largest subreddit for the US Republican party.
   (c) r/democrats: the largest subreddit for the US Democratic party.
   (d) r/TheNewRight: alt-right subreddit for the "New Right" ideology.
   (e) r/dsa: left-wing subreddit for the Democratic Socialists of America.

3. Controversial

   (a) r/The_Donald: alt-right subreddit for supporters of Donald Trump.
   (b) r/ChapoTrapHouse: left-wing subreddit for the Chapo Trap House podcast, which promotes dirtbag left (aggressive and vulgar) tactics.
   (c) r/MGTOW: acronym for "Men Going Their Own Way," which is a misogynistic subreddit that promotes the separation of men and women.
   (d) r/antiwork: subreddit that promotes anti-work ideologies.
   (e) r/KotakuInAction: right-wing subreddit that centers on gaming related discourse and Gamer-Gate.

4. Control

   (a) r/MadeMeSmile: users post heartwarming content.
   (b) r/explainlikeimfive: users post questions and receive responses that attempt to explain the answer simply.
   (c) r/aww: users post pictures of cute animals.
   (d) r/Damnthatsinteresting: users post about topics that are novel or generally interesting.
   (e) r/HobbyDrama: long-form content that documents niche drama, with a focus on accurately and neutrally describing said drama.

3.2 Classification

From the SQLite database, for each subreddit, user month, and moderator month, the following columns are generated:

1. body: the raw text of the entry.
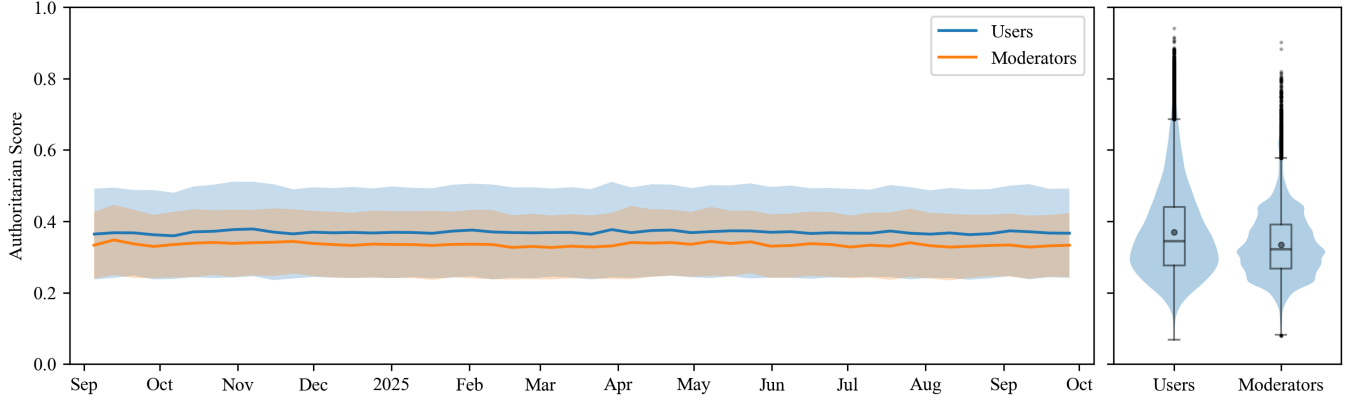2. subreddit: the subreddit in which the entry was posted.

Figure 2: Left: The mean and one standard deviation (shaded) of the moderator and user posts from 09/2024 to 09/2025. Right: The total distribution of authoritarian score for moderator and user posts. (N=129,988)

3. `id`: the unique identifier associated with the entry.

4. `parent_id`: the `id` of the entry the entry is replying to; empty if the entry is a submission.

5. `created_utc`: a Unix timestamp representing the time an entry was posted.

6. `score`: the net upvotes minus downvotes the entry has.

7. `num_sentences`: a calculated value (not part of the Pushshift dumps) indicating the number of sentences the entry.

8. `distinguished`: an integer indicating whether this entry was sent by a user, moderator, or admin.

The first part of the classification process is converting the SQLite database table into a CSV, which is done simply by a script which copies each column of the table into a corresponding column in the output CSV.

Once the data is in CSV form, it is classified using the authoritarian language classifier, resulting in a new CSV with an extra column `auth`. This column is a floating-point number ranging from 0.0 to 1.0, which represents the authoritarian language score the classifier has assigned to the string in the `body` column, where 0.0 is least authoritarian (defined as democratic by the classifier) and 1.0 is most authoritarian. This was the most time consuming part of the data pipeline due to the significant time it took for a entry to be classified.

4 Results

Our work relies on the interpretation of the authoritarian scores reported by the classifier. Thus, this analysis inherits much of the same biases and issues present in the classifier itself. The most prevalent is the classifiers reliance on keywords (often related to religion or nations) to inform its reported score. This can cause posts that are not authoritarian to be classified as such due to specific, out of context keywords:

> Erm. That is no Doxxie. That is no angel either. That is obviously the legendary Good Boy of 'Who's a good boy?' Clearly he is, that is Mr. Good Boy.

(`auth`=0.677, in `r/aww` referring to a pet)

Furthermore, the inherent difference between the UN speeches the classifier was trained on and the comments and submissions within the Reddit sphere means that there is not a one-to-one correlation between the authoritarianism exhibited in the UN and in Reddit. This is partially alleviated by the ingestion filter by attempting to only classify long-form, well-structured posts that are more likely to resemble the classifier's training data. Thus, we consider authoritarian scores to only be meaningful relative to each other. Finally, the relatively small sample size means that the following analysis may omit some outliers.

4.1 Users

Figure 2 shows the authoritarian score for both users and moderators from 09/2024 to 09/2025. Despite similar means at 0.367 for users and 0.335 for moderators, their distributions are noticeably different with $p < 0.01$. The users had a heavier right skew, leading to a higher standard deviation of 0.127 compared to the moderators' 0.092, despite the distribution to the left of the median being very similar. Examining the user/moderator quantiles reflects this, with both the 25% (0.276, 0.268) and 50% (0.345, 0.322) quantiles being very similar (users, moderators). A divergence is only seen at 75%, with users at a higher 0.440 compared to the moderators' 0.391. Furthermore, examining the right violin plot shows that both users and moderators had similar most common authoritarian score at around 0.3.

This suggest that there exists some baseline tone/type of post that is used by both users and moderators alike across all of Reddit. Despite different distributions, both groups have

the same most common score and very similar lower quantile values across the entire period. The difference is that users, being a larger and more diverse group, had more extreme users which shifted the distribution upwards. However, a major confounder is that moderators tend to post copy-pasted responses to many users, which may bias their distribution downwards given such responses are often neutral in tone:

> If you feel your post was removed by mistake or is an exception, feel free to message the moderators using this link. Please also give a short explanation.
> (`auth`=0.322, quantile=0.5, moderators)

Higher authoritarian scores are usually correlated with more political posts, as the classifier is likely picking up on certain keywords:

> At min 6, they show deportations. Obama deported 3 million to 3.5 million people. That is 1100 people per day. There is NO evidence that Obama sent 1100 people before a judge per day. That didn't happen. Obama used the FISA courts, and deported people without due process. These historians are hypocrites. They didn't complain when Obama deported people, ILLEGALLY
> (`auth`=0.746, quantile=0.99, users)

And:

> Not all opinions matter. Do you have the freedom of speech? Yes. Would I take you seriously? No.
> (`auth`=0.746, quantile=0.99, users)

But, this keyword approach means that some posts are incorrectly classified:

> Alternative strategy:
> 1. Have all black orcs
> 2. Send them in all at once
> 3. Kill everything before WAGGGHHHH is needed
>
> (`auth`=0.746, quantile=0.99, users, referring to the *Warhammer 40,000* franchise)

The lower authoritarian scores are usually correlated with more neutral posts about non-political topics:

> well 1st the owner has to take his share. then all the managers that you totally need to manage a complex task like division. ofc they get their share too for handling such complex tasks. And they just had to rent one external consultant for the outcome you see in the picture OP has provided.
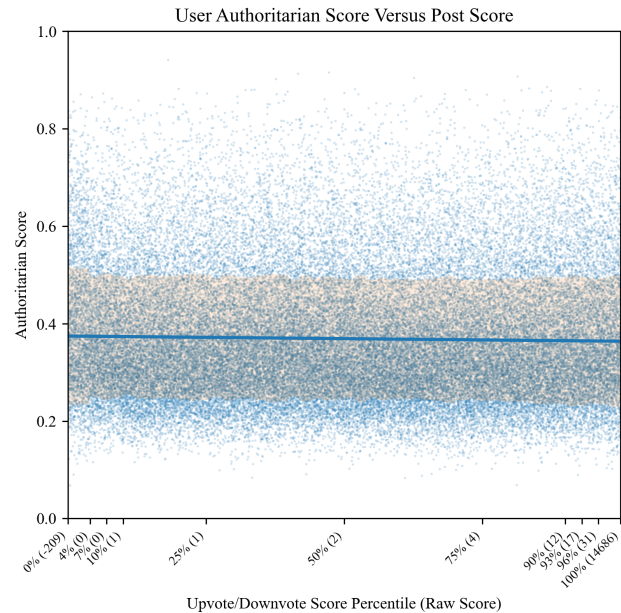> (`auth`=0.276, quantile=0.25, users)

And:

Figure 3: The combined user and moderator distribution of authoritarian score versus post score (net upvotes and downvotes) over the total time range from 09/2024 to 09/2025. (N=129,988)

> I've been happy with my Xboxes over the years. Started out with a PS2 as a kid but then moved to a 360 so we could play Halo. Have stuck with it ever since (360 in like 05 then XB One in 2013 and now a Series X since like 2022 or so?). Like other people said, take a look at the big name games that are exclusives. And if you have gamer friends, see what they run. My 3 main gamer buddies are all on Xbox so we just roll with that.
> (`auth`=0.276, quantile=0.25, users)

There appears to be little to no correlation between the score (the net upvotes and downvotes) of a post and the authoritarian score, as seen in Figure 3. The distribution of authoritarian score is near to be constant for all post scores. Instead, it is more likely that authoritarian score is dependent upon the host subreddit.

4.2 Subreddits

The difference in categorization between political and non-political speech is especially clear in Figure 4. The political category, consisting of `r/politics`, `r/Conservative`, `r/democrats`, `r/TheNewRight`, and `r/dsa` had a significantly higher mean score at 0.485 compared to the other categories 0.376, $p < 0.01$. Interestingly, all political subreddits had a very consistent distribution (including `r/The_Donald` and `r/ChapoTrapHouse` which also center around political
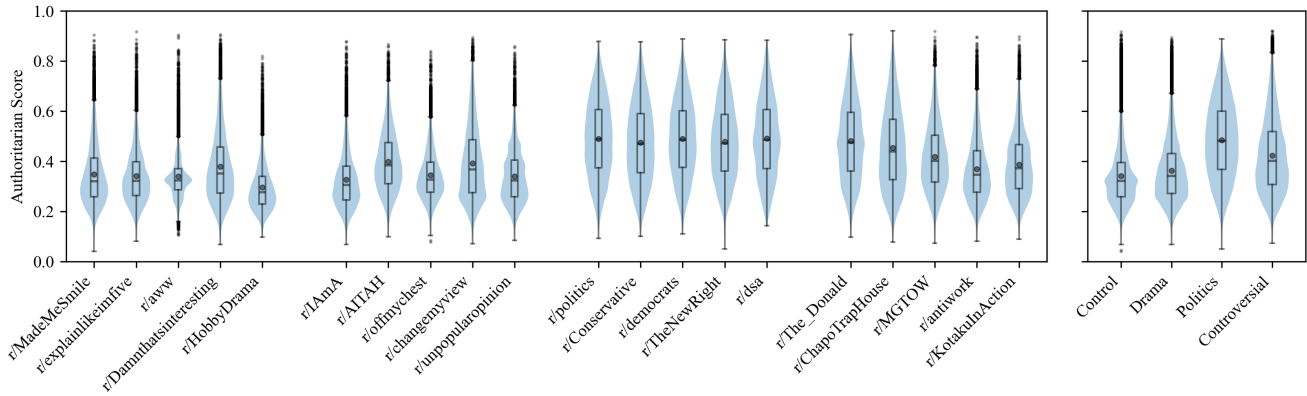
Figure 4: The distribution of authoritarian scores across selected subreddits. Subreddits are organized by group: (in order) control, drama, politics, and controversial. (N=221,380)

discussion). The other subreddits congregate around the 0.3 mark noted in the previous section.

This is expected, as political subreddits are more closely related to the type of authoritarian rhetoric used in the UN given that many of the topics, i.e. geopolitics, are shared. It is also worth noting that some of the non-political subreddits explicitly ban political discussion, e.g. r/Damnthatsinteresting, which may artificially lower their distribution. As the mean user authoritarian score in the previous section is closer to the mean in non-political subreddits, political subreddits do not appear to dominate discourse on Reddit.

High authoritarian scores within the political subreddits explicitly mention various political topics, though are not biased towards either conservative or liberal views (most of the political subreddits sampled center around American politics):

> Oh, something is coming, and this time we're prepared for that. The only question is what is coming. And I'm not shocked that Biden is up, because no version of Biden will save him in the general. He has been yelling at Garland to ramp up the prosecutions but it hasn't worked. He's a decaying old figure who, no matter what he does won't become younger by election day. People have realized that it's his DEI marxist staffers that are running the show.
> (auth=0.813, quantile=0.99)

And:

> .. not to mention mental acuity, generosity and above all, class .. that Drumpf could now never possess in his geriatric insanity .. as far as winning in November .. that could be up to the Electoral College and fuckery by Drumpf's domestic terrorist Cult of MAGA and foreign enemies of the United States of America ..
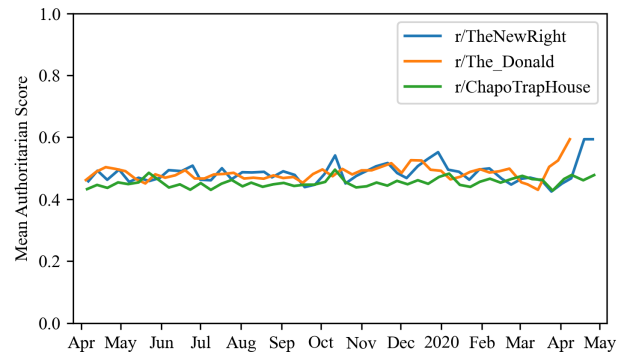> (auth=0.814, quantile=0.99)

Figure 5: The mean authoritarian score of r/TheNewRight, r/The_Donald, and r/ChapoTrapHouse before their ban from Reddit, covering the period from 04/2019 to 04/2020. (N=28,160)

High authoritarian scores in non-political subreddits are generally due to incorrect classification as a result of the keyword sensitivity exhibited by the classifier, though they are still lower than those in the political subreddits:

> A big. Viscous. Toothy. Freaky tailed. BIG. Hamster. Lol
> (auth=0.676, quantile=0.99, in r/aww)

And:

> This is Lucifer. Lucifer hates me. I adore Lucifer and feed him twice a day. If I go near him he attacks me.
> (auth=0.680, quantile=0.99, in r/aww)

Despite the misclassifications, the authoritarian score does appear to be loosely correlated with extremism. In June of 2020, Reddit banned some subreddits for

hate speech, including `r/The_Donald`, `r/TheNewRight`, and `r/ChapoTrapHouse` [4]. Figure 5 shows the period immediately before each subreddit's ban. Both `r/The_Donald` and `r/TheNewRight` saw significant increases in mean authoritarian score 1-2 months before their ban, which may reflect the hate speech/extremist views Reddit banned the subreddits for. This implies that there is a correlation between hate speech/extremist content and increasing authoritarian scores.

5 Conclusion

We show that the language used by users and moderators are different, with moderators using less authoritarian language likely on account of neutral templated responses that are frequently copy-pasted. However, both groups are still closely related (with similar means), suggesting that outside of the templated posts, users and moderators do interact with the wider community similarly.

While there is no noticeable pattern between the score of a post and its authoritarian score, there is a large difference between subreddits. Political subreddits have significantly higher mean authoritarian scores but are a minority within the entirety of Reddit. The banning of `r/The_Donald` and `r/TheNewRight` shows that increasing scores are loosely correlated with hate speech/extremist content.

6 Future Work

Some other features were planned, including:

1. authoritarian score when a user replies to a moderator (and vice versa),

2. if user/moderator language changes based on the language of the post being replied to,

3. if a user's/moderator's language changes depending on the subreddit they are posting in,

4. sorting moderators by activity and/or how many subreddits they moderate,

5. a longer time analysis to see if major world events correlates to spikes in authoritarian score.

More work could also be done on filtering the posts, e.g. stripping advanced markdown syntax or using an advanced language detect to remove all non-English posts, and excluding other bots in addition to AutoModerator (though this may be difficult as Reddit treats bots as identical to user accounts).

The main challenge that restricted us from covering these topics was a lack of compute and storage resources. In total, the raw data for the 13 month period for the users category took approximately 520 GiB, and the subreddits took an extra 74 GiB. This represents approximately 6.6 billion entries, which took upwards of 80 hours to gather, process, and classify. Furthermore, much of the pipeline described in Figure 1

was specially developed for this project, notably the ingestion and filtering from Pushshift and the scripts to wrap the classifier. As such, more extensive analysis or data coverage was infeasible for this report.

7 Artifacts

The source code for this report is available at `https://github.com/Andallfor/396H-Project` and includes the classifying scripts and the data ingestion module. The actual processed dumps or classified text is not included due to storage limitations [12].

References

[1] Hind Almerekhi, Supervised by Bernard J. Jansen, and co-supervised by Haewoon Kwak. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 294–298, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *CoRR*, abs/2001.08435, 2020.

[3] Joshua Hendricks. Alt-right of the_donald and authoritarian communists on reddit: Internet memes to build community. Master's thesis, University of Southern Mississippi, 2022.

[4] David Ingram and Ben Collins. Reddit bans hundreds of subreddits for hate speech including trump community, 2020.

[5] Vincent Marrazzo. The federalists of the internet? what online platforms can learn from reddit's decentralized content moderation scheme. *Nebraska Law Review*, 2023.

[6] Michal Mochtak. Chasing the authoritarian spectre: Detecting authoritarian discourse with large language models. *European Journal of Political Research*, 64(3):1304–1325, 2025.

[7] Elia Pizzolitto, Ida Verna, and Michelina Venditti. Authoritarian leadership styles and performance: a systematic literature review and research agenda. *Management Review Quarterly*, 73:841–871, 2023.

[8] RaiderBDev. Reddit comments/submissions 2025-07, 2025.

[9] RaiderBDev. Reddit comments/submissions 2025-08, 2025.

[10] RaiderBDev. Reddit comments/submissions 2025-09, 2025.

[11] stuck_in_the_matrix, Watchful1, and RaiderBDev. Reddit comments/submissions 2005-06 to 2025-06, 2025.

[12] Leo Wang, Kyle Lin, and David Li. Cmsc396h-project, 2025.

[13] Watchful1. Subreddit comments/submissions 2005-06 to 2024-12, 2025.

A User Metadata

Month	Raw count	Filtered count
09/2025	364,400,451	18,329,833
08/2025	374,731,044	18,984,206
07/2025	377,926,771	19,457,589
06/2025	361,955,878	18,771,520
05/2025	365,727,232	19,378,542
04/2025	349,247,132	18,849,741
03/2025	365,416,129	19,439,652
02/2025	340,567,900	18,153,859
01/2025	371,887,689	20,005,184
12/2025	338,056,561	17,917,922
11/2024	339,726,491	18,128,014
10/2024	342,994,866	18,572,378
09/2024	328,811,163	17,899,637

Table 1: Processed Pushshift monthly dumps for the user category.

B Subreddit Metadata

Category	Subreddit	Time span	Raw count	Filtered count	Sample size
Drama	r/IAmA	01/2024–12/2024	26,418,432	3,498,061	12,000
	r/AITAH	01/2024–12/2024	23,751,957	3,481,885	12,000
	r/offmychest	01/2024–12/2024	11,245,621	1,960,112	12,000
	r/changemyview	01/2024–12/2024	15,924,864	4,390,646	12,000
	r/unpopularopinion	01/2024–12/2024	49,914,916	4,626,244	12,000
Politics	r/politics	01/2024–12/2024	206,228,369	20,588,197	12,000
	r/Conservative	01/2024–12/2024	20,126,511	1,329,385	12,000
	r/democrats	01/2024–12/2024	2,241,816	139,199	11,806
	r/TheNewRight	04/2019–04/2020	182,732	14,612	4,160
	r/dsa	01/2024–12/2024	99,283	7,930	1,414
Controversial	r/The_Donald	04/2019–04/2020	53,111,792	3,512,943	12,000
	r/ChapoTrapHouse	04/2019–04/2020	10,320,276	510,208	12,000
	r/MGTOW	02/2020–01/2021	5,740,257	787,775	12,000
	r/antiwork	01/2024–12/2024	24,104,472	2,346,889	12,000
	r/KotakuInAction	01/2024–12/2024	8,627,189	1,372,307	12,000
Control	r/MadeMeSmile	01/2024–12/2024	17,818,417	598,701	12,000
	r/explainlikeimfive	01/2024–12/2024	21,128,493	3,702,010	12,000
	r/aww	01/2024–12/2024	47,354,286	1,325,022	12,000
	r/Damnthatsinteresting	01/2024–12/2024	25,960,408	960,695	12,000
	r/HobbyDrama	01/2024–12/2024	729,125	104,058	12,000

Table 2: Processed Pushshift subreddit dumps for the subreddits category.