

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

1 Abstract

We present a dual analysis on the usage of authoritarian language in Reddit communities (subreddits), focusing on distinctions between users versus moderators and between a selection of drama, controversial, and/or political subreddits. While existing literature examines the usage of such language across governments and influential figures ([source](#)), this analysis provides a first look at how such rhetoric may be reflected in the interactions between the common man, as measured through the submissions and comments on Reddit. At its core, this work is a novel extension of the authoritarian language classifier developed by Mochtak [1] to the Reddit platform.

We find that ...

2 Introduction

Reddit is a popular internet forum consisting of many user created and moderated communities (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 (colloquially known as a "post") is a top-level entry that may appear on the front-page of a subreddit, while comments can be posted underneath a submission or in response to other comments. Both submissions and comments have an aggregated score metric determined by users/moderators, which is used as a proxy for popularity.

Authoritarian language, especially as it relates to governments or elites, is a well-researched topic([source](#)). Less is known, however, 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 [1]

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 or not it mirrors authoritarian language. We utilize the output probability as an "authoritarian score", with 1 indicating very authoritarian and 0 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. For clarity, 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 same values for users, moderators, and each subreddit:

- Distribution of authoritarian score,
- Mean and standard deviation of authoritarian score over time,
- Mean authoritarian score per sentence in a given comment/submission,
- Mean authoritarian score with respect to a comment's/submission's score.

3 Related Work

TODO - should include how the classifier works here

4 Methodology

The data processing pipeline is described in [Figure 1](#). Submission and comment data was derived from the Pushshift dumps [2–6], 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, which is then feed into the classifier. This returns an authoritarian score for both the entire entry as well as per sentence.

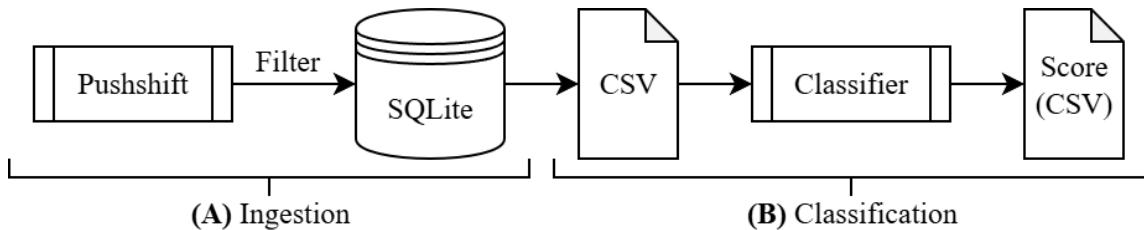


Figure 1: The data pipeline, split into **A.** ingestion (filtering and preparing raw Pushshift data) and **B.** classification.

Submissions are treated as top-level comments, and thus are not differentiated in the classification step and differ only slightly in their respective filtering rules during the ingestion step. In total, 351,368 entries were ingested and classified, 129,988 for the users category and 221,380 for the subreddits category.

4.1 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. Additionally, the ASCII restriction doubles as a proxy to filter out non-English posts.

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 typically manually posted by moderators).

4.1.1 Users

The submission and comment data for the user category was sourced from the Pushshift monthly dumps [2–5] 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 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.

4.1.2 Subreddits

The submission and comment data for the subreddit category was sourced from the Pushshift subreddit dumps [6], 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.

TODO - should explain what each subreddit is/represents.

4.2 Classification

From the SQLite database, for each subreddit, user month, and moderator month, we have the following columns:

1. `body`: the raw text of this message
 2. `subreddit`: the subreddit this message was posted
 3. `id`: the unique string associated with this message
 4. `parent_id`: the `id` of the message this message is replying to; empty if this message is a submission
 5. `created_utc`: the time which this message was posted, in Unix

6. score: the net upvotes minus downvotes this message has
7. num_sentences: the number of sentences this message has; sentences defined by .?! followed by whitespace, or any character followed by a new line
8. distinguished: an integer indicating whether this message 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 can be classified using the authoritarian language classifier [1], 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 message to be classified. Furthermore, preliminary results seemed to indicate that the classifier model is biased towards messages which have mention of religion or nations, rating them more authoritarian than what might be expected; this is likely due to the fact that the classifier was trained on United Nation speeches, which frequently include mention of religion and nations.

5 Results

caveat about interpreting results, reference what was said in related work

5.1 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 $p < 0.01$. This is most notable in their standard deviations, where users had a higher 0.127 standard deviation compared to the moderators' 0.092. Interestingly, both groups had near identical values at minus one standard deviation (*I'm not sure what the correct phrasing is, I'm referring to the fact that on the graph the bottoms are the same*). 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. 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 0.3.

This suggest that there exists some baseline tone/type of message 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 time 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)

However, 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 messages 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,

"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)

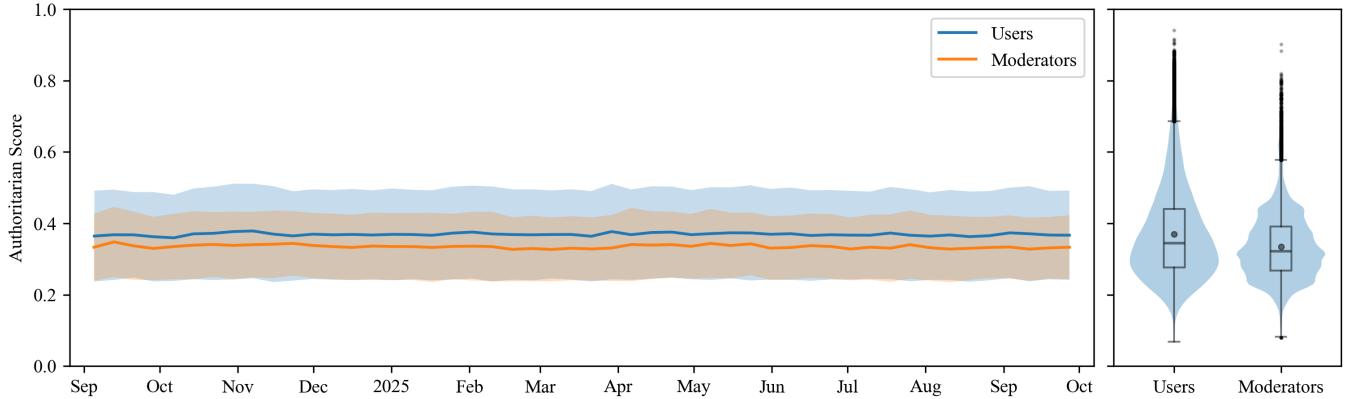


Figure 2: Left: The mean and one standard deviation (shaded) of the moderator and user posts from 09/2024 to 09/2025. Right: The total distribution of authoritarian score for moderator and user posts. (N=129,988)

5.2 Subreddits

The difference in categorization between political and non-political speech is especially apparent in Figure 3. 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 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/Damnthatssinteresting, which may artificially lower their distribution.

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

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

6 Conclusion

7 Future Work

References

- [1] Michal Mochtak. Chasing the authoritarian spectre: Detecting authoritarian discourse with large language models. *European Journal of Political Research*, 64(3):1304–1325, 2025.
- [2] RaiderBDev. Reddit comments/submissions 2025-07, 2025.
- [3] RaiderBDev. Reddit comments/submissions 2025-08, 2025.
- [4] RaiderBDev. Reddit comments/submissions 2025-09, 2025.

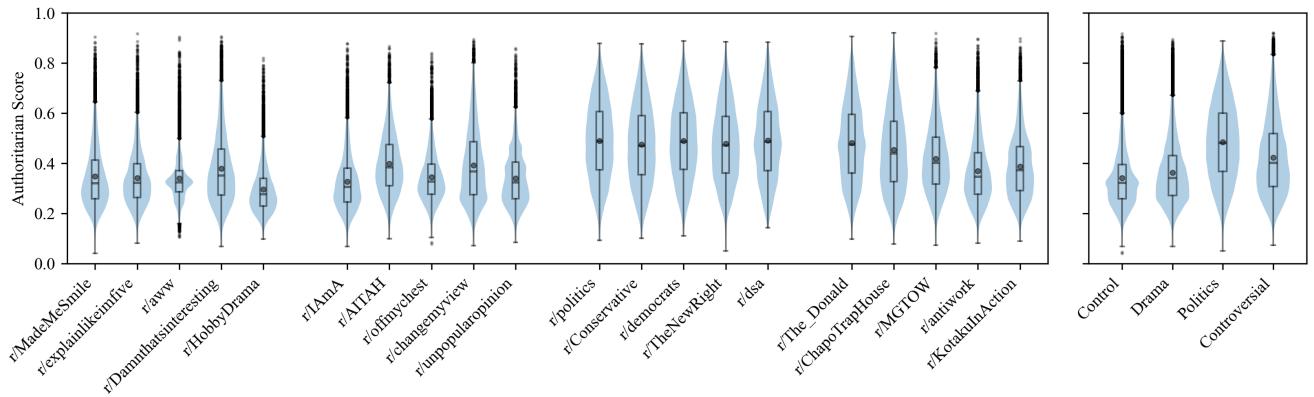


Figure 3: The distribution of authoritarian scores across selected subreddits. Subreddits are organized by group: (in order) control, drama, politics, and controversial. (N=221,380)

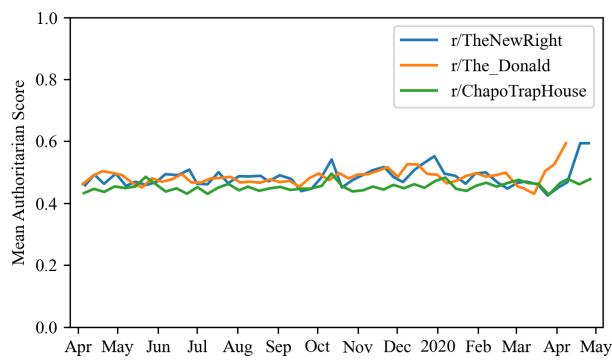


Figure 4: The mean authoritarian score of r/TheNewRight, r/The_Donald, and r/ChapoTrapHouse prior to their ban from Reddit, covering the period from 04/2019 to 04/2020. (N=28,160)

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/ama	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
	r/politics	01/2024–12/2024	206,228,369	20,588,197	12,000
Politics	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
	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
Controversial	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
	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
Control	r/damnthatstinteresting	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.