Authoritarian Language Use by Reddit Moderators: Intermediate Report

Leo Wang
leowang@terpmail.umd.edu

Kyle Lin
klin1215@terpmail.umd.edu

David Li
dl1@terpmail.umd.edu

1 Problem Description

Reddit is an internet forum where users discuss a broad range of topics. Reddit moderators are users with moderation powers in certain "subreddits," or subforums. Our goal is to determine how holding power in an online community affects a user's diction.

We specifically analyze the use of authoritarian language in Reddit, with an expanded, dual focus on moderators compared to non-moderators and subreddits compared to other subreddits. For ease, these two categories will be referred to as "users" (moderator versus non-moderator) and "subreddits" (subreddits versus other subreddits). We plan to test for correlation with the following factors for the users category:

- size of subreddit,
- differences in interactions when replying to a normal user or moderator,
- number of upvotes,
- trends over time.

For the subreddits category, planned factors include:

- political affiliation,
- controversiality,
- activity/popularity.

Because this is an open-ended, exploratory problem, we will base evaluation on how holistic the resultant model is by incorporating as many secondary variables as possible. We may also base evaluation on the confidence level of the correlation of variables in the resultant model, as well as how reasonable the results appear under manual inspection.

2 Related Work

There is limited prior research about Reddit's social dynamics. Almerekhi detected and categorized toxic posts and comments [1]. Hendricks examined how self-proclaimed "authoritarianist communists" Reddit users define authoritarianism [3]. In its third chapter, Marrazzo's master's thesis compares Reddit's decentralized moderation practices, such as delegating subreddit moderation to community moderators, to a federalist government system [4]. Marrazzo's thesis resembles this project's goal of comparing moderator authoritarian language to global authoritarian discourse.

Our work builds off Mochtak, using their large language model to score authoritarian language of a text passage [5]. Mochtak trained the model on United Nations speeches and the V-Dem Electoral Democracy Index. Pushshift is an API designed to assist Reddit moderators [2]. The maintainers of Pushshift provide dumps containing many posts or comments obtained through the API [6]. We have used Mochtak's model to score comments from a Pushshift dump.

3 Progress

3.1 Data Collection

Submission (posts) and comment data was derived from the Pushshift dumps [6], then filtered and written into an SQLite database for the classification step. We consider submissions to be top-level comments and thus are treated largely the same as comments. For both the user and subreddit categories, the same set of filters were applied:

1. The body must have at least 5 sentences (as our classifier expects long-form text).
2. Comments may not be from AutoModerator, as it is a default moderation bot and not an actual user. We allow submissions from AutoModerator, as these are typically manually created by moderators.
3. The body must contain only ASCII characters. The classifier does not handle non-ASCII text (e.g. emojis) as it is based off UN speeches. This also filters non-English text.
4. Common markdown formatting is adjusted. This includes:

Month	Raw count	Filtered count
07/2025	377,926,771	19,457,589
08/2025	374,731,044	18,984,206
09/2025	364,400,451	18,329,833

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

(a) Links in the form [display](url) are replaced with "display".
(b) Links in the form https:\\ or http:\\ are removed.
(c) Superscript (^), blockquote (>), and header (#) notation is removed.

3.1.1 Users

The data was sourced from the Pushshift monthly dumps [6], including both submissions (posts) and comments, from 09/2024 to 09/2025. A given month has ∼370 million entries, and so the entire data range encompasses ∼4.4 billion entries. After filtering the data, ∼5% remained or ∼19 million entries. 10,000 random samples are then taken, 5,000 from non-moderator users and 5,000 moderators. In total, this section of the analysis makes use of 120,00 entries evenly distributed across a year, with 60,000 moderator entries and 60,000 non-moderator entries. Table 1 describes the collected months.

3.1.2 Subreddits

The data was sourced from the Pushshift subreddit dumps [7], 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 some cases, a subreddit may have less than 12,000 total sampled posts. All subreddits have been processed and are detailed in Table 2.

3.2 Classification and Analyzation

After the data is filtered into the SQLite database, an intermediary script converts it into a CSV file. This is then passed into the classifier, resulting in a new CSV file with extra columns including a per-text authoritarian score, which can be aggregated into various graphs. So far, a script that aggregates CSV files into bar charts for the two categories has been created. All these processes are efficient besides the classification process itself, which is a limitation of the Hugging Face model being used. This limits the total amount of data that can be aggregated, but it is still within an acceptable sample size. This means this bottleneck should not affect our results significantly outside of potentially missing outliers, which the project does not focus on.

4 Challenges

Given the lack of compute resources (such as an external server), all programs and processing must be done on the team's personal computers. Thus, the large volume of data ingested presented three main issues: file size, performance of the data collector, and performance of the classifier. The Pushshift dumps for a single month are typically upwards of 50 GB, and so cannot all be stored concurrently. Instead, the current pipeline processes such files individually before deleting the dumps. The processed data (i.e. the sampled and classified entries) and the dumps for the subreddit category are significantly smaller and can be stored locally.

Next, the original implementation of the data collector would have required over a full week to process the full year of data. This has since been significantly optimized and can now be completed within a day.

The final issue is the performance of the classifier itself. It cannot be easily optimized (as it is developed by a third-party) and classifies at a relatively slow 10,000 entries per hour. Given the size of the input data, this forces the usage of a heavy filter and small sample size. Currently, a total of 360,000 entries are classified between the two categories, which will take an estimated 36 hours.

Lastly, it is worth noting potential biases in Mochtak's model. While classifying a small sample, based on looking over the resulting scores the model produced, the model assigns higher authoritarian scores to texts that mention countries and religion, even when the context and content do not seem authoritarian on inspection. This may be because the model is trained on United Nations speeches, which frequently mention countries and religion. Thus, the model may not be directly translatable to text, such as internet comments, that falls outside its domain of minutes-long spoken political rhetoric. This is partially mitigated by enforcing a 5-sentence minimum for each entry.

5 Plan

The data collection side of the project is essentially complete. The only remaining task is to generate the data for the remaining 9 months for the user category. Much of the groundwork for the classification and analyzation of the data has also been completed.

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.

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

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

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

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