# Usage of Authoritarian Language in Reddit Community Moderators

David Li
*dl1@terpmail.umd.edu*

Kyle Lin
*klin1215@terpmail.umd.edu*

Leo Wang
*leowang@terpmail.umd.edu*

## 1 Motivation

Authoritarianism is a common field of research, especially in the context of global politics and leaders. However, understanding how it propagates in and is used by the common person is understudied. To this end, this project will be analyzing the usage of authoritarian language by community (subreddit) moderators across Reddit. The bulk of the work will be testing for correlations to a variety secondary factors, including but not limited to: size of subreddit, number of subreddits moderated, comparison to the general non-moderator population (globally and within their specific subreddit), differences in interactions with subreddits the user does/does not moderate, and how active a moderator is.

Language will be classified by its relation to the language used by commonly recognized political authoritarian figures. This allows the paper to draw a larger conclusion about how (or if) traditional authoritarian language maps to the power dynamics present within subreddits, and if so, what factors could cause the common person to trend towards authoritarian practices.

## 2 Prior Work

This work is centered around the application of an existing authoritarian language classifier [5]. There is also an existing corpus of work surrounding the Reddit userbase, such as Almerekhi's work on categorizing toxicity [1] or Marrazzo's study showing how Reddit's decentralized moderation practices (that is, delegating subreddit moderation to community moderators) mirrors a federalist governmental system [4]. This is similar in nature to this project's goal of comparing moderator authoritarian language to the global authoritarian discourse. Finally, Hendricks has examined how self-proclaimed "authoritarianist communists" users define authoritarianism [3].

## 3 Methodology

The central work of this project is analyzing existing databases and supplementing it as needed. Authoritarian language of a user will be sourced from their comments and posts and measured through an existing authoritarian language classifier [5]. These comments and posts will be sourced through the Pushshift comment dumps and its related API [2, 6], with additional required metadata gathered either through other such databases or manually scrapped.

## 4 Evaluation

Because this is an open-ended, exploratory problem, evaluation will be based on how holistic the resultant model is (i.e. incorporating as many secondary variables as possible). The final goal is to be able to build a conclusion about the trends of authoritarian language in subreddit moderators.

## 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. *Aquila*, 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.