Authoritarian Language Use by Reddit Moderators: Proposal

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. Yet, there is little prior work investigating how it spreads in and is exercised by common people. This project will analyze the use of authoritarian language by moderators, who have moderation powers in specific "subreddits," on Reddit. The bulk of the work will be testing for correlations to secondary factors including

- 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
- how active a moderator is.

This offers insights on the psychological aspects of holding a position of power in an online community and how a moderator's diction reflects this.

We will classify language by its relation to that used by authoritarian political figures. This allows us to test whether 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

We will center this work around the application of an existing authoritarian language classifier [5]. Also, there is existing 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, i.e., 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 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. We will determine authoritarian language of a user from their comments and posts and measured through an existing authoritarian language classifier [5]. We will source these comments and posts through the Pushshift comment dumps and its related API [2, 6], with extra required metadata gathered either through other such databases or scraped. We can then measure the text from the Pushshift comment dumps for authoritarian language using the existing classifier [5] to test and draw conclusions. Much of this work will consist of analyzing and testing comments and posts from users who meet specific criteria against those from Reddit moderators.

4 Evaluation

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 when compared to a manual inspection. The goal is 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. 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.