Auditing Automated Content Moderation Tools

Geralyn Chong[1], Yash Patel[2], Danish Pruthi[2], and Arjun Bhagoji[1]

[1]University of Chicago
[2]Indian Institute of Science

Abstract

We describe and audit content moderation tools provided by common platforms such as Discord and Twitch.

1 Introduction

2 Background

- History of content moderation

- Automated content moderation (research and practice)

- Investigation of content moderation broadly
[1]
- Investigation of automated content moderation

3 Setup

3.1 How is moderation organized?

3.2 Channel-level moderation

3.3 User-level moderation

4 Methodology

4.1 Data used

4.2 Automated testing

5 Experiments

5.1 Simulation setup

5.2 Research Questions and Results

6 Discussion

References

[1] M. Ozanne, A. Bhandari, N. N. Bazarova, and D. DiFranzo. Shall ai moderators be made visible? perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, 9(2):20539517221115666, 2022.

7 Appendix

1. Platform Exploration

2. Experimental Workflow

Bot Creation and Commands Twitch provides 'tmi.js', a Javascript library known as the Twitch Messaging Interface. tmi.js handles the bot's IRC protocol for instant messaging on Twitch Stream Chats. Each chatbot requires a supporting user account that must be verified via email and a valid phone number for 2FA as well as a user-account external app registration. This ensures that Twitch Bots are human-created and are for human-designed intentions such as mass moderation large streams. After implementing the Token Renewal system that ensures a stable connection to a specified channel, we implement specific functions for receiving commands that are declared by an *!<command>* and sending messages as a user. This allows us to activate specific bots implemented with differing commands such as *'!audit1', '!audit2'* for chat bots one and two.

Dataset Curation and Input We automate the process of parsing our curated datasets using a data formatting script that will convert csv to json files and standardize each dataset's heading to include 'text' as representative of the messages we will feed to our chat bots.

Audit Workflow <Insert Flowchart here>

3. Token Renewal

Method This section describes the process of utilizing user access tokens and renew tokens to stabilize and automate bot connections. Under Twitch Developer Documentation for authenticating external chat bots, user access tokens are required when permission to access user resources such as making a chat bot under the user's account. We use the Authorization code grant flow as opposed to the other manual user sign in or device-linked grant flows, to best automate process. The user access tokens generated via the POST method and cUrl command provide an expiry time of around four hours of which the renew tokens provided can help to update. With that we are able to utilize the exec package in JavaScript to run embedded user access tokens, renewal tokens, and the client-ids of the bot automatically.