Reducing the Spread of Racist Audio Messages on Discord

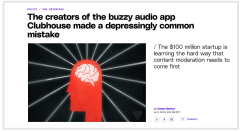
Team 7: Sonny Young, Matt Wolff, Ana Nguyen, Tyler Smith

CS152 / COMM122 / INTLPOL267: Trust and Safety

Stanford Computer Science

Objectives

- Hate speech, or offensive language targeting a group or an individual based on inherent characteristics[1], is on the rise since Elon Musk's takeover of the media platform X[2].
- **Nearly a third of internet users have experienced hate speech online.** Online platforms are an avenue for extremists to spread harmful ideologies.[3]
- Audio content–such as music, podcasts, and voice memos–pose unique moderation challenges due to its unstandardized nature, allowing hate speech to spread widely and often undetected.[4]



- **Goal:** to mitigate the dissemination of racist hate speech on Discord. Our content moderation system targets racist speech shared through audio memos and text messages on Discord servers.

Policy

Our platform is committed to upholding the principles of free expression, recognizing that the diverse perspectives of our users enrich our online community. Empathy and tolerance are crucial to ensuring every member of our community feels safe, respected, and valued.

As part of our platform's goal to create an inclusive environment, **we do not tolerate the promotion, advocacy, or incitement of hatred or violence**[6] against users based on their protected characteristics. These characteristics include, but are not limited to, race or ethnicity, gender identity or sexual orientation, religious affiliation, and disability status.

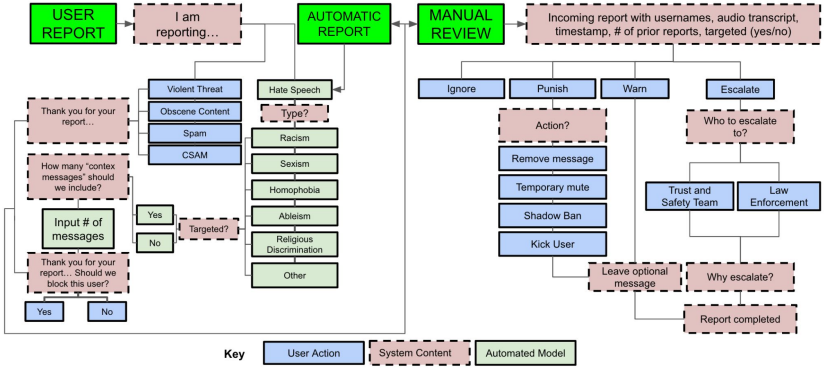
Enforcement Actions

If a user violates this policy, we reserve the right to take one or more of the following actions, depending on the severity and frequency of the violation:

- Immediate removal of reported message(s).
- Temporarily muting the user
- Reduction in the visibility of the user's content (shadowbanning)
- Suspension or termination of of user's account

Maintaining a respectful and inclusive community is a collective responsibility, and we appreciate the cooperation of all users in upholding these standards.

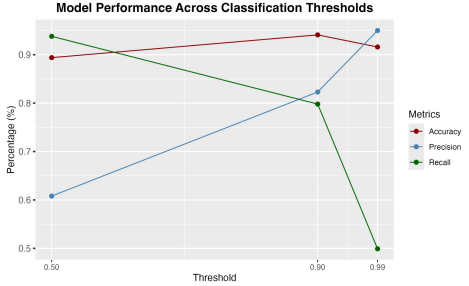
Reporting Flows



Technical Back-end

- **Model Architecture:** We fine-tuned the Qwen2.5-0.5B language model with a randomly initialized classification head to detect racist hate speech.
- **Training Data and Labeling:** We trained the model using the Measuring Hate Speech dataset produced by the UC Berkeley D-Lab.[5] Messages containing a racial target and a hate speech score above 0.5 was labeled 'racist.'
- **Audio Transcription:** We used OpenAI's Whisper model to transcribe audio files into text for analysis.

comment_id	platform	hate_speech_score	target_race_asian	target_race_black	target_race_latinx
47,777	3	-3.9	TRUE	TRUE	TRUE
22,819	2	-1.85	TRUE	FALSE	FALSE
39,276	2	-0.6	FALSE	TRUE	FALSE
48,140	3	0.23	FALSE	FALSE	TRUE



Threshold	Accuracy	Recall	Precision
0.5	0.894	0.938	0.608
0.9	0.941	0.798	0.823
0.99	0.916	0.499	0.95

Evaluation

- Our initial classification threshold of 0.5 resulted in many false positives, reflected in the roughly 61% precision score. Given the severity of punishments imposed by our auto-moderation system, this threshold was over-inclusive.
- We decided to increase the threshold to 0.99, significantly improving our precision by more than 21% to the detriment of recall.
- While our system may fail to flag some instances of hate speech, we prioritized minimizing wrongful punishments and only automatically penalizing messages that the model determines is highly likely to be harmful

Looking forward

- Integrate a LLM capable of reasoning through borderline cases, making more context-dependant moderation decisions.
- Incorporate Whisper's translation functionality to expand across languages. Even with these tools, understanding cultural nuances, regional slang, and implicit speech will require human oversight