Mitigating Cyberbullying on Instagram

Group 5: Daniel Shin[1], Eastan Giebler[2], Lauren Winkley[3], Saumya Goyal[4], Sophie Cline[5]

{dshin7[1], eastan[2], lwinkley[3], saumg[4], clines[5]}@stanford.edu

Stanford University

Stanford
Department of Computer Science/Politics

Problem Description

When a victim is cyberbullied, they are subjected to harassment online, often publicly through social media platforms. This harassment can encompass many behaviors including offensive name-calling, spreading false rumors, threats of physical harm, and sharing private information and images without consent[1]. We selected cyberbullying as our abuse type for this project because of its profound impact on social media users, especially young adults. Cyberbullying is incredibly prevalent among tween and teen users, and victimization rates have been steadily increasing over the past decade, with 59% of surveyed middle and high schoolers reporting having experienced some sort of cyberbullying as of 2023[2]. Additionally, being cyberbullied can have profound effects, with victims of cyberbullying being significantly more likely to think about, attempt, and complete suicide[3]. In other scenarios, cyberbullying can lead victims to harm others, as research has shown bullying linked to gun carrying amongst middle school and high school students[4].

Stanford University

Policy Language

Cyberbullying Guidelines

Cyberbullying is **strictly prohibited** on Instagram. This includes content that targets an individual with one or more threatening or abusive messages, doxxes or exposes private information about an individual, and/or shares one or more nonconsensual images of an individual with malicious intent.

Consequences of Violation

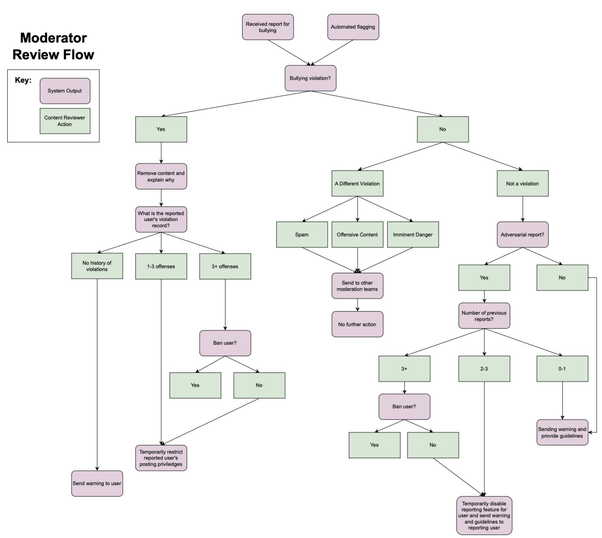
No prior offenses: Warning

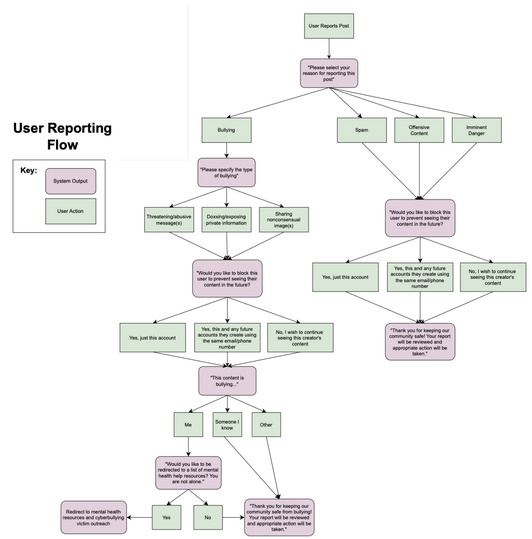
1-3 prior offenses: Posting capabilities temporarily suspended

3+ prior offenses: Eligible to be banned at discretion of moderator. If not banned, ability to post is temporarily suspended

Reporting

We encourage users to help keep our community safe by actively reporting cyberbullying. Users will be notified if the content they reported did not violate guidelines and sent the Community Guidelines for their reference. Users found by moderators to be adversarially reporting are at risk of consequences including suspension of reporting capabilities and permanent removal from the platform, based on prior number of offenses.





Technical Back-end

1. User and Moderator flow functionality implemented into bot
2. Two APIs: **gpt 4o** and **gemini-1.0-pro-vision-001**

Automated flagging using LLM(gpt 4o)

Inputs and outputs

- Inputs: post(image), policy language, comment(text and/or image), prompt
- Output: 'yes' or 'no' determining if a post is a violation

Multi-modal support

- Offensive comments
  - Ex: "trans men are not women"
- Offensive images/emojis
  - Ex: 🤢 🖕 🙈 🐷

Tailored Mental Health Resources using LLM(gemini 1.0)

Cyberbullying encompasses a wide range of abusive behaviors and the goal is to automatically provide relevant mental health resources

Inputs and outputs

- Inputs: post(image), list of resources, report summary, prompt
- Output: a list of mental health resources
  - Ex: The Trevor Project (LGBTQ+ Youth), Trans Lifeline

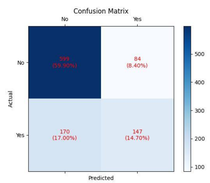
Providing violation explanation(gemini 1.0)

User may not know why their comment has violated platform policies. To protect all users, LLM sends violation explanation to reported users.

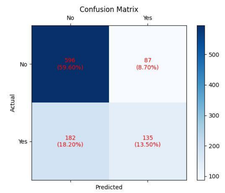
Inputs and outputs

- Inputs: post(image), comment(text and/or image), prompt
- Output: explanation on why a comment was auto-flagged/reported
  Ex: "The reported message violates our policy because it targets an individual (trans men) with a hateful and discriminatory message."

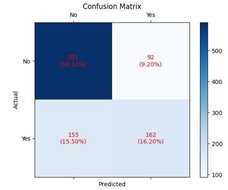
Auto-flagging Metrics



No policy - precision(63.64%), recall(46.34%)



Short policy - precision(60.81%), recall(42.52%)



Full policy - precision(63.78%), recall(51.10%)

We test three different policies to include in our prompt and ran our metric on twitter cyberbullying dataset. For cyberbullying, we prioritize precision instead of recall and we want to minimize false positives in order to not overload moderators. The **full policy** gives the **highest precision** and the **highest recall**, which is used during production.

Qualitative evaluation

Threatening/Abusive Messages

- Removes negative comments towards general users, but not public figures (expected behavior)
- Recognizes ambiguity in edge cases and has non-deterministic result based on context
- Attempts to account for sarcasm/humor - potential loophole for bad actors

Doxxing/Private Information

- Removes specific street addresses, but not large, general locations (expected behavior)
- Does not differentiate between public and non-public figures for doxxing posts (expected behavior)
- Has some biases in training – fails to recognize some specific foreign locations

Nonconsensual Images

- Removes images displaying nudity (expected behavior)
- Has trouble accounting for consensual nudity - could impact sex workers and others

Looking Forward

For the Future

- Consider public nature of addresses in doxxing cases/implement LLM that can actively search web and determine nature of address
- Train LLM further on worldwide addresses/landmarks to eliminate some of the current biases we discovered in testing
- Implement "buddy system" for automatic notification of a trusted support contact when a victim is cyberbullied

Sources:

[1] Vogels, E. A. (2022, December 15). Teens and Cyberbullying 2022. Pew Research Center. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/
[2] Patchin, J. W. (2024, February 16). Summary of Our Cyberbullying Research (2004-2022). Cyberbullying Research Center. https://cyberbullying.org/summary-of-our-cyberbullying-research
[3] Schonfeld, A., McNiel, D., Toyoshima, T., & Binder, R. (2024). Cyberbullying and Adolescent Suicide. The Journal of the American Academy of Psychiatry and the Law, 52(1). https://doi.org/https://doi.org/10.29158/JAAPL.220078-22
[4] Sumner, R., Ganz, M., Jacobs, M., Alessandro, C., Fuchs, D., Gamss, S., & Miller, D. (2022). Bullying Victimization as a Risk Factor for Gun Carrying Among US Adolescents. Cureus, 14(11). https://doi.org/10.7759/cureus.31785