Moderating and Automatic Flagging to Detect and Report Child Sextortion Online

Team 29: Blessing Opoku, Caroline Van, Fabian Luna, Lara Franciulli, Weston Kirk

Stanford
Computer Science/Policy

Problem Description

Sextortion is a crime in which an abuser threatens to expose intimate images or videos of their target unless the victim complies with their demands. Anyone can be a victim of sextortion, but children and teenagers are the most targeted. In 2023, Snap Inc. found that 65% of Gen Z teens have either been victims of online sextortion schemes themselves or have friends who have. Since 2022, there has been a significant rise in financially motivated sextortion for which victims are typically teenage males between the ages of 14 to 17. Based on data from the NCMEC and the FBI, we operate under the following assumptions:

- We are a direct messaging platform (Instagram, and SnapChat are the main platforms used for child sextortion today).
- Child sextortion has 3 stages:
  1. The abuser sends an intimate picture to the child.
  2. The abuser asks for an intimate picture of the child.
  3. The abuser threatens (either the child or someone else) to share intimate pictures of the child.
- Reports can be user-generated (our primary users of the report flow would be children or their guardians) or through our NLP model.
- All reports are manually verified by a moderator to comply with law enforcement.

Technical Backend

Backend Database: Stores information on reports and users.

- Reports generated by users stored in the 'Reports' table.
- User statistics on number of reports submitted by them + against them stored in 'UserStats'.

Report Priority Queue:

- Reports are assigned a priority when submitted by users.
- Priority values are based on the stage of the sextortion indicated by the reason of the report.
- Moderators can fetch the highest priority report from the 'Reports' table.
- Priority ties are broken by report timestamps.

User Statistics:

- 'UserStats' tables tracks the number of times a user has submitted a report, has been reported, has submitted false reports, or has been banned.
- Moderators can detect adversarial reporting and act accordingly.

Microsoft Azure NLP Model: Informs automated content flagging

- The Pan12 dataset provides labeled data for training and evaluation.
  - Documents submitted for model development entail real conversations
- Using Azure SDK to link the model to the bot, a "predatory" classification with >0.95 confidence deletes the message and submits a report to the moderator.

Policy on Child Sexual Exploitation, Abuse, and Nudity

We do not allow content or activity that sexually exploits or endangers children (i.e., individuals under 18).

If you violate this policy, your account will be banned and reported to the National Center for Missing and Exploited Children (NCMEC), in compliance with the law.

If you become aware of a violation of this policy, you should report it to us. To initiate a report, you can send "report" to the Team 29 Bot and follow the prompts. If you are under 18 and know or suspect intimate images of you have been leaked, visit Take It Down (takeitdown.ncmec.org) for help.

Violations of this policy include:

Child sexual exploitation: Content that threatens, depicts, praises, supports, provides instructions for, makes statements of intent, admits participation in, or shares links of the sexual exploitation of real or non-real children.

Inappropriate interactions involving children

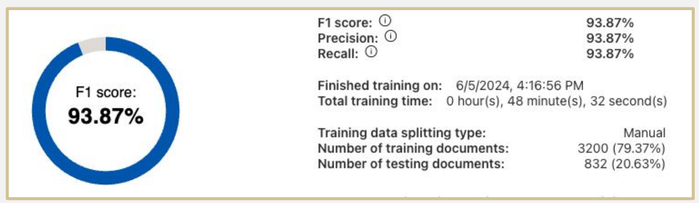
- Engage in sexual conversations with children that attempt to exchange sexual material or involve real-world encounters with them.
- Solicit sexual content involving real or non-real children.

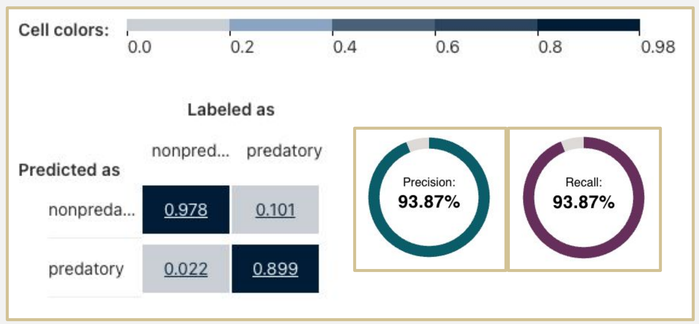
Exploitative intimate imagery and sextortion

- Coerce others for money, favors or intimate imagery with threats to expose real or non-real private sexual conversations or intimate imagery of children.

Evaluation/Metrics

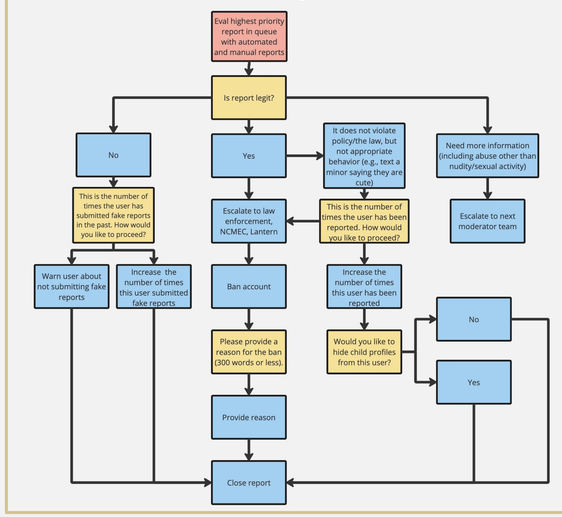
Below are evaluation metrics following training the NLP model on the PAN12 dataset.



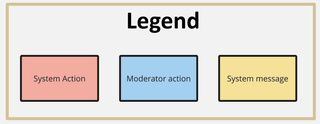


Reporting Flows

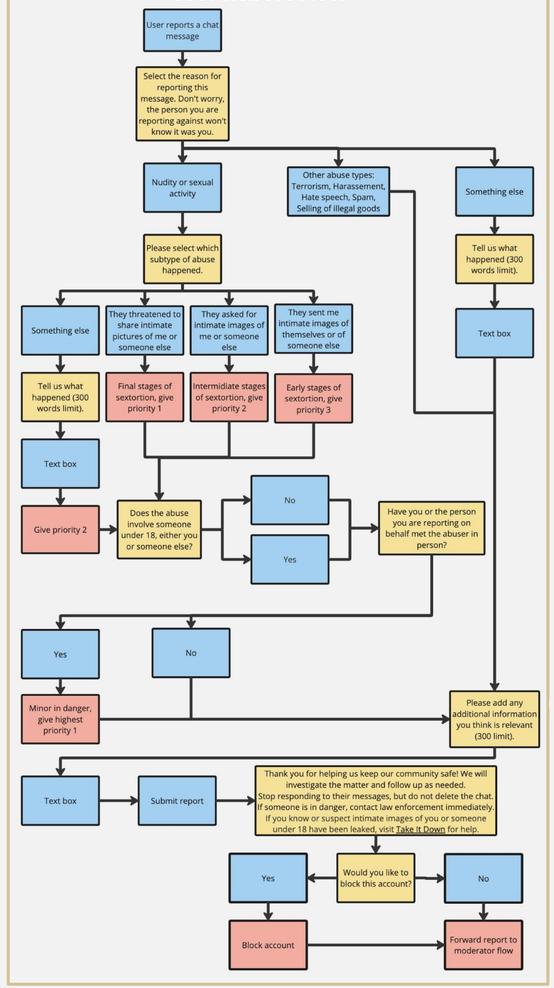
Moderator Report Flow



Legend



User Report Flow



Looking Forward

Limitations:

- Model optimizes for recall over accuracy, but having many false positives may overwhelm moderators.
- Deepfakes and AI-generated imagery can also overwhelm the system and de-prioritize reports involving real children.

Next Steps:

- Allow users to select multiple messages (e.g., text, audio and images) to provide moderators with more context on the conversation.
- Train a classifier to detect when CSAM is being sent and prevent its delivery.
- Detect whether abuser suggests to migrate to another platform that is less safe.
- Finetune the NLP model to better account for conversational nuances.
  - Language used by predators to gain victim's trust can be similar to that of a kind friend-to-be.
  - Adults sexting can resemble predatory language.

Bibliography/Notes:

We used Meta's Instagram report system and Community Standards regarding child sexual exploitation as a benchmark.

Instagram Updates to Prevent Sextortion: https://about.instagram.com/blog/announcements/new-tools-to-help-protect-against-sextortion-and-intimate-image-abuse.

NCMEC on Sextortion: https://www.missingkids.org/sextortion.

FBI on Child Sextortion: https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-scams-and-crimes/sextortion

Snap Inc. study: https://www.weprotect.org/blog/two-thirds-of-gen-z-targeted-for-online-sextortion-new-snap-research

PAN12 dataset to train NLP model: https://pan.webis.de/clef12/pan12-web/sexual-predator-identification.html