Leveraging New Language Models to Combat Sextortion on Discord

Avi Gupta
Computer Science, Stanford University

Cristobal Garcia
Computer Science, Stanford University

Riley Pittman
Computer Science, Stanford University

Sally Wang
Economics, Stanford University

Problem Description

Sextortion is a cybercrime associated with the extortion of money or sexual favors from someone threatening to reveal alleged evidence of some sexual activity.

Victims of sextortion are often teenagers and young adults. Both males and females can be victims, though young boys are increasingly targeted. They can be located anywhere in the world due to the proliferation of social media. Some common characteristics of victims are high usage of social media, low self-esteem, loneliness, and a trusting nature.

A common sextortion **procedure** is:

1. Extorter will befriend or catfish their victim in order to build trust.
2. Leverage that trust to obtain intimate photos and/or videos.
3. Threaten to release the content unless victim produces more intimate content.

Other **patterns** in sextortion messages include:

- Claims of Hacked Device: sender claims to have hacked victim's device to obtain intimate content.
- Ransom Demand: demand for payment, usually in cryptocurrency such as Bitcoin.
- Urgent Language: such as consequences of non-compliance.

Policy Language

Abuse Types: Our system is constantly searching for four high-level abuse types – sexual threats (sextortion), offensive content, spam, and urgent safety threats.

Enforcement: Automatically detect abusive content and automatically generate detailed reports for all abuse types using language models. Plus, a manual reporting flow for human-generated reporting and a content moderation flow.

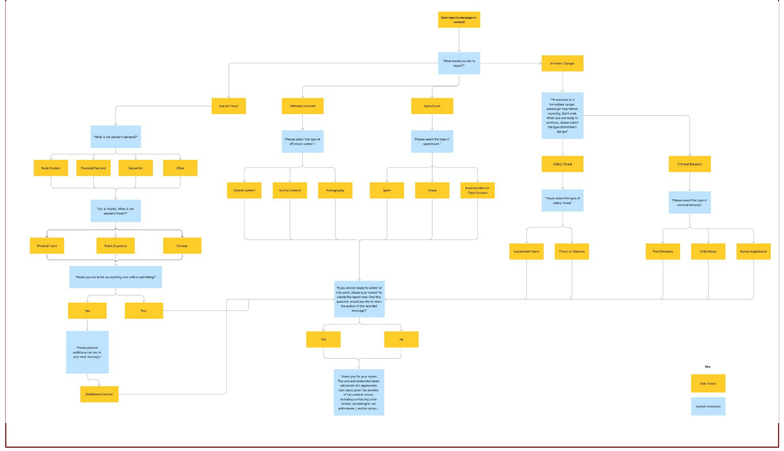
Reponses: Content moderation team will will categorize reports into three severity levels:

- **Tier 1 (Low Severity):** flag user for submitting fake report.
- **Tier 2 (Medium Severity):** delete message and kick user from platform.
- **Tier 3 (High Severity):** contact law enforcement.

Primary Policy on Sextortion: We prohibit any message that threatens a recipient with public shame or physical harm unless the sender is paid or given intimate content. Any message that falls into this category will immediately result in the user being banned from the platform (Tier 2), and law enforcement will be contact in cases of imminent danger (Tier 3).

No Warnings: Due to the severity of sextortion cases, we do not provide users with warnings before removing them.

Reporting Flow



Technical Back-end

Manual Reporting
Users can manually report sextortion, offensive content, spam, or imminent danger. However, recognizing that users may know how to report or may not feel comfortable reporting, we also implement automatic detection and report generation for all abuse types. Below, we discuss our technical process for detection and generation in the specific context of sextortion.

Automated Sextortion Detection and Report Generation
Goal: reduce barriers to reporting by automatically detecting sextortion and creating reports for content moderation review.

Implementation:

1. Run any *direct message* through a language model.
2. If sextortion is detected, create a report.
   a. Do so by ask the model the same series of questions that would be asked of a human reporting the message.
3. Publish the report in moderation channel for review by content moderators.

Model Comparison

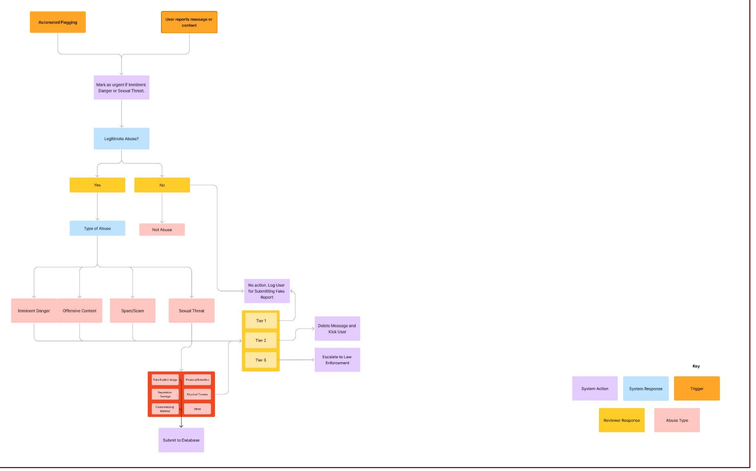
- Both Gemini Pro 1.0 and GPT 3.5 Turbo are highly capable and cost efficient text-based language models.
- Gemini gave us more control over safety and permissions, but GPT simpler to work with.
  - Gemini provides modifiable safety settings, all of which we had to override in order to invoke the model.
  - GPT uses API keys, where as Gemini uses more fine-grained IAM controls.

Model Prompting

- Our prompt to both language models was:
  "Please tell me if you detect any sextortion in the message below. Sextortion occurs when the message contains both a request for explicit material and a threat if the receiver does not comply. For example, asking for nude images alone is not sufficient for sextortion; the message must also include a threat, such as releasing potentially incriminating or sensitive content, or physical harm. Respond in the following format: Please only say 'yes' or 'no' to indicate if you detect sextortion. Here's the message: [insert reported message here]."

- The language models we used were very responsive to the prompts we gave them; small wording adjustments in the prompt made a noticeable difference.
  - Example 1: without providing the example in the prompt, the models would detect sextortion on simple requests for nude images. We do not want to overwhelm content moderators with safe flirtatious messages.
  - Example 2: using only the phrase 'sensitive images' would result in incorrect classification for video content. Modified prompt to 'potentially incriminating or sensitive content, or physical harm' to cast a wider net.

Moderation Flow



Metrics

Table 1. Classification Performance on Sextortion Detection Dataset

	Gemini	OpenAI
Accuracy	70.3%	84.2%
Precision	95%	100%
Recall	39.5%	66.7%
F1-Score	55.9%	80%

Table 2. Gemini Confusion Matrix

n=101	Predicted: Non-Sextortion	Predicted: Sextortion
Actual: Non-Sextortion	52	1
Actual: Sextortion	29	19

Table 2. OpenAI Confusion Matrix

n=101	Predicted: Non-Sextortion	Predicted: Sextortion
Actual: Non-Sextortion	53	0
Actual: Sextortion	16	32

Key Takeaways:

- The OpenAI GPT 3.5 Turbo model outperformed Gemini Pro 1.0 in all aspects.
- The dataset was generated by the Claude 3 Sonnet model from Anthropic AI
  - Of the 101 messages, 48 contained content related to sextortion and 53 did not
- When either model says that a message contains sextortion, they are extremely accurate (the precision score)
- There are almost no false positive predictions for either model, but there are false negatives. This is an acceptable result, as the user may create a manual report in these situations.

Qualitative Evaluation

Automation for Sexual Threats:

- Our most detailed reports are produced for sexual threats (sextortion).
- Using an LLM, we include the sender's demand, the threat, and additional context for the report reason.
- This report is then sent to the content moderation team.

Examples and Outcomes:

- Basic Sextortion: "I have compromising photos of you, pay me or I'll release them."
  - Both models correctly identify sextortion.
- In some messages without explicit sexuality, both models incorrectly predicted they did not contain sextortion.
  - "Pay me $500 in bitcoin or I'll release your personal photos."
  - "I have a video of you, pay me or it will go public.
- Though overall metrics were similar, models made different predictions in some straightforward cases.
  - "Send me money if you don't want your private videos leaked." – Gemini predicted no, but OpenAI said yes.
  - "Pay $700 or I'll send your explicit photos to all your contacts." – Gemini predicted yes, but OpenAI said no.

Takeaway: Language models are powerful, but certainly not bulletproof! Attackers can find workarounds.

Looking Forward

Impact: Currently, we believe our system will have a positive impact on our community's safety. We expect our technical implementation to reduce the number of potential sextortion abuses and victims through automatic and manual moderation.
Improvements:

1. At scale, finding a more cost-effective technologies. Running every DM through a third-party LLM is not economical.
2. Implementing shadow banning to avoid retaliation from perpetrators.
3. Reduce inaccurate user reporting, though ways such as but not limited to providing legitimate sample reports
4. Addressing false user reporting, through ways such as but not limited to revoking the user's reporting functionality and further disciplinary action.
5. Explore more sophisticated automatic detection mechanisms to intercept sextortion at an earlier stage, such as before a threat is made, without overwhelming content moderators with benign or flirtatious messages.

Stanford | ENGINEERING
Computer Science

Contact

Avi Gupta
Computer Science
Stanford University
agupta07@stanford.edu

Cristobal Garcia
Computer Science
Stanford University
cgarcia0@stanford.edu

Riley Pittman
Computer Science
Stanford University
rileywp@stanford.edu

Sally Wang
Economics
Stanford University
sallyw02@stanford.edu

References

1. Markham, Kevin. "Simple Guide to Confusion Matrix Terminology." *Data School*, Data School, 25 Mar. 2014, www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.