Mitigating Online Sextortion through Automated Message Classification

Team 11: Allie Griffith Ethan Katz Neil Shen Carmel Limcaoco Benny Pan

Problem Description

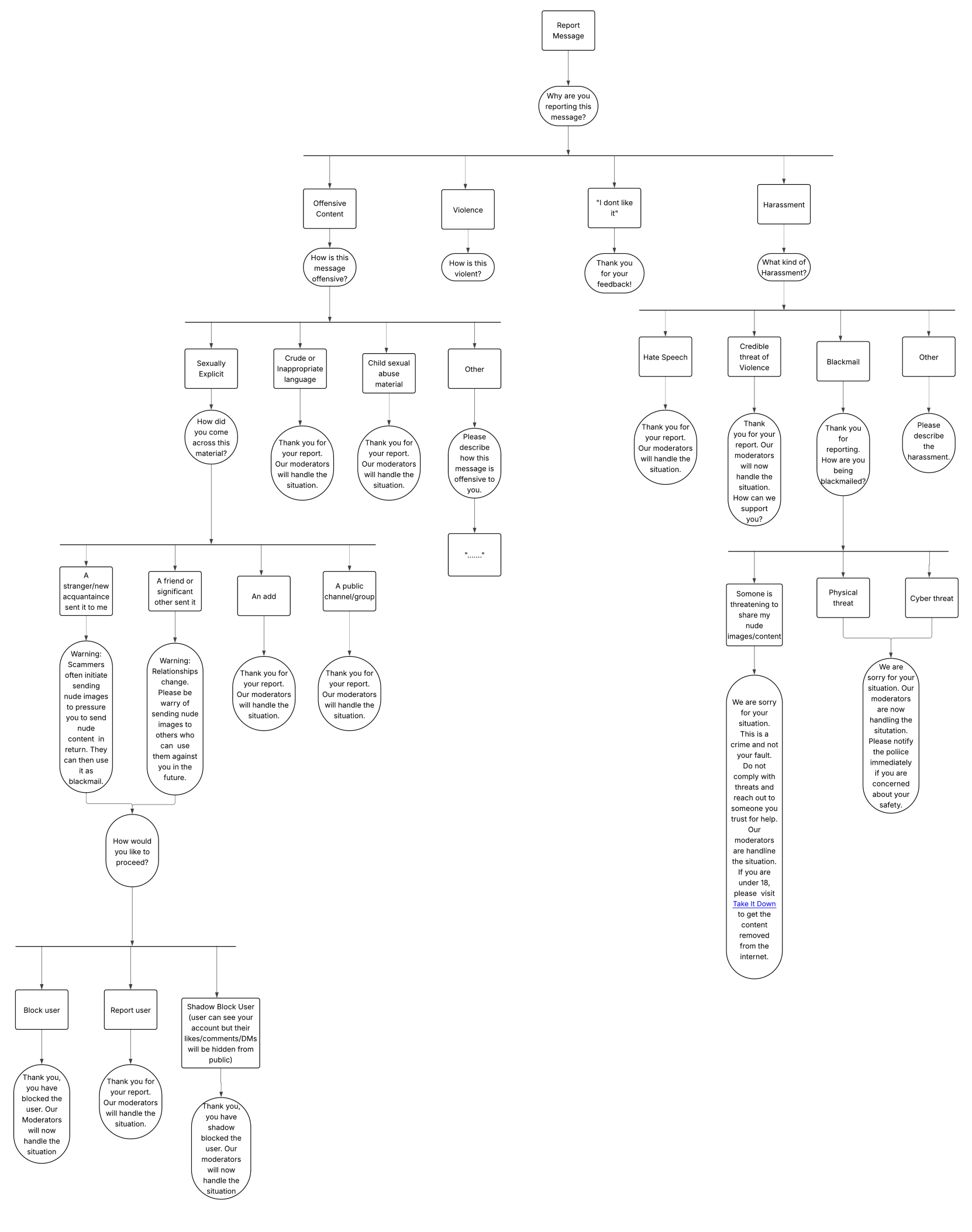
Sextortion is characterized by a perpetrator threatening to share a victim's sexually explicit content unless the victim makes financial payments or complies with their demands. Sextortion is an **increasingly prevalent** crime that can be extremely **traumatizing, stressful, and potentially life threatening to its wide range of victims**. Since a large portion of sextortion crimes are perpetrated by repeat offenders, prudent moderation to identify criminals is especially fruitful. To combat sextortion we implement a comprehensive and clear reporting flow for every stage of the sextortion life cycle. By issuing early warnings, connecting victims to help resources, allowing users to block/shadow block perpetrators, automatically detecting potential perpetrators and manually reviewing elevated cases, we hope to mitigate the harms of sextortion and protect platform users.

Policy Language

Our platform is a place where users should feel safe and protected. We therefore have a zero-tolerance policy for sextortion (demanding money or acts in exchange for not sharing intimate images). To work towards this goal, all messages will run through our automatic classification bot to flag potential instances of sextortion and take decisive actions in more severe cases. Sextortion often starts small and then escalates as the malicious actor gains traction and leverage over the user, so our policy is to catch cases early and prevent the situation from getting to a place where the user feels powerless. We do this by warning users against the dangers of sharing intimate content when our bot has moderate confidence there are early signs of sextortion. While sending nude images is allowed on our platform, users will receive warnings about the dangers of sharing nude images to encourage thoughtfulness.

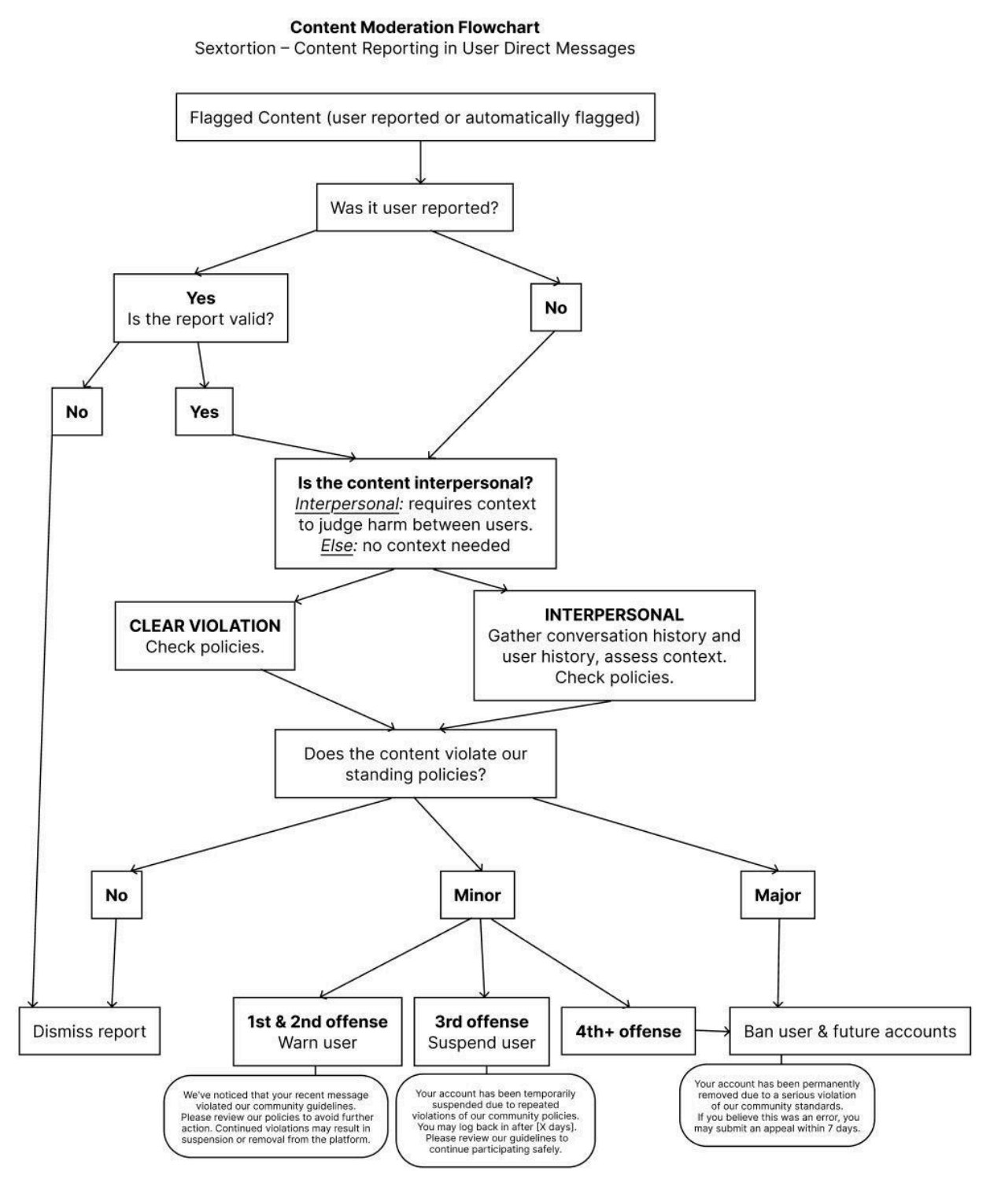
Messages that fall into categories that are common to sextortion, including a combination of requests to send money, threats, and mentions of having sexually explicit content of the user, will be automatically flagged as severe instances of sextortion by our classifier. We will warn the user that they are potentially being sextorted, send the user advice and help resources (such as NMEC's Take It Down service) and initiate a manual (human) moderation reporting flow. Once reported, our human moderators will review the message and determine its severity. In minor offenses, we will issue a warning to the perpetrator. After three minor offenses, the perpetrator will be suspended. After four minor offenses or one major offense, we will permanently ban the perpetrator and all of their future accounts.

Manual Reporting Flows



Report Message
Why are you reporting this message?
Offensive Content
How is this message offensive?
Sexually Explicit
How did you come across this material?
A stranger/new acquantaince sent it to me
Warning: Scammers often initiate sending nude images to pressure you to send nude content in return. They can then use it as blackmail.
A friend or significant other sent it
Warning: Relationships change. Please be warry of sending nude images to others who can use them against you in the future.
How would you like to proceed?
Block user
Thank you, you have blocked the user. Our Moderators will now handle the situation
Report user
Thank you for your report. Our moderators will handle the situation.
Shadow Block User (user can see your account but their likes/comments/DMs will be hidden from public)
Thank you, you have shadow blocked the user. Our moderators will now handle the situation
An add
Thank you for your report. Our moderators will handle the situation.
A public channel/group
Thank you for your report. Our moderators will handle the situation.
Crude or Inappropriate language
Thank you for your report. Our moderators will handle the situation.
Child sexual abuse material
Thank you for your report. Our moderators will handle the situation.
Other
Please describe how this message is offensive to you.
"......"
Violence
How is this violent?
"I dont like it"
Thank you for your feedback!
Harassment
What kind of Harassment?
Hate Speech
Thank you for your report. Our moderators will handle the situation.
Credible threat of Violence
Thank you for your report. Our moderators will now handle the situation. How can we support you?
Blackmail
Thank you for reporting. How are you being blackmailed?
Somone is threatening to share my nude images/content
We are sorry for your situation. This is a crime and not your fault. Do not comply with threats and reach out to someone you trust for help. Our moderators are handline the situation. If you are under 18, please visit Take It Down to get the content removed from the internet.
Physical threat
Cyber threat
We are sorry for your situation. Our moderators are now handling the situtation. Please notify the poliice immediately if you are concerned about your safety.
Other
Please describe the harassment.

User Reporting Flow



Content Moderation Flowchart
Sextortion – Content Reporting in User Direct Messages
Flagged Content (user reported or automatically flagged)
Was it user reported?
Yes
Is the report valid?
No
Yes
No
Is the content interpersonal?
Interpersonal: requires context to judge harm between users.
Else: no context needed
CLEAR VIOLATION
Check policies.
INTERPERSONAL
Gather conversation history and user history, assess context.
Check policies.
Does the content violate our standing policies?
No
Minor
Major
Dismiss report
1st & 2nd offense
Warn user
We've noticed that your recent message violated our community guidelines. Please review our policies to avoid further action. Continued violations may result in suspension or removal from the platform.
3rd offense
Suspend user
Your account has been temporarily suspended due to repeated violations of our community policies. You may log back in after [X days]. Please review our guidelines to continue participating safely.
4th+ offense
Ban user & future accounts
Your account has been permanently removed due to a serious violation of our community standards. If you believe this was an error, you may submit an appeal within 7 days.

Moderator Reporting Flow

Automatic Detection

For **proactive** mitigation of sextortion at **scale**, we trained an AI classifier to estimate the probability that a given message was sent by a sextortion perpetrator. We then integrated this classifier into our moderation flow: when it detects signs of Sextortion, the bot issues a helpful warning to the user; if the confidence exceeds a certain threshold, it triggers a manual reporting workflow to ensure human oversight.

Data

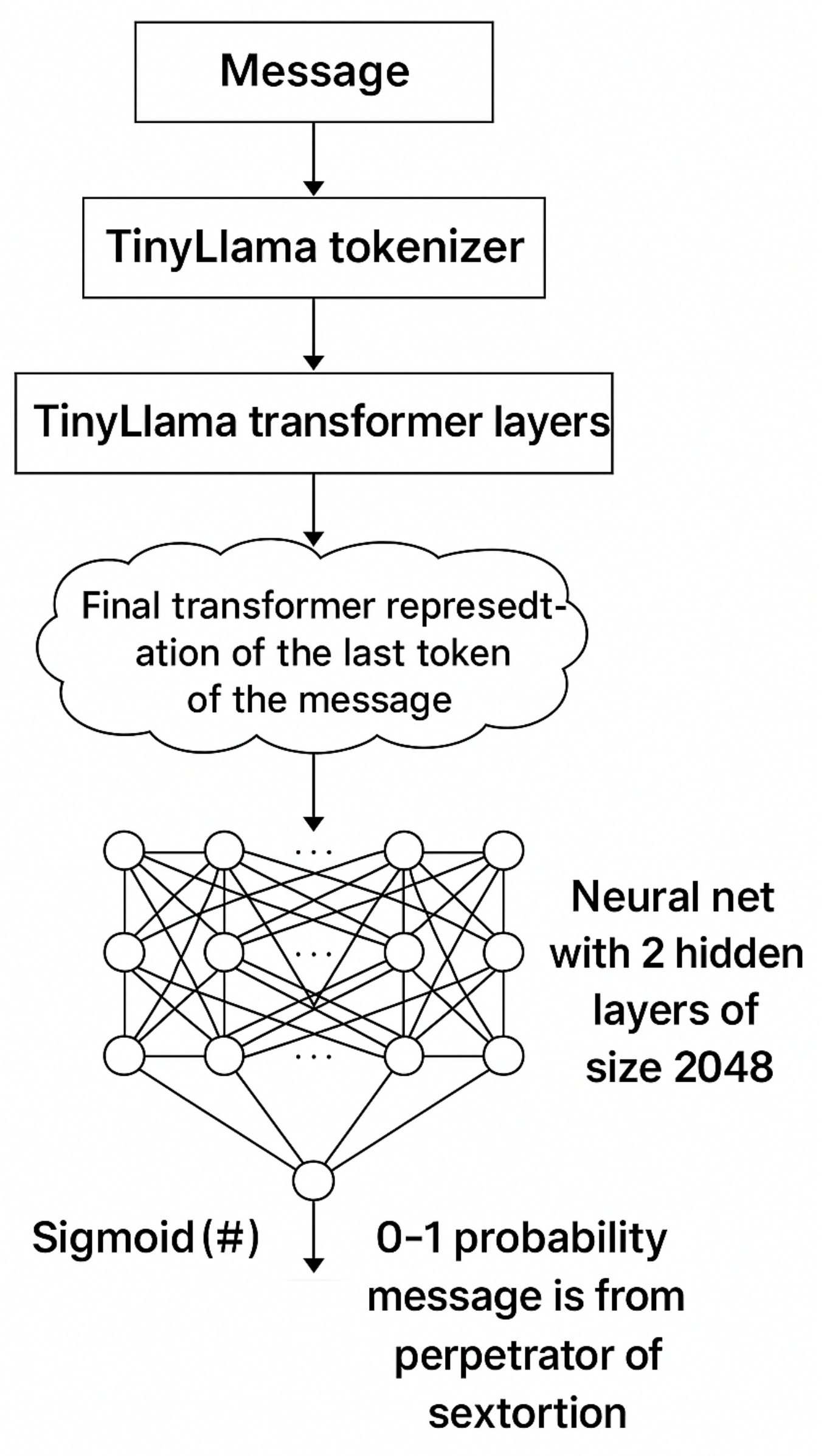
Our team utilized a synthetically generated dataset of 400 messages from ChatGPT-4o. Each message is labeled as an instance of sextortion or not sextortion, and includes message variations where characters are obfuscated and non-text characters are used. We manually read through the dataset to ensure it was sound and reasonable. We then divided the data into training and evaluation sets using 80-20 splits. Examples of the synthetic data used for bot training can be found below.

message	label
Thereâ€™s no turning back now. Your files and videos have been downloaded. Everyone will know your secrets. Do not reply. Just pay $1100.	sextortion
Urgent: Pay your toll online to avoid extra fees.	not sextortion
This is your final warning. We took over your camera while you were online. We will make you a public example. $1400 in BTC is the only way out. ðŸ˜³	sextortion
Happy birthday! Hope you have an amazing day.	not sextortion

Table 1. Example of training messages and their corresponding labels

Designing and Training the Classifier Bot

Model Architecture

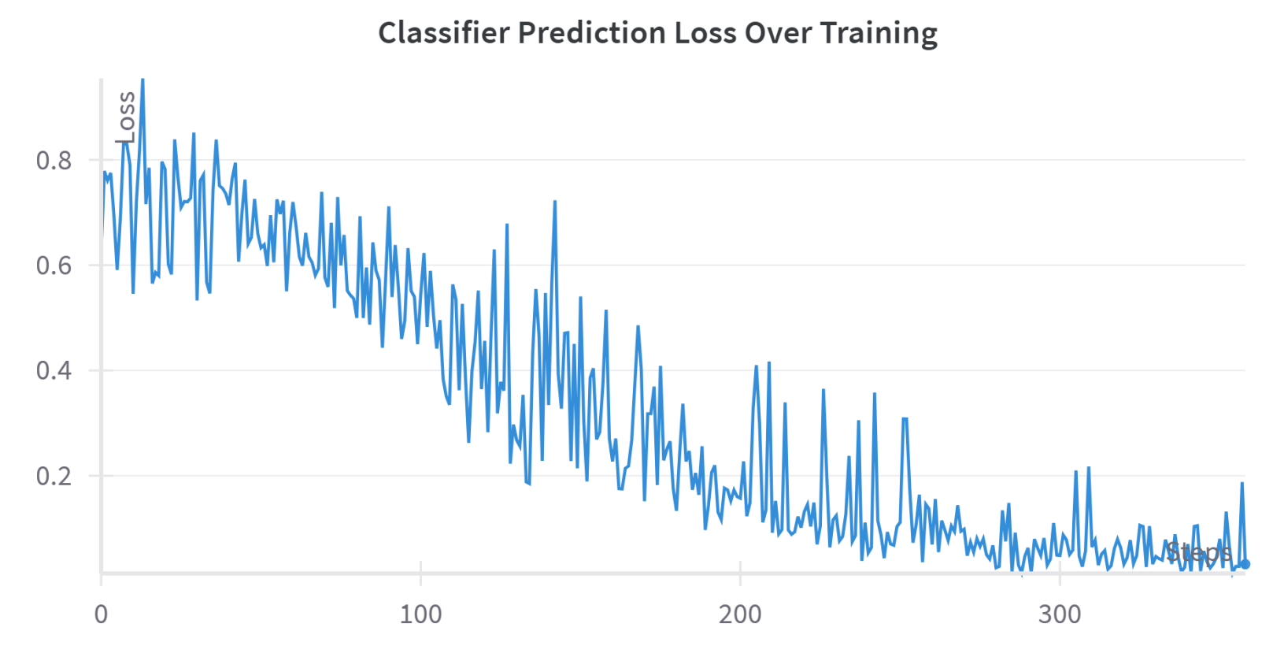


Message
TinyLlama tokenizer
TinyLlama transformer layers
Final transformer represent-ation of the last token of the message
Neural net with 2 hidden layers of size 2048
Sigmoid (#)
0–1 probability message is from perpetrator of sextortion

(a) Model Architecture.

Training

We train the neural network classifier with the training split of the data (360 examples/steps). The bot is given TinyLlama's final transformer representation of the final token of each training message. The bot predicts a score and receives a loss from the true label using a built-in PyTorch loss function for binary classification. We perform gradient descent using the Adam optimizer to update neural net weights. By 300 training steps the bot converges to nearly 0 loss, signaling successful training.

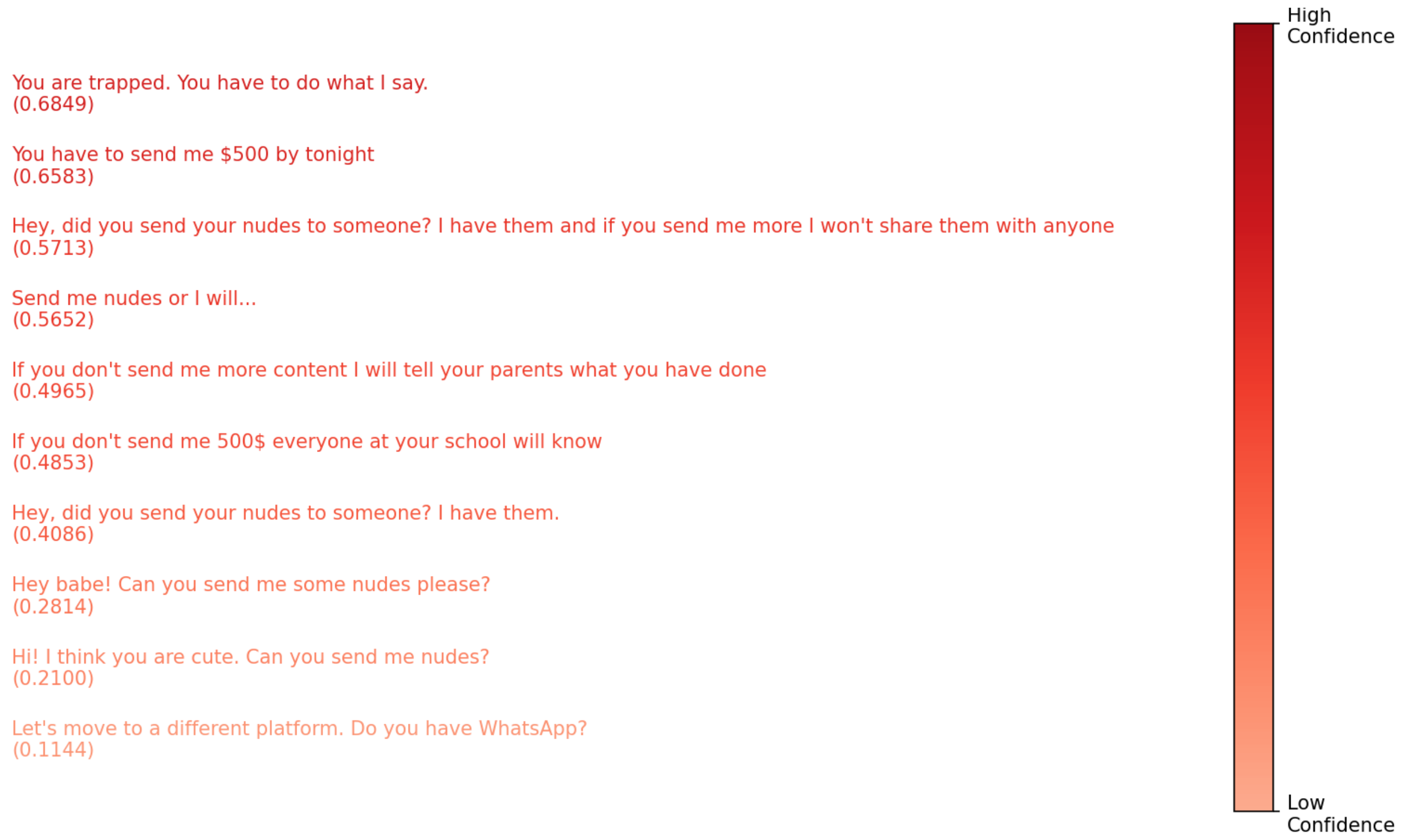


(b) During supervised training, the neural-network classifier exhibits classic learning behavior: its classification loss steadily decreases toward zero over 360 training iterations.

Automatic and Manual Evaluation

On the test split of the sextortion dataset, our bot achieved 100% accuracy with perfect recall and precision.

To further assess our model's performance, we tested it on a set of manually curated examples exhibiting varying degrees of sextortion indicators. The results (shown below) offer valuable guidance for selecting a confidence threshold that balances precision and recall on the moderator side.



You are trapped. You have to do what I say.
(0.6849)

You have to send me $500 by tonight
(0.6583)

Hey, did you send your nudes to someone? I have them and if you send me more I won't share them with anyone
(0.5713)

Send me nudes or I will...
(0.5652)

If you don't send me more content I will tell your parents what you have done
(0.4965)

If you don't send me 500$ everyone at your school will know
(0.4853)

Hey, did you send your nudes to someone? I have them.
(0.4086)

Hey babe! Can you send me some nudes please?
(0.2814)

Hi! I think you are cute. Can you send me nudes?
(0.2100)

Let's move to a different platform. Do you have WhatsApp?
(0.1144)

Figure 2. Manual evaluations of bot showed accurate and granular scores

Strengths:

- Lightweight architecture (1.1B parameters) is much smaller than modern LLM solutions.
- Fast classification that takes fractions of a second.
- Granular confidence scores allow moderators to control what thresholds trigger certain responses.
- Though manual evaluation found classification errors, the model aligns with human evaluations over 90% of the time.

Weaknesses:

- Requires a GPU to "encode" messages with TinyLlama
- Fails to detect some obfuscated harmful messages (ie $end nude$)

Implementing Automatic Moderation

We established score thresholds to determine the course of action the bot takes once a message has been classified. If the message's score is 0.4 or less, our bot does not take action. If the score is within the range of 0.4-0.5, our bot will send a cautionary message to the user, warning them of the harms of sending sensitive content to unknown users and urging them to reconsider before proceeding. We have chosen a relatively low threshold for the warning message to optimize for recall over precision, as we'd like to lean on the safer side to ensure we catch as many potentially harmful messages as possible. If the score is >0.5, the warning is sent to the user, and the conversation is flagged for manual content moderation so an intervening human can evaluate the context of the message to determine if the user is in danger.

Looking Forward

Our classifier provides a promising foundation for detecting sextortion. Future improvements that could be implemented are exploring multi-modal inputs, like images or voice messages, or fine-tuning with real-world messages that may be obfuscated or include nuanced slang. The system itself could be improved by taking human moderator decisions as input in a feedback loop to continuously retrain the model. Ultimately, advancing tools like our model is a step toward integrating proactive detection and thoughtful intervention to meaningfully reduce the harm caused by sextortion.