Preventing Cryptocurrency Scams through Automatic and Manual Moderation

Stanford University

Team 16: Ben Alexander[1], Tracey (Xiyuan) Chen[1], Gordon Downs[1], Silvia Gong[1], Valeria Gonzalez[2]

CS152: Trust and Safety Engineering / POLISCI 243C: The Politics of Internet Abuse

[1]Department of Computer Science, Stanford University, [2]Department of Political Science, Stanford University

Problem Description

Cryptocurrency is popular among scammers because crypto transactions are largely anonymous, irreversible, unregulated, and easily accessible across international borders. Not only does it pose relatively little risk to the scammers, but it also requires fairly minimal technical capabilities. Crypto scam victims can be anyone with Internet access, especially young, Internet-savvy, and active social media users. Many victims are exposed to crypto scams on social media before they understand the market, technology, and scam environment, so it is especially important for our platform to support them by providing anti-scam moderation. In this project, we primarily address "social engineering" scams such as giveaway scams ("Free Bitcoin giveaway!") and extortion scams ("Send money to this address …") carried out through Discord DMs.

Cryptocurrency Scam Policy

Overview:

We want our platform to be a place where community is built, rather than broken. Therefore, we prohibit any content that in any way facilitates, solicits, promotes, or encourages, the deception of others for personal financial gain.

With the rise in popularity of cryptocurrency, we specifically warn users of scams related to crypto. Common crypto scams and fraudulent behavior includes, but is not limited to:

- *Giveaway scams*
- *Phishing/fake URL scams*
- *Extortion/blackmail scams*
- *Pump-and-dump scams*

Our platform is taking on various measures to prevent such harmful behavior from existing on our platform. These measures include auto-flagging scam content using machine learning and other fraud detection tools, responding to user reports, and performing human moderation.

If you violate this policy:

Depending on the severity of the violation and/or the previous history of violations, we reserve the right to:

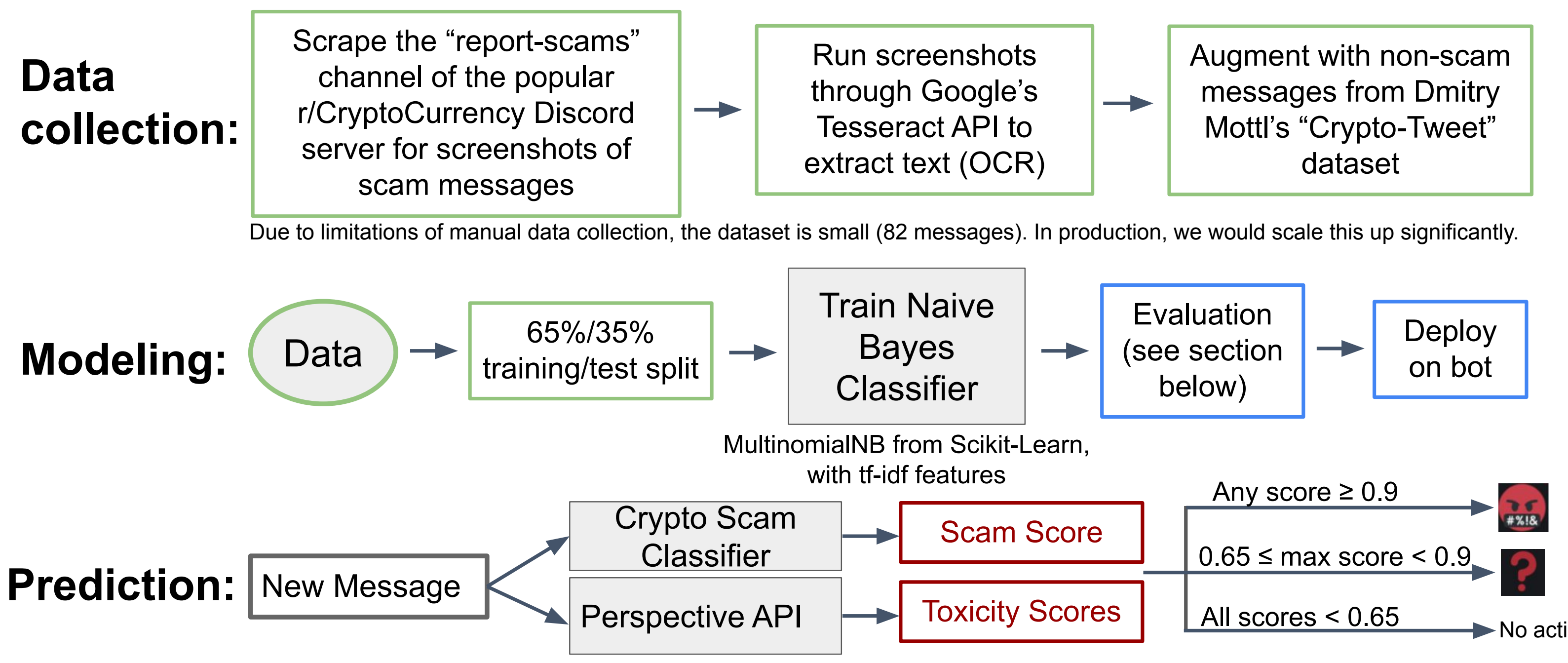
1. Remove (or gray out and add warnings on top of) messages containing crypto scam content.
2. Blacklist and provide warnings about URLs and Bitcoin addresses we believe to be fraudulent. If you believe your URL or address has been wrongly blacklisted, you may make an appeal.
3. Warn and/or temporarily suspend offending accounts.
   a. We have a "three strikes" policy. For the first two violations, we warn and temporarily suspend offending accounts for a certain amount of time, depending on the severity.
4. Permanently suspend offending accounts.
   a. For the third violation, we permanently suspend the account. Note that we do not permanently suspend accounts based on automatic flagging; the three strikes must come from human moderators.

How you can help:

If you see content that you believe to be a scam, or is otherwise fraudulent, please follow our user reporting flow to report the content to moderators. We can all work together to keep our community safe!

Technical Back-end

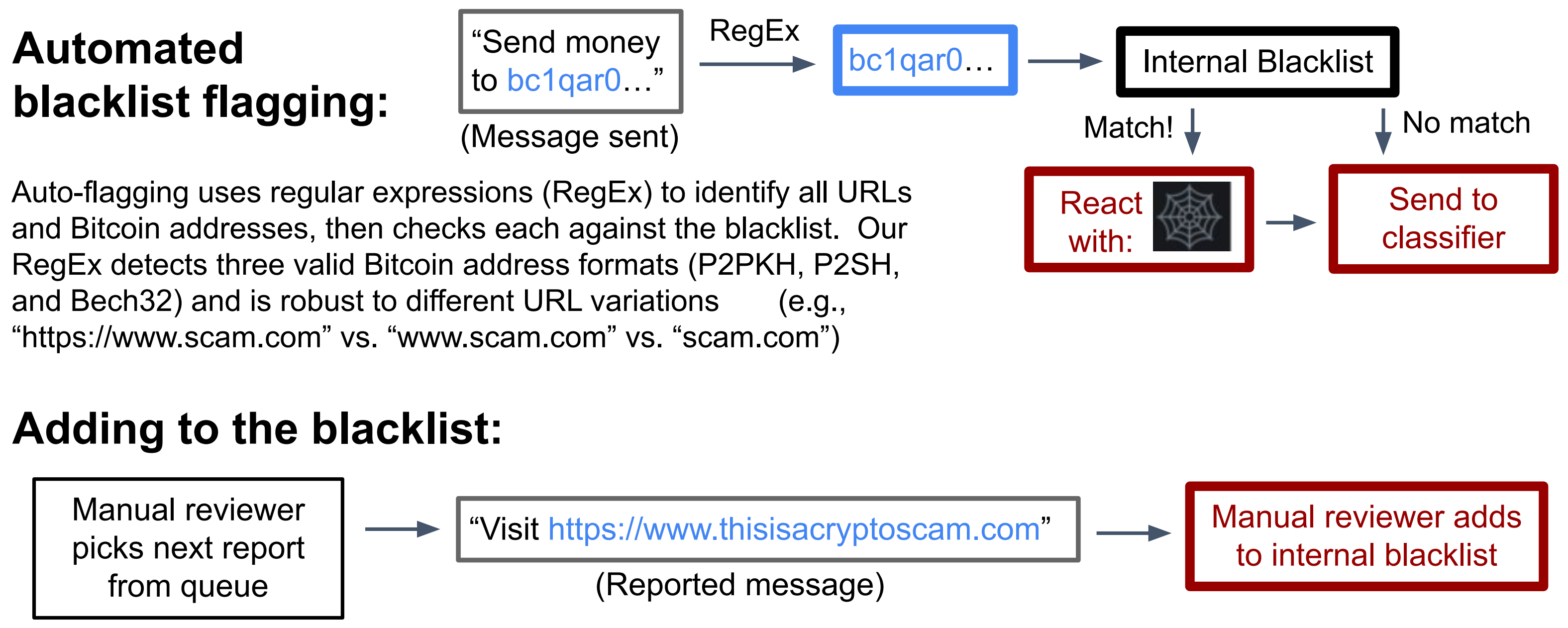
Auto-flagging: Cryptocurrency Scam Classifier


Data collection:
Scrape the "report-scams" channel of the popular r/CryptoCurrency Discord server for screenshots of scam messages
Run screenshots through Google's Tesseract API to extract text (OCR)
Augment with non-scam messages from Dmitry Mottl's "Crypto-Tweet" dataset
Modeling:
Data
65%/35% training/test split
Train Naive Bayes Classifier
Evaluation (see section below)
Deploy on bot
Prediction:
New Message
Crypto Scam Classifier
Perspective API
Scam Score
Toxicity Scores
Any score ≥ 0.9
0.65 ≤ max score < 0.9
All scores < 0.65
No action

Due to limitations of manual data collection, the dataset is small (82 messages). In production, we would scale this up significantly.

MultinomialNB from Scikit-Learn, with tf-idf features

Auto-flagging: Internal Blacklist (scam Bitcoin addresses + URLs)


Automated blacklist flagging:
"Send money to bc1qar0…"
(Message sent)
RegEx
bc1qar0…
Internal Blacklist
Match!
No match
React with:
Send to classifier
Adding to the blacklist:
Manual reviewer picks next report from queue
"Visit https://www.thisisacryptoscam.com"
(Reported message)
Manual reviewer adds to internal blacklist

Auto-flagging uses regular expressions (RegEx) to identify all URLs and Bitcoin addresses, then checks each against the blacklist. Our RegEx detects three valid Bitcoin address formats (P2PKH, P2SH, and Bech32) and is robust to different URL variations (e.g., "https://www.scam.com" vs. "www.scam.com" vs. "scam.com")

Adversarial Cases

Our system handles several adversarial cases:

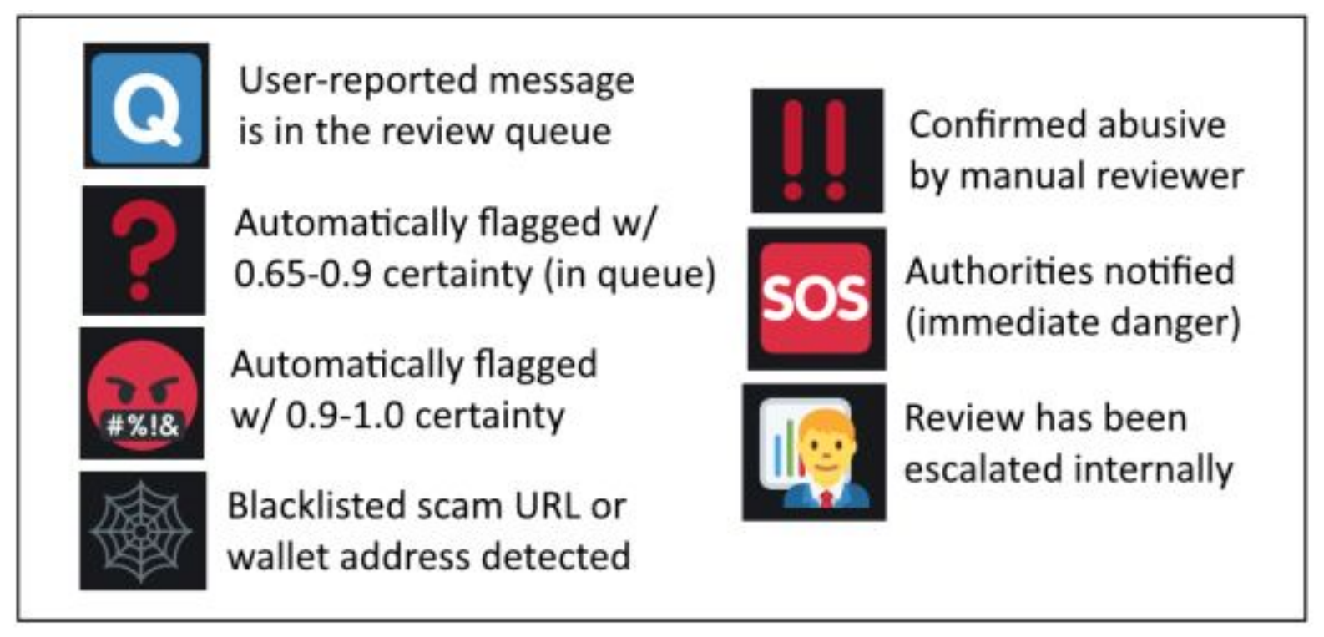
1. If a user disguises a scam message in unicode characters, we still recognize it by passing it through the "unidecode" library to convert it to regular English text.

   e.g., 𝓫ɪᴛⒸ𝑜𝕀∩ → bitcoin

2. If a user edits an innocuous message to something abusive, we run the automatic flagging tools on the edited message again, so that we detect the new abusive message.
3. The Perspective API can be tricked by strange casing, like "eXaMPle." So, we run the lowercase version of each message (in this case, "example") through the API to avoid this problem.

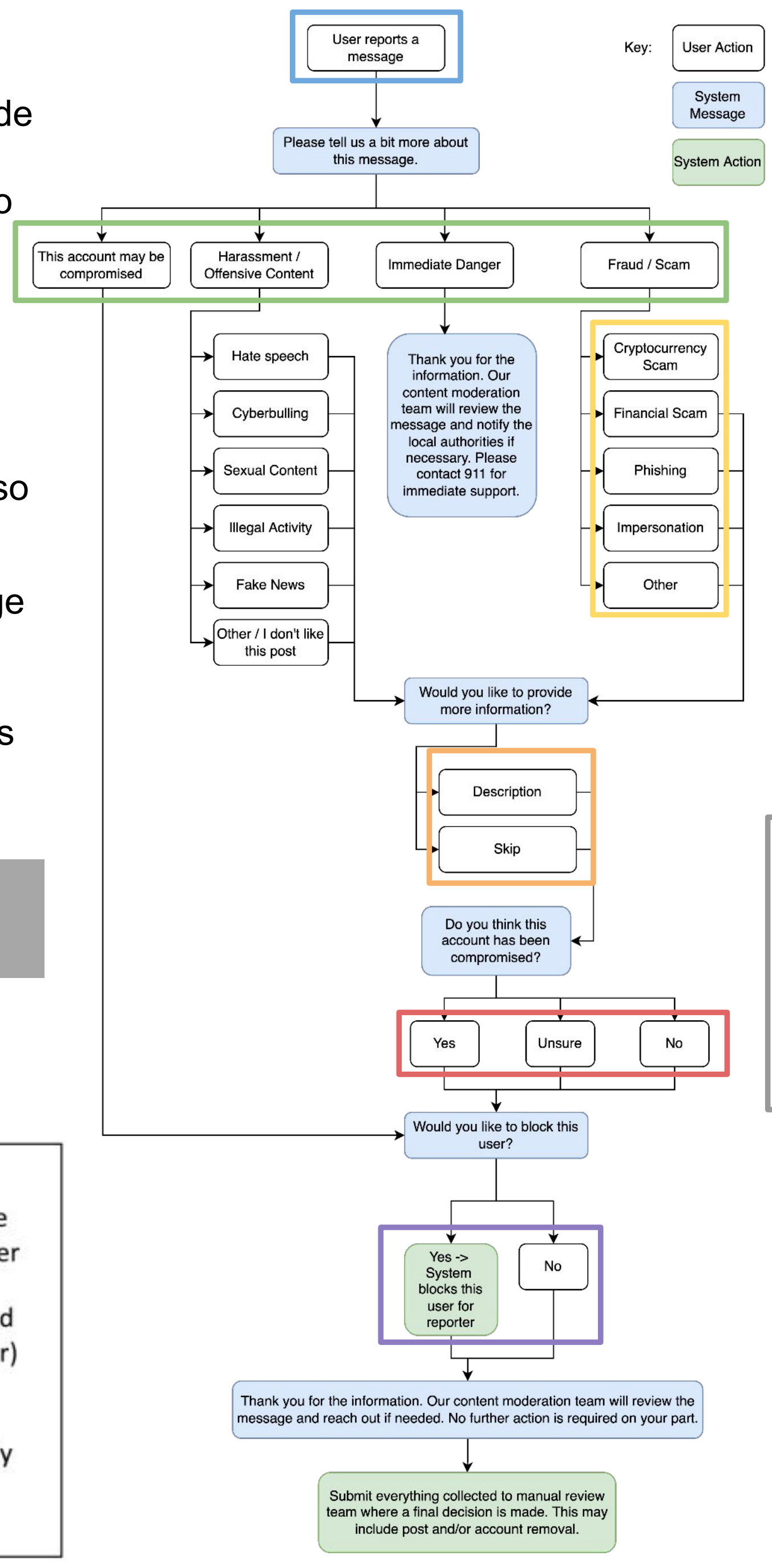
Emojis/Reactions (Legend)

These are the emojis that we use to represent different reactions in our platform:



- Q User-reported message is in the review queue
- ? Automatically flagged w/ 0.65-0.9 certainty (in queue)
- 🤬 Automatically flagged w/ 0.9-1.0 certainty
- 🕸 Blacklisted scam URL or wallet address detected
- ‼ Confirmed abusive by manual reviewer
- SOS Authorities notified (immediate danger)
- 🧑‍💼 Review has been escalated internally

Manual Reporting: User Reporting Flow


User reports a message
Key:
User Action
System Message
System Action
Please tell us a bit more about this message.
This account may be compromised
Harassment / Offensive Content
Immediate Danger
Fraud / Scam
Hate speech
Cyberbulling
Sexual Content
Illegal Activity
Fake News
Other / I don't like this post
Thank you for the information. Our content moderation team will review the message and notify the local authorities if necessary. Please contact 911 for immediate support.
Cryptocurrency Scam
Financial Scam
Phishing
Impersonation
Other
Would you like to provide more information?
Description
Skip
Do you think this account has been compromised?
Yes
Unsure
No
Would you like to block this user?
Yes -> System blocks this user for reporter
No
Thank you for the information. Our content moderation team will review the message and reach out if needed. No further action is required on your part.
Submit everything collected to manual review team where a final decision is made. This may include post and/or account removal.

Example User Report:

```
report_time: 2022-03-06 11:55:31
mod_report:
    reporter: gordizzle
    message:
        author: silvgong
        content: Congrats! You have been
                 selected for a free bitcoin
                 giveaway!
    category: Fraud / Scam
    sub-category: Cryptocurrency Scam
    justification:
        I think it is a suspicious message.
        Looks like a giveaway scam!
    account_status: Reported may be
                    compromised.
    user_action: Reporter blocked silvgong.
```

```
scores:
    PROFANITY: 0.015,
    IDENTITY_ATTACK: 0.020,
    TOXICITY: 0.048,
    SEVERE_TOXICITY: 0.015,
    THREAT: 0.027,
    CRYPTO_SCAM: 0.711
```

```
auto_flagged: false
reported_account_abusive_strikes: 0
reporter_account_malicious_strikes: 0
```

Use machine learning to automatically detect if the message is abusive or crypto scam related.

Manual Review Process

Review Priority Queue:

1) Messages involving immediate danger are given the highest priority; others are ranked by reporting time (oldest to newest).
2) Always review the message at the front of the queue next.
3) Must finish processing each message before starting reviewing a new one.

Malicious/Frivolous User Report Outcomes:

1) Warn: remove Q / ?; automatically DM the reporter with a warning. (≤ 5 malicious/frivolous reports)
2) Suspend: remove Q / ?; temporarily disable reporting feature for the malicious reporter; automatically DM the reporter with a suspension notification. (> 5 malicious/frivolous reports)

Message with Immediate Danger:

Q / ? → SOS

Message to be Escalated to Higher-Level Reviewers:

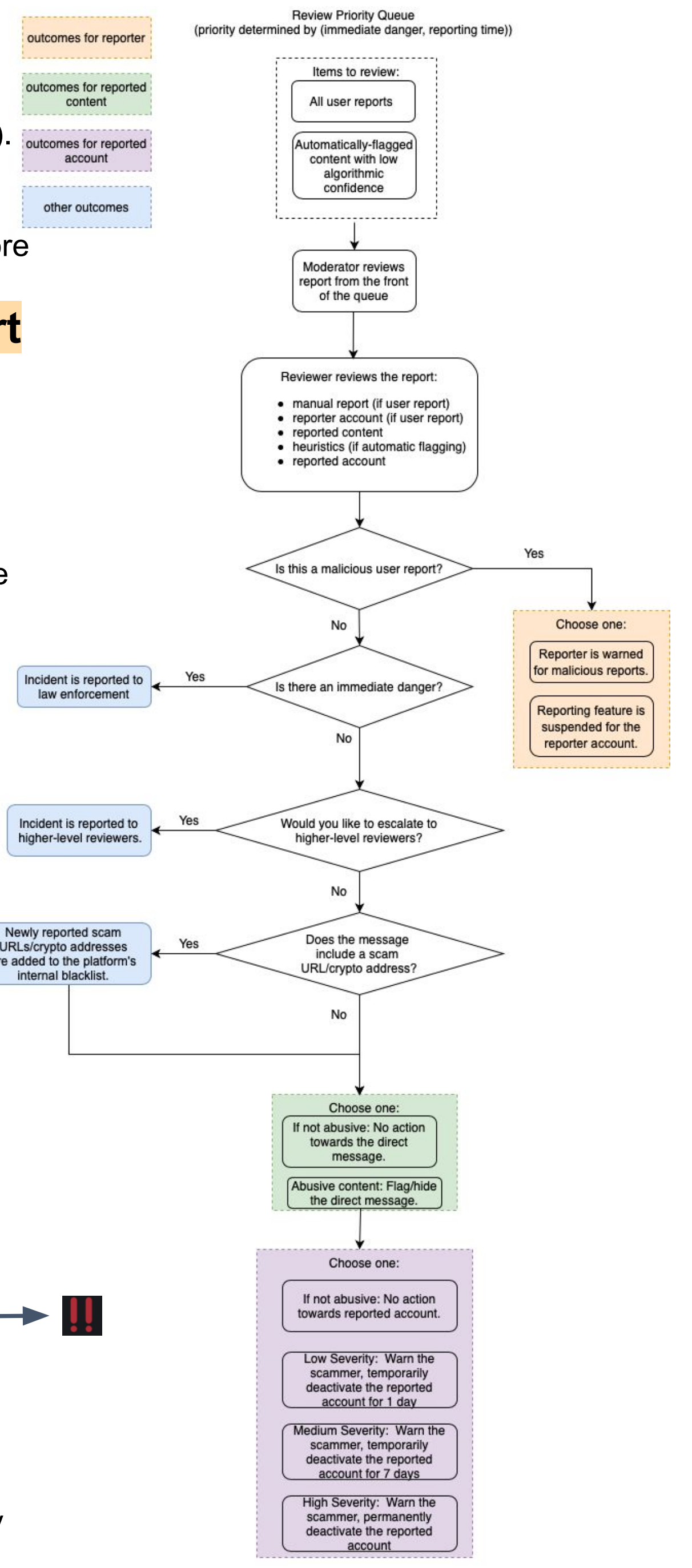
Q / ? → 🧑‍💼

Reported Content:

- Add new scam URL/crypto address in the reported message to internal blacklist (improves auto-flagging)
- **Outcomes**
  1) No action: remove Q / ?
  2) Flag/hide reported message: Q / ? → ‼

Reported Account Outcomes:

1) No action
2) Deactivate based on severity (three-strike policy): temporarily or permanently deactivate scammer account (simulated by automatically DMing the abuser)


outcomes for reporter
outcomes for reported content
outcomes for reported account
other outcomes
Review Priority Queue
(priority determined by (immediate danger, reporting time))
Items to review:
All user reports
Automatically-flagged content with low algorithmic confidence
Moderator reviews report from the front of the queue
Reviewer reviews the report:
manual report (if user report)
reporter account (if user report)
reported content
heuristics (if automatic flagging)
reported account
Is this a malicious user report?
Yes
No
Choose one:
Reporter is warned for malicious reports.
Reporting feature is suspended for the reporter account.
Is there an immediate danger?
Yes
Incident is reported to law enforcement
No
Would you like to escalate to higher-level reviewers?
Yes
Incident is reported to higher-level reviewers.
No
Does the message include a scam URL/crypto address?
Yes
Newly reported scam URLs/crypto addresses are added to the platform's internal blacklist.
No
Choose one:
If not abusive: No action towards the direct message.
Abusive content: Flag/hide the direct message.
Choose one:
If not abusive: No action towards reported account.
Low Severity: Warn the scammer, temporarily deactivate the reported account for 1 day
Medium Severity: Warn the scammer, temporarily deactivate the reported account for 7 days
High Severity: Warn the scammer, permanently deactivate the reported account

Evaluation

Classifier: Performance Metrics

This confusion matrix displays the performance of our classifier on our custom dataset (described above). We use a 65%/35% training/test split.

		Predicted Label	
		Non-scam	Scam
True Label	Non-scam	0.929	0.071
	Scam	0.0	1.0

(normalized by row)

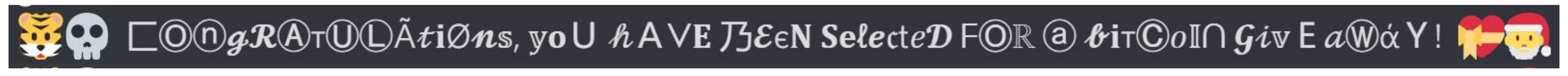
Performance on the test set is quite good, with few false positives and no false negatives; however, due to the small dataset, the classifier does not generalize well outside of this data distribution, as we see in the qualitative analysis (right). However, our training pipeline should scale well to the much larger datasets that we would have in production. Given more time and resources, data collection would be our next step.

User Report/Manual Review Flows: Qualitative Analysis

- + The user reporting flow allows users to add a justification for their report, allowing for nuanced discussion
- + The priority queue system allows moderators to check out one report at a time, keeping the channel clean
- + The manual review flow allows moderators to contact law enforcement or escalate to higher-level reviewers
- - Our classifiers can be overly sensitive (see below), putting a larger burden on the human moderators

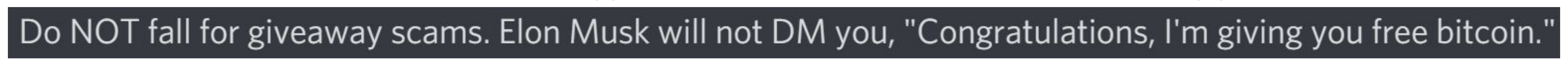
Auto-flagging: Qualitative Analysis

- + The classifier is not tricked by adversarial text with unicode characters and strange casing:



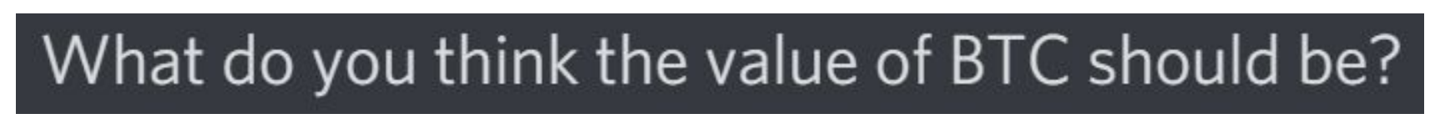
🐯💀 ⊏Ⓞⓝ𝙜𝓡ⒶᴛⓊⓁÃ𝓽𝓲Ø𝓷𝓼, 𝘆𝗼U 𝓱AVE 𝕭𝕰€N 𝐒𝐞𝐥𝐞𝐜𝐭𝐞𝐃 FⓄR ⓐ 𝓫ɪᴛⒸ𝑜𝕀∩ 𝙂𝙞𝙫 E 𝙖Ⓦά Y ! 💝🎅

- + Real-world scam messages scraped from r/CryptoCurrency's Discord server are detected very accurately
- + Blacklist system effective at notifying users of fraudulent BTC addresses and URLs in various formats 🕸
- - Counterspeech is often incorrectly flagged as abusive, disproportionately affecting good samaritans:



Do NOT fall for giveaway scams. Elon Musk will not DM you, "Congratulations, I'm giving you free bitcoin."

- - Can be overly sensitive to cryptocurrency jargon:



What do you think the value of BTC should be?

Looking Forward

Currently, we believe our system would have a positive impact on the safety of our community. It should dramatically reduce the number of cryptocurrency-related scam messages/accounts, through both automatic (classifier + blacklist) and manual (user reports + human moderation) methods.

Given more time and resources, we would consider improving our system in the following ways:

- Collect a larger dataset to make our classifier more robust and accurate.
- Automatically follow URLs and analyze target webpage content to detect scam websites. Currently, we use only our internal blacklist to determine whether a URL is a scam.
- Consider making use of external blacklists of scam URLs/crypto addresses (such as CryptoScamDB, which collects crypto scam URLs) in addition to our internal blacklist—although we would need to vet these blacklists thoroughly before deploying them.
- Make our URL/Bitcoin address detections more robust to adversarial formatting, e.g. breaking up URLs with spaces.
- Allow users to report multiple messages at once to give reviewers better context.
- Conduct large-scale user studies to assess the ease-of-use of our reporting and review flows.