Links

Git Repo: https://github.com/jackle3/cs152

Demo Video (6 minutes): 152_milestone2_6_minutes.mov

Longer Demo Video (explaining user/moderator flows): cs152_milestone2_final.mov

Writeup

Our group chose to focus on fraud, which we defined as any deceptive practice that results in financial or personal gain, as our harm type. Given this broad definition, we categorized types of fraud into four subtypes: 1) phishing: attempts to steal a user's information; 2) investment scams: attempts to coax users into providing immediate payment; 3) e-commerce scams: fake online stores that promote paid goods or services that are never provided or messages advertising counterfeit products; and 4) account takeovers: attempts to control a user's account through unauthorized logins or messaging from that account. When a user reports a message, the bot prompts the user to pick a subtype.

For phishing reports, users select one of four common targets: identity, location, payment information, or social security number. For investment scam reports, users indicate whether the message promotes fake investments into cryptocurrencies, whether the message is part of a romance scheme, or something else. For e-commerce scam reports, users indicate whether they have encountered a fake online store or the advertisement of a counterfeit item. For account takeover reports, users indicate whether they have dealt with an unauthorized login into their account or an unauthorized message sent from their account.

After a user selects the fraud type and subtype, the user is asked whether they would like to provide any additional information in a free-text space. Our group reasoned that users who wished to provide more details would appreciate a more interactive experience, and the moderation team may benefit from additional details for cases that may not neatly fall into one category.

Finally, users submit their report, which produces a confirmation message for the user and some information about the next steps. Our message states that the user will receive an update when the report has been reviewed internally and actioned appropriately. Further, it provides immediate steps the user can take to secure their accounts or prevent further unwanted interactions with specific users.

On the moderator side, once a report has been received, a moderator reviews the report details and decides whether to take action or dismiss the report based on whether the moderator believes the fraud occurred. Because of this binary choice, we did not include an escalation

pathway for further review. Should the moderator choose to dismiss the report, a dismissal summary displays to the moderator, and a message is sent to the reporting user, letting them know the moderator has finished their review. Should the moderator action the report, the moderator must decide whether to remove the message or to keep the message. Then, the moderator decides which of four actions to take against the infringing user: First-time offenders receive a warning. After a second offense, users are not allowed to use the service for 24 hours. A third offense results in a kick off the server. Finally, a fourth offense results in a complete ban from the platform. Moderators are also asked to indicate how severe the violation is based on the content of the message: Low, Medium, High, Critical. For high and critical severity violations, the system automatically notifies a specialized investigations individual to follow up on the report. Once a moderator has reviewed and confirmed the summary of their report, a message is sent to the reported user, detailing the offense and the action taken.