Addressing the Coordinated Harassment of Journalists

Aditya Iswara, Alex Hurtado, Arvind Subramanian, Sarah Jaques

{iswara, hurtado, arvindvs, sjaques}@stanford.edu

Problem Description

We focused our implementation towards the coordinated harassment of journalists. As the world becomes increasingly dependent on online communication, we've observed a dangerous trend of increasing abuse towards journalists by a variety of groups - for example, terrorist organizations, governments, white supremacists, and even ordinary readers. Journalists can be targeted for a wide variety of reasons but are disproportionately targeted with respect to their race, religion, ethnic background, gender or sexual orientation. These intimidation tactics are extremely effective and pose a grave threat to society and freedom of expression.

Policy Language

Users are not permitted to take part in behavior targeted at harassing another person, whether done individually or in collective action. We consider harassment to be any attempt to silence another individual, prevent them from sharing their views, or belittle them. Do not post content that targets individuals based on certain characteristics, encourages coordinated harassment, or expresses a desire for harm or intimidation. Our platform is dedicated to tolerance and respect and being an open environment for users to share their views. Any attempts to harm this environment will result in the content being removed from our platform and potential consequences for the users involved. If you see or experience harassment or another abuse, please report it to our content moderators and seek additional help as necessary.

Technical Back-End

The original goals of our back-end technology were to create comprehensive user reporting, review, and automated detection systems to handle general abusive content and coordinated harassment against journalists.

User Report

Accepts an abusive message, a classification of the abuse type, and a status indicating if a user is in imminent danger. Meant to address general abusive content and provide information for the moderators to act.

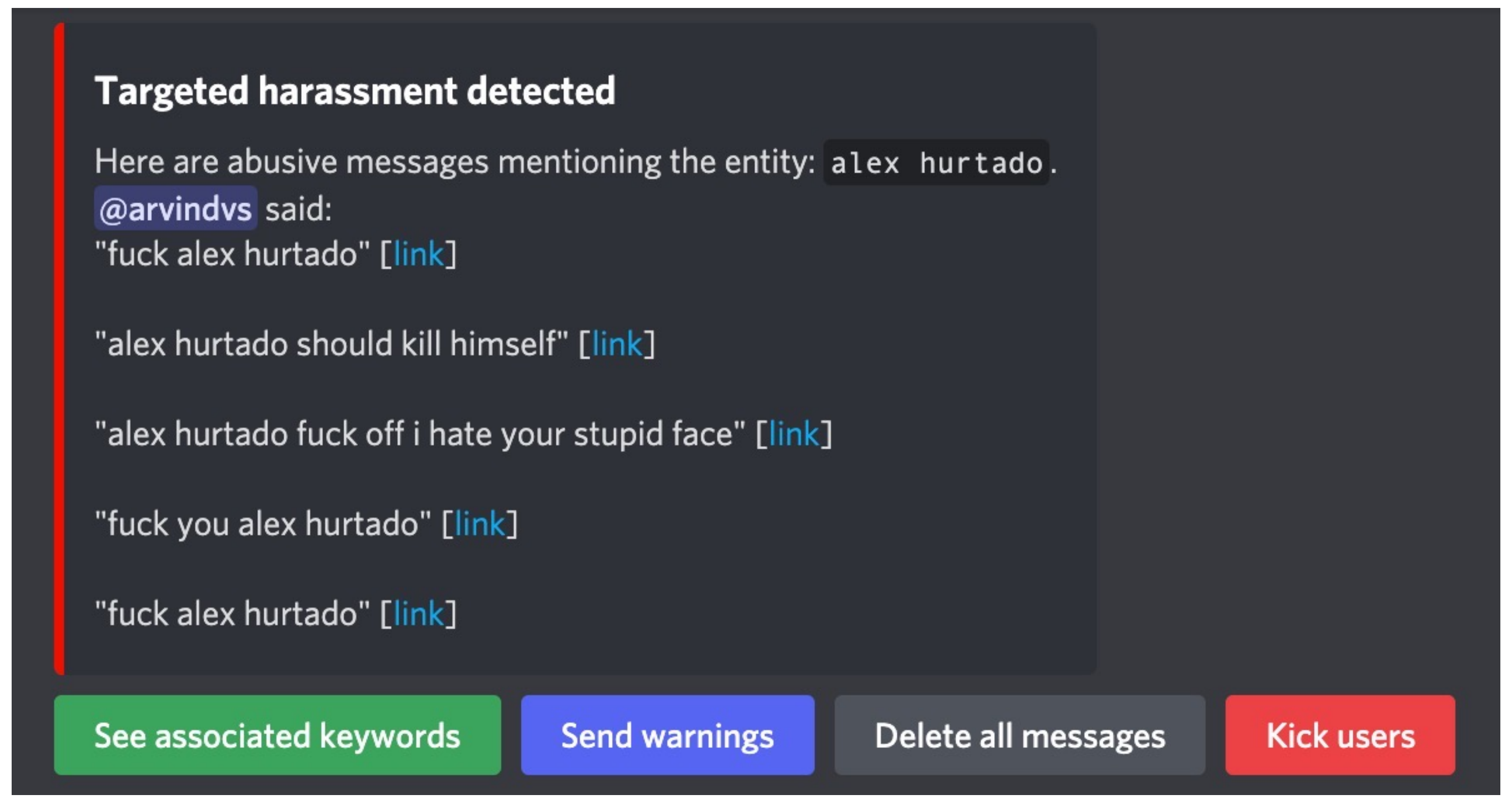
Flow extends to optionally accept information specific to targeted harassment campaigns, which is useful in addressing coordinated harassment of journalists. Collects all messages that the user believes is part of the harassment campaign. Accepts the Twitter handle of the harassment campaign victim, checking if the reported Twitter handle is valid.

Targeted Harassment Campaign Detection + Prevention

Builds on the Perspective API and techniques in natural language processing (NLP) to automate the detection of a harassment campaign and automate the flagging of associated keywords.

In addition to evaluating the abuse score of each message, we leverage named entity recognition to extract all peoples mentioned in its text (e.g. Putin, Ben Shapiro). Our backend keeps a record of all abusive messages referencing a particular entity and surfaces these messages as part of a targeted harassment campaign once they exceed an abuse threshold. Moderators can then warn the users, kick the users, and delete their messages.

Since we keep a record of all abusive messages that reference a particular entity, we can identify the keywords that seem specific to a particular harassment campaign. This would be useful, for example, if a harassment campaign uses a substitution word to circumvent typical slur detection. We use a modified implementation of term frequency-inverse document frequency to automatically identify these abuse-adjacent keywords. Moderators can then flag these words in the chat as abuse, thus surfacing messages containing these words.



Targeted harassment detected

Here are abusive messages mentioning the entity: `alex hurtado`.

@arvindvs said:

"fuck alex hurtado" [link]

"alex hurtado should kill himself" [link]

"alex hurtado fuck off i hate your stupid face" [link]

"fuck you alex hurtado" [link]

"fuck alex hurtado" [link]

Abusive User Detection

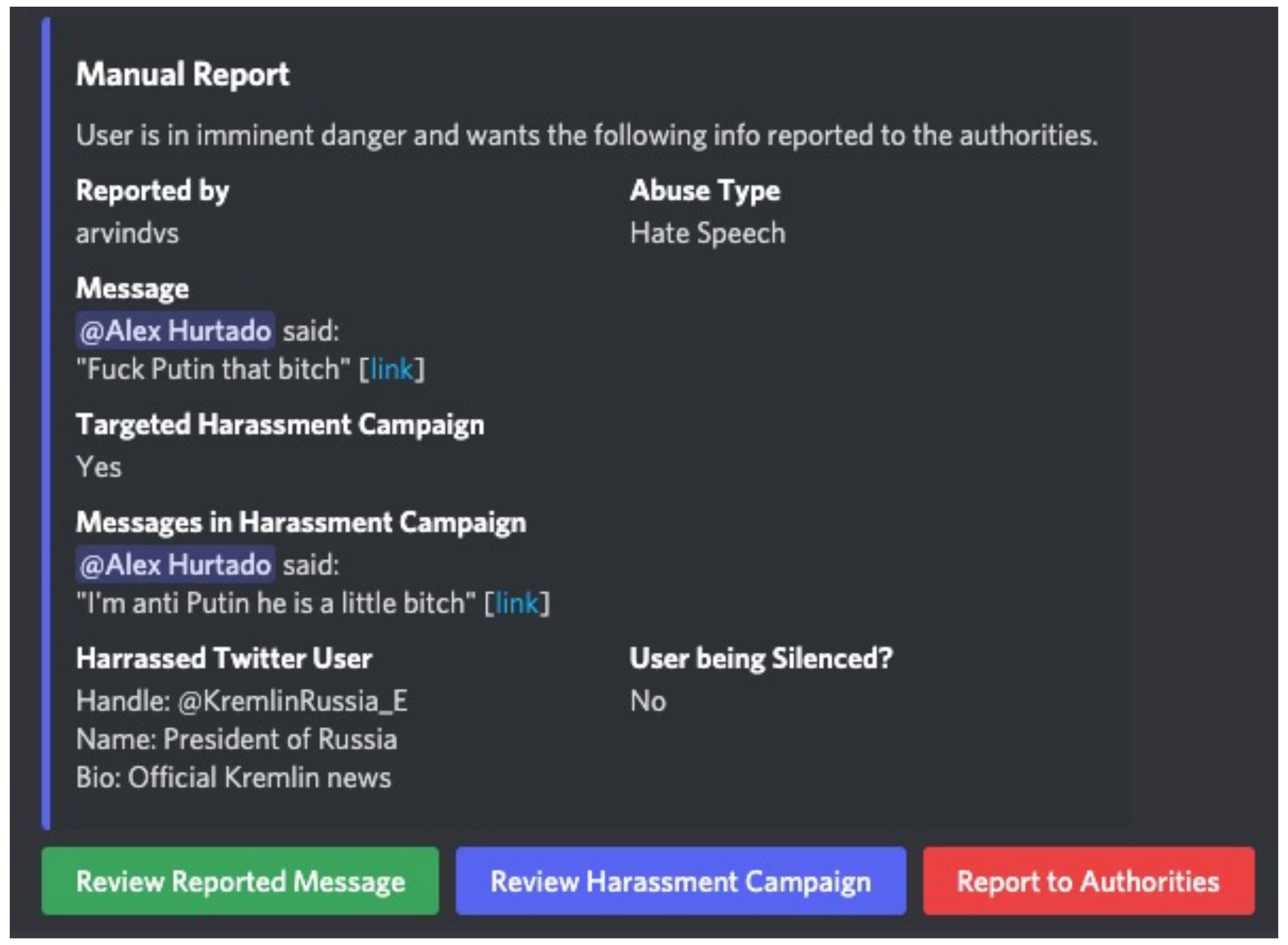
Utilizes Jigsaw's Perspective API to evaluate each message and record any message that exceeds a certain abuse score. Users that exceed a threshold of abusive messages are flagged to moderators. Moderators can then warn the user, kick the user, and delete their messages.

Manual Review

Uses a series of button interactions to moderate content. Information is split into two categories: general abuse and targeted harassment campaign. Each has a separate review flow. Reviewers can alert authorities if the user is in imminent danger.

With general abuse, the reviewer can either kick a user from a reported channel or delete the reported message. If the content isn't abusive, a reviewer can indicate to the user that the content doesn't violate guidelines.

With targeted harassment campaigns, a reviewer can act on the optional information provided by the user. If the user reports a list of messages, a reviewer can either delete all messages or kick all users who posted the message. If the user provides a targeted Twitter account, a reviewer can forward that information to Twitter's investigation team to encourage cross-platform cooperation. We use Twitter accounts as a proxy for contacting/identifying journalists since Twitter is popular in the profession.



Manual Report

User is in imminent danger and wants the following info reported to the authorities.

Reported by
arvindvs

Abuse Type
Hate Speech

Message
@Alex Hurtado said:
"Fuck Putin that bitch" [link]

Targeted Harassment Campaign
Yes

Messages in Harassment Campaign
@Alex Hurtado said:
"I'm anti Putin he is a little bitch" [link]

Harrassed Twitter User
Handle: @KremlinRussia_E
Name: President of Russia
Bio: Official Kremlin news

User being Silenced?
No

Evaluation

For this evaluation, we scraped Twitter for abusive tweets mentioning Ben Shapiro. After gathering 100 tweets, we checked to see if our model can recognize Ben Shapiro as the one being abused in each. In **71% of tweets**, "Ben Shapiro" or "@benshapiro" was correctly recognized as the target of abuse.

Overall, our bot is effective on common sentence structures. However, consider the following tweet:

- "@benshapiro, ben just wanted it to flop so r patts would do porn ;) we see you ben you dirty dawg"

Our model recognized "ben" as the entity and not "@benshapiro." While this is technically accurate, it is difficult to automatically detect that this Ben is equivalent to "Ben Shapiro" and not another Ben. This is a misclassification of the named entity.

Another example is

- "@benshapiro Bro you are a dog shit film critic, The Batman is bad? Parasite is bad? Snowpiercer bad?"

This model recognized "Bro" as the entity being attacked here. This is likely based on the overall structure of the sentence. In English, Bro is the subject of this sentence, but this is a tweet, so Bro refers to @benshapiro. The model is trained on general language and not tweet structures.

With more time and resources, we could use a more complex named entity model, such as a state-of-the-art NLP model like BERT, which could be trained specifically on abusive language and on tweets. This would resolve the issues discussed earlier as well.

Looking Forward

Without the right safety and reporting tools, any amount of abuse could potentially go undetected or unreported. The ability to alert law enforcement ensures that users who post abusive content face proper legal repercussions. Furthermore, automated detection methods that learn from recent trends in abusive language protects the user from experiencing a deluge of both blatant and coded abuse. These implementations should greatly increase platform safety and ease of moderation.

Our group has several avenues for improving our technical approach, including:

- Incorporating conference resolution to better identify indirect references of people in abusive messages
- Cross-referencing targeted entities against a database of known journalists
- Allowing targeted people to suggest words to add to the chat's blacklist
- Identifying clusters of user behavior to allow targeted people to block their harassers en masse