Mitigating Terrorist Radicalization and Recruitment Online

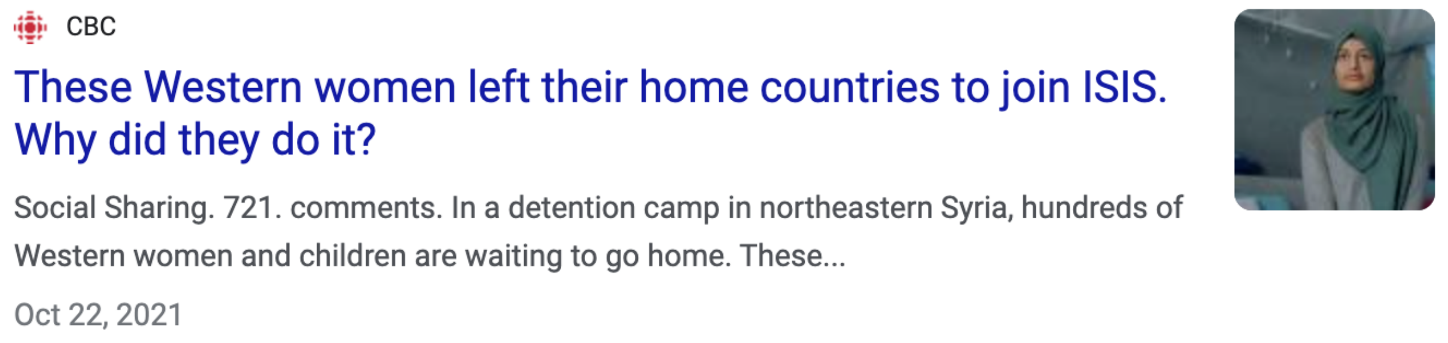
Group 10: Agnes Liang, Jennifer Lin, Rishi Sreekanth, Alex Bradfield, Nick Riedman

Problem Description

Violent extremist activity has been an unrelenting problem on social media platforms. For example, by 2017, ISIL had used social media to recruit over 30,000 foreign fighters from around 100 countries.

We identified two umbrella online abuse types relevant to terrorist activity: **threats of attacks** and **extremist content**. The latter includes content that serves as **propaganda**, **radicalization**, **recruitment**, and **documented violence**.

Through widespread calls to action, narrowcasting specific demographics, developing phony friendships to gain trust, end-to-end recruitment of lone actors, and evading capture by including links to other websites, extremists and terrorists have historically leveraged social media platforms to recruit new members.



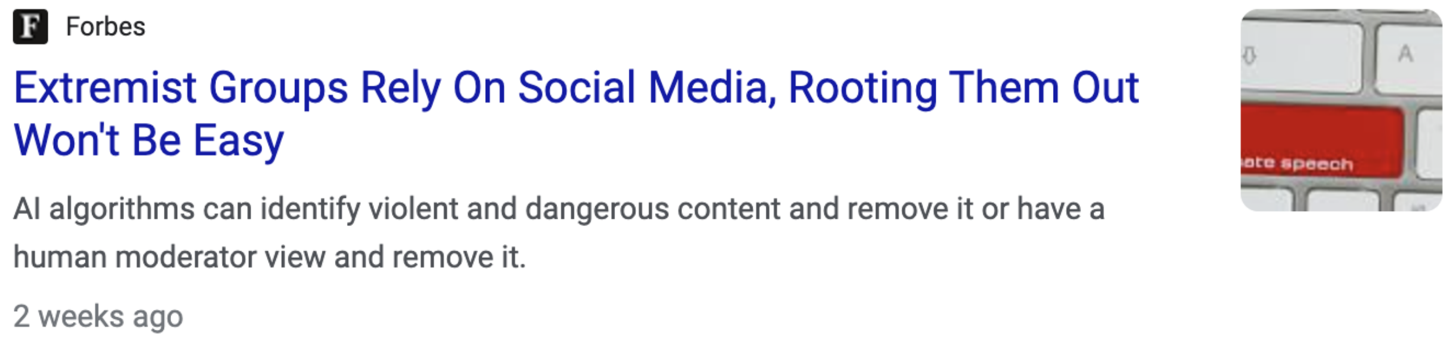
CBC

These Western women left their home countries to join ISIS. Why did they do it?

Social Sharing. 721. comments. In a detention camp in northeastern Syria, hundreds of Western women and children are waiting to go home. These...

Oct 22, 2021

Our group's goal is to identify extremist content and reduce their presence on platforms while minimizing false positives, which is presently an aspect of trust and safety that companies still struggle with.



Forbes

Extremist Groups Rely On Social Media, Rooting Them Out Won't Be Easy

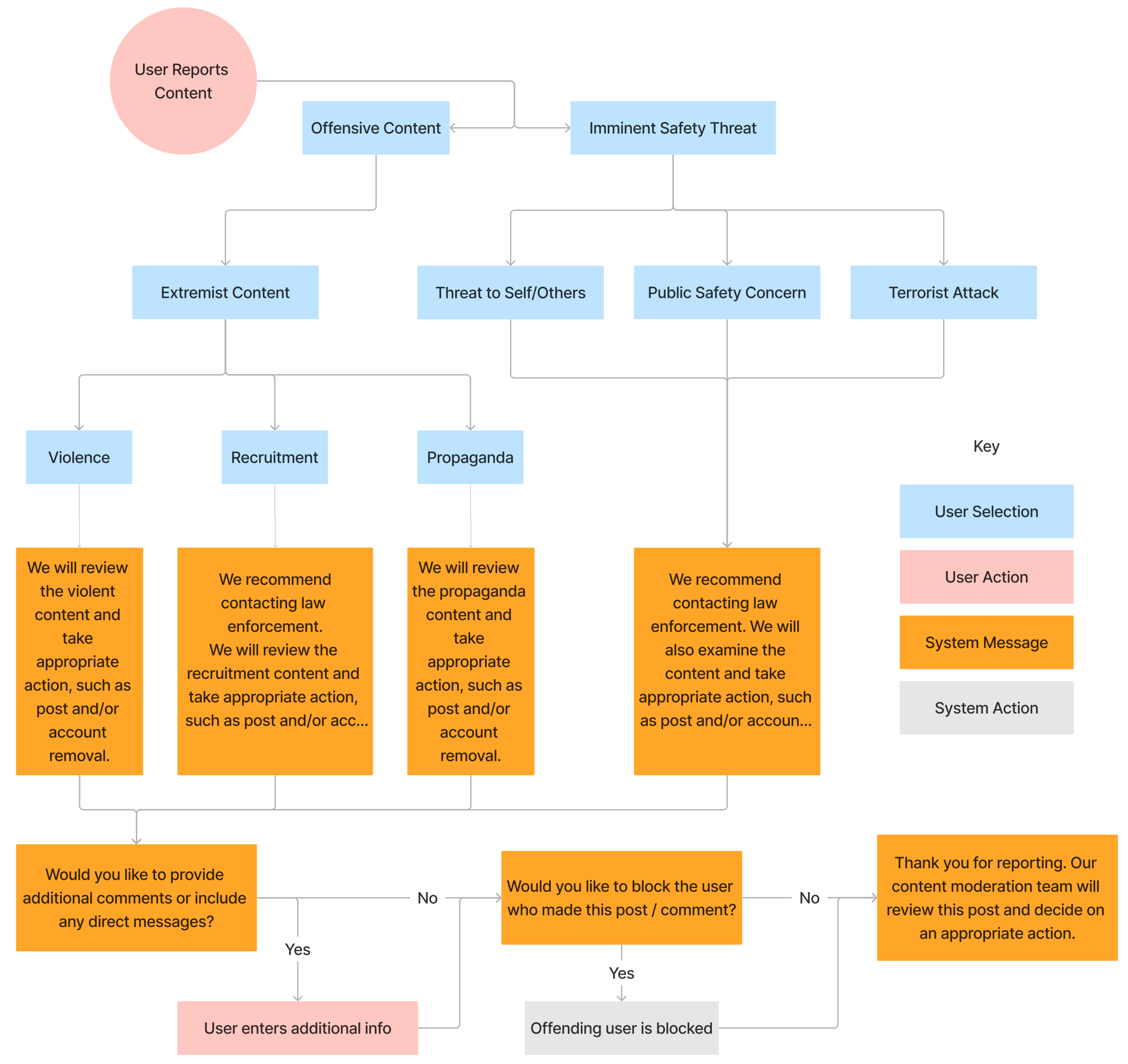
AI algorithms can identify violent and dangerous content and remove it or have a human moderator view and remove it.

2 weeks ago

Policy Language

- Trust and Safety Goals
  - Balance protecting the freedom of expression and the well-being of our users.
  - Femove content, users, and groups flagged as propagating violent extremism and violence.
- Prohibited Content
  - Promotion of Violence: Extremist content that incites violence, including terrorist attacks, lone-wolf violent acts, and violent extremist ideologies.
  - Violent Propaganda: Dissemination of graphic and non-graphic materials that glorifies violence and associated terrorist groups.
  - Recruitment for Violent Extremist Groups.
- Enforcement
  - Content that violates our policy against promotion of any form of terrorism will be taken down.
  - Content that poses a significant threat to the safety of individuals or society will be passed to the appropriate authorities.
  - Multiple violations lead to account removal.
- Reporting and Review Process
  - User Reporting: Users can report content that they believe violates our policies on violent extremism and violence.
  - Review Process: All reports are reviewed by our expert moderation team to ensure fair and consistent enforcement of our policies.
  - False Reporting: If our system detects a pattern of reports that our moderators determine are false, we will first warn the user, then remove their account if they continue.

User Reporting Flow



User Reports Content
Offensive Content
Imminent Safety Threat
Extremist Content
Threat to Self/Others
Public Safety Concern
Terrorist Attack
Violence
Recruitment
Propaganda
We will review the violent content and take appropriate action, such as post and/or account removal.
We recommend contacting law enforcement. We will review the recruitment content and take appropriate action, such as post and/or acc...
We will review the propaganda content and take appropriate action, such as post and/or account removal.
We recommend contacting law enforcement. We will also examine the content and take appropriate action, such as post and/or accoun...
Would you like to provide additional comments or include any direct messages?
Yes
No
User enters additional info
Would you like to block the user who made this post / comment?
Yes
No
Offending user is blocked
Thank you for reporting. Our content moderation team will review this post and decide on an appropriate action.
Key
User Selection
User Action
System Message
System Action

Figure 1. User Reporting Flow

Moderator Flow

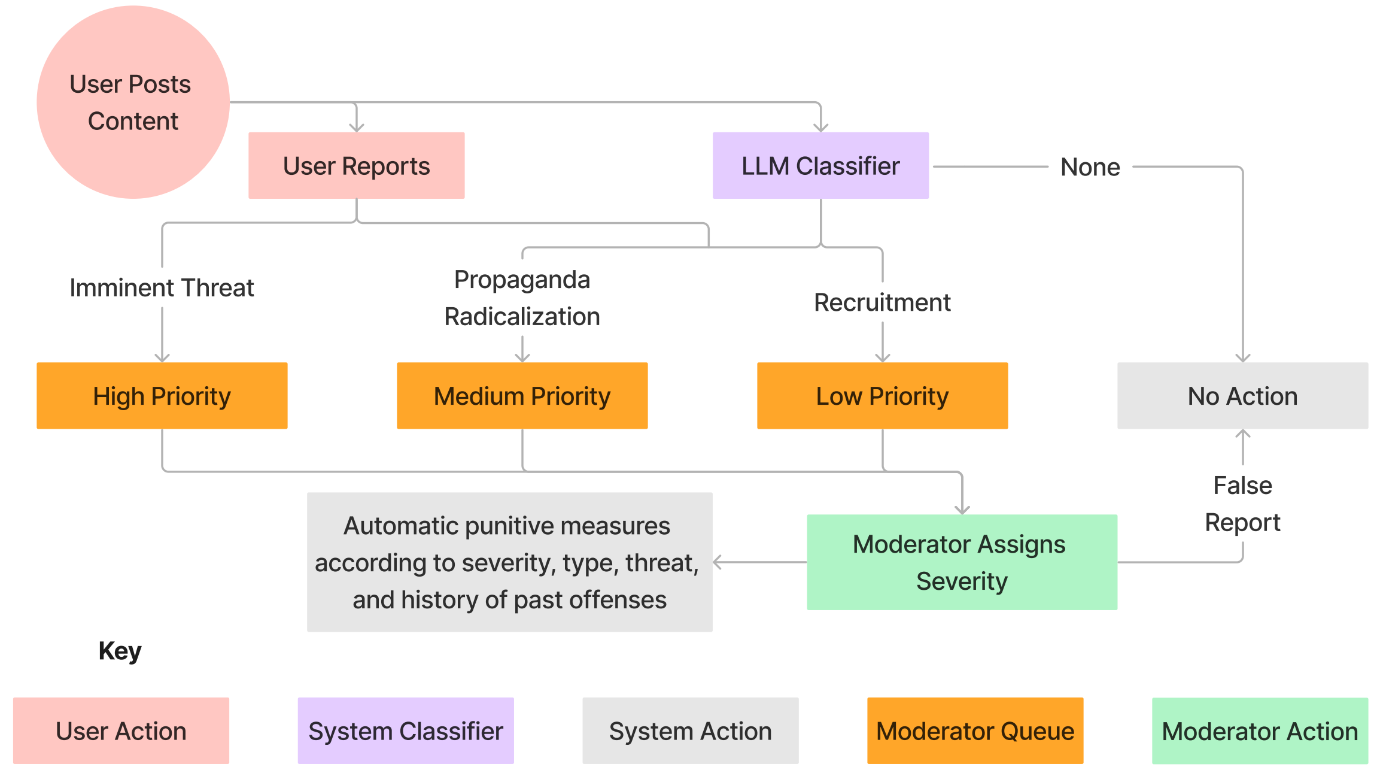


Figure 2. Moderator Flow

Technical Back-End

The back-end, moderator-facing process for automatically detecting extremist content is summarized as follows:

1. **User make a post**. An arbitrary user creates a post and shares it on the platform.
2. **Gemini classifies the text**. Given the text of a post, Gemini classifies the post as either "Radicalization," "Propaganda," "Recruitment," or "None", with the first three constituting extremist content and the latter being not extremist content.
3. **The post goes to the Moderator Flow**. If the LLM classifies the content as extremist, it is placed in the same priority queue as user-reported posts, entering the Moderator Flow to be examined manually by a human moderator.

Evaluation Dataset

We evaluated our model over 1,397 tweets. 398 of the posts are from the **Merged ISIS-White Supremacist dataset (MIWS)**, which is a multi-ideology online extremist dataset composed of tweets that fall under the categories Propaganda, Radicalization, and Recruitment [2].

1. **Propaganda:**
   - White Supremacist: "America is dead! long live America!"
   - Jihadist - ISIS: "Coalition planes massacred these children in airstrikes on Hit, Anbar"
2. **Radicalization:**
   - White Supremacist: "Only right is right of lead"
   - Jihadist - ISIS: "Himalaya888 you do realize IS wants to destroy every single nation-state, Arab or Kurd or communist does not matter, that they come across?"
3. **Recruitment:**
   - White Supremacist: "Join or die with rest"
   - Jihadist - ISIS: "This is the time for muslim kurds in Turkey to show whether they can ever counter PKK."

The remaining 999 tweets in the evaluation set are randomly sampled from **sentiment140**, a sentiment analysis dataset composed of over 1 million twitter posts labeled according to their positive/negative sentiment [3]. Tweets from this dataset are labeled "None" in our system. For example:

1. @Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
2. @octolinz16 It it counts, idk why I did either. you never talk to me anymore

Qualitative LLM Evaluation

- Both LLMs had high accuracy and F1 scores (see top right) for classifying extremist vs. non-extremist content.
- Gemini vastly outperformed GPT in recall, which was largely due to GPT's reluctance to label borderline content as extremist.
- Both LLMs struggled with correctly sub-categorizing extremist content.
- Fine-tuning our prompting by providing the LLM more detailed definitions of the categories along with examples of each might improve sub-categorization and reduce false negatives.
- We don't want to over-restrict the LLM in our prompt engineering, so considerate and diverse curation of examples is a must.

Looking Forward

Implementing user and automated reporting for terrorism-related abuse is vital for making social media platforms safer. Eliminating certain posts and messages can prevent attacks both in the US and abroad. As an FBI executive assistant director stated before the Senate Committee on Homeland Security and Governmental Affairs, "the foreign terrorist now has direct access into the United States like never before."

While our model for classifying terrorism-related content is highly accurate, we can strengthen our defenses by employing additional strategies. These include automated content image and video checks, analyzing user tendencies, and educating users on spotting extremist content.

To improve our model's quality, we should run it on a more diverse dataset that includes phrases about ethnic conflict, interstate wars, intrastate wars, and religious groups. This comprehensive approach will enable platforms like Twitter to better protect users and combat terrorism effectively.

Quantitative LLM Evaluation

We focused on evaluating the performance of OpenAI's GPT (3.5) and Google's Gemini (1.5 Flash) large language models (LLMs) classifying benign and extremist tweets (see Evaluation Dataset). We prompted both GPT and Gemini with instructions to behave like it was on a social media content moderation team, and would be tasked with reviewing the following content and categorizing it as one of the aforementioned 4 categories, and providing a reason for its categorization.

Categories	Gemini	GPT
Accuracy	96.5%	91.1%
Precision	99.7%	100%
Recall	87.9%	68.6%
F1 Score	93.4%	81.4%

Table 1. Gemini, GPT Performance Metrics for Extremist/Non-Extremist Classification

Gemini	Predicted Propaganda	Predicted Radicalization	Predicted Recruitment	Predicted None
Actual Propaganda	**154**	53	0	20
Actual Radicalization	37	**23**	2	8
Actual Recruitment	22	27	**31**	20
Actual None	1	0	0	**998**

Table 2. Gemini Sub-Confusion Matrix (n = 1396, one blocked by safety filters)

GPT	Predicted Propaganda	Predicted Radicalization	Predicted Recruitment	Predicted None
Actual Propaganda	**66**	82	12	66
Actual Radicalization	12	**34**	4	21
Actual Recruitment	0	28	**34**	38
Actual None	0	0	0	**999**

Table 3. GPT Sub-Confusion Matrix (n = 1397)

References

[1] Federal Bureau of Investigation.
Isil online: Countering terrorist radicalization and recruitment on the internet and social media.
Testimony.
Accessed April 18, 2024.

[2] Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, Ketan Kotecha, and Shalli Rani.
Multi-ideology, multiclass online extremism dataset, and its evaluation using machine learning.
Computational Intelligence and Neuroscience, 2023.

[3] Bhavik Jikadara.
Tweets dataset: Sentiment analysis using 1m+ tweets dataset.
Kaggle, 2024.