US Online Domestic Violent Extremism Monitor

Bi-Monthly Data Snapshot of Cross-Platform Social Media Activity

Executive Summary

This bi-monthly data snapshot overviews key trends in US-related violent extremist activity online, providing a contemporary picture of the evolving digital threat landscape. The report aims to provide practitioners and policymakers with actionable and up-to-date trends insights to help inform prevention efforts. Analysis is based on a dataset of over 760 US-linked accounts and channels across a broad range of platforms and violent extremist ideologies, manually vetted by experts as engaging in clear violent extremist behavior. Data was collected through platforms' public Application Programming Interfaces (APIs), with analysis incorporating the use of Large Language Models (LLMs) to identify key trends in violent extremist discourse, including prominent narratives, the targets of violent extremist activity, and the nature and extent of targeted hate against minority communities. All findings are anonymized and presented in aggregate, with personally identifiable information removed at the point of data collection. A full methodology is included at the end of this report.

Key Findings

During December 2024 and January 2025, US violent extremist actors produced over 1.28 million posts, with a further 987,000 posts recorded within US violent extremism-linked online spaces. A summary of key findings is provided below:

Platform Activity:

• Telegram remained the leading platform for violent extremist activity, despite a 22% drop in posts since October-November. Meanwhile, language indicative of extremism or targeted hate from US users on 4chan's /pol/ board increased by 45%. Violent extremists on Bluesky, a new entry, posted over 35,000 times, demonstrating the importance for practitioners in staying across trends on emerging platforms.
• Violent extremist activity increased by 42% between December and January, peaking in the days following Inauguration Day. Practitioners should note this close relationship between offline political disturbance and online violent extremist activity.

Threat Categories:

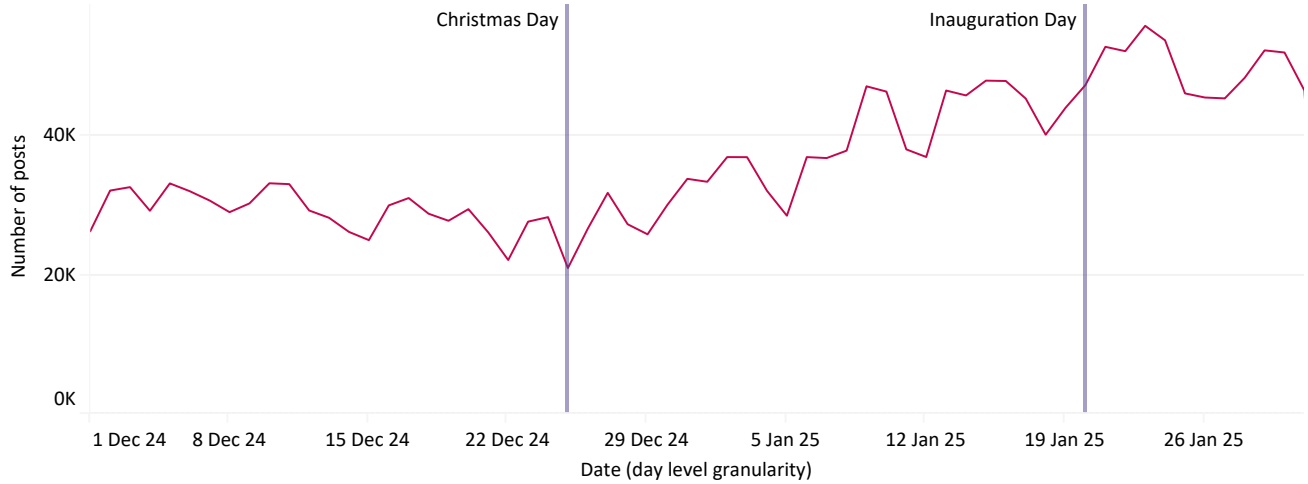
• While violent extremist actors without a clearly defined ideology garnered the most engagement, users classified under violent extremism targeting political opponents generated the most posts, representing a 28% increase since last monitoring period.
• Pro-foreign terrorist organization (FTO) accounts were the threat actor group with the second-most message engagement. Recent ISD analysis found that pro-Islamic State (IS) Facebook pages and TikTok videos garnered tens of thousands of followers and views, respectively, while Similarweb analysis ranked US-based users second among pro-IS site visitors (17% of traffic). This demonstrates the continued need for practitioners to understand the pro-FTO and domestic violent extremism nexus online.

Mobilizing Narratives:

• Bespoke LLMs showed that while only 0.8% of messages from violent extremist actors online were overtly violent in nature, broader conversation topics included Technology and Internet Culture (15% of thematic discussion) and Culture and Society (14%).
• Classifiers detected targeted hate within 4.72% of messages, with 3% of messages antisemitic, 1.65% anti-LGBTQ+, and 0.5% anti-Muslim.
• Violent extremists were especially galvanized by violent incidents in New Orleans and Las Vegas on New Year's Day, the January 6 anniversary, and the change in presidential administration on January 20. Analysis of evolving online narratives espoused by violent extremists are key to informing targeted prevention efforts.

Domestic and International Interplay:

• Domestically, violent extremists across the ideological spectrum were animated by the assassination of the United Healthcare CEO in early December. The response shows potential for such attacks to inspire others with personalized grievances and shows the importance of practitioners intervening with individuals who may look to conduct "copycat" style attacks.
• Internationally, violent extremists in the US reacted to the fall of the Assad regime in Syria with antisemitic rhetoric and speculation about the alleged involvement of the "deep state." The spike in violent extremist interest in Syria shows the importance of practitioners understanding the dynamic relationship between global developments and online violent extremist activity, key to upstream prevention.



Volume-over-time graph showing the total number of posts per platform

This research was supported by the U.S. Department of Homeland Security, Science and Technology Directorate, under Grant Award Number 23STFRG00021. Any opinions or conclusions contained herein are those of the authors and do not necessarily reflect those of the Department of Homeland Security, Science and Technology Directorate.

Platforms

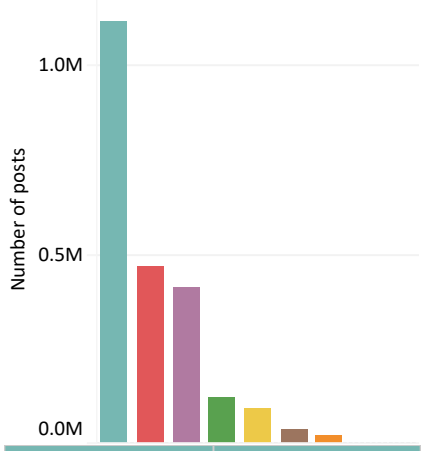
Overview

Telegram continued to be the most active platform among violent extremists, with over 1.1 million posts collected during this monitoring period. Telegram comprised almost half of all messages in the violent extremism dataset, however the overall volume of posts by violent extremists decreased by over 22% compared to the last report.

Notably, violent extremism-relevant activity from US users on 4chan /pol/ increased by 45% compared to the last report, showing the imageboard remains a prominent platform for anonymous discussion among violent extremists. This report also includes nearly 100,000 posts from four forums relevant to US violent extremist mobilization (which we have opted not to name to avoid undue amplification). With all these fora, we collected only posts which mentioned keywords related to hate and extremism or associated with groups that are regularly targeted with violent extremist harassment.

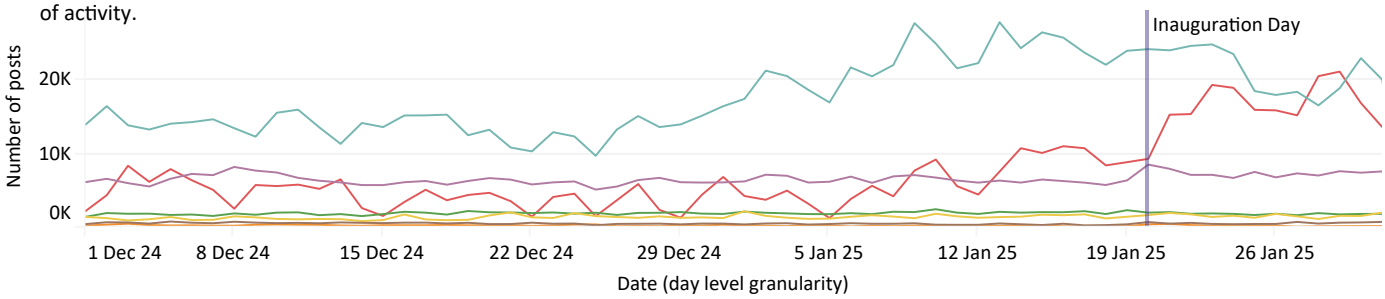
Newly introduced to this monitor is content from Bluesky, which hosted over 35,000 posts from violent extremist accounts, constituting 2% of total activity across all platforms. Notably, compared to the last report there was a 22% increase in comments on YouTube videos produced by violent extremist actors, with content from violent extremist actors on X and within violent extremist sub-Reddits also increasing by 4% and 9% respectively. Facebook activity remained relatively stable compared to the previous monitoring period, while Instagram featured a negligible amount of content during this period due to accounts becoming inactive or abandoning their focus on violent extremism.

In aggregate, across all platforms posts from online violent extremists steadily increased over the monitoring period and peaked in the four days following President Donald Trump's inauguration. Also of note was that there was 55% more violent extremist-related activity in January than December. This suggests that practitioners must prepare for potential upticks in extremist activity around major transfers of political power in the US, and that December may be a relatively quiet period for online extremists, with the holiday season marked by low levels of activity.



Telegram	1,117,431
YouTube	464,936
4Chan	412,707
X (Twitter)	121,791
Forums	91,035
BlueSky	35,758
Reddit	18,777
Facebook	565
Instagram	14

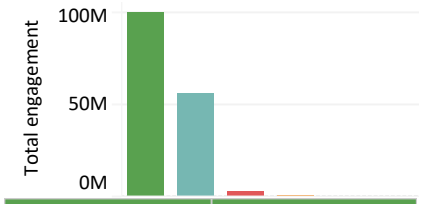
Bar graph & corresponding tabular view showing total posts per platform



Platforms

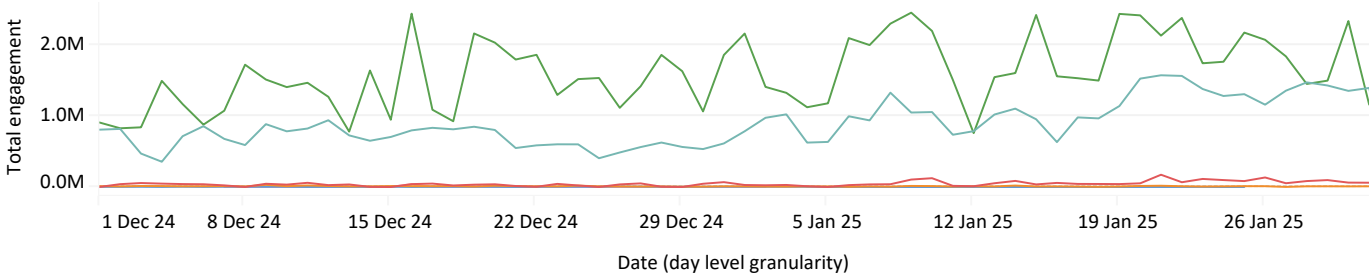
Engagement

The types and availability of engagement data (e.g., comments, likes, shares) vary dramatically by platform, but still provide valuable insight into narratives gaining traction among violent extremists online. Continuing previous trends, X and Telegram saw considerable engagement during this review period. Violent extremists on X totaled more than 18.3 million followers (an increase of over 24% from the last report), and their posts received nearly 74.3 million likes (up by over 13%). Meanwhile, on Telegram, violent extremist content received over 9.7 million likes and 2.7 billion views, a drop of nearly 61% and 70% respectively. YouTube generated substantial engagement in the form of comments (nearly 91,000), followed by Reddit (over 17,300) and Facebook (over 11,500). The aforementioned forums and Bluesky did not provide engagement metrics.



X (Twitter)	100,349,211
Telegram	55,968,766
YouTube	2,844,896
Reddit	672,082
Facebook	95,449
Instagram	52,362

Bar graph & corresponding tabular view showing total engagement per platform

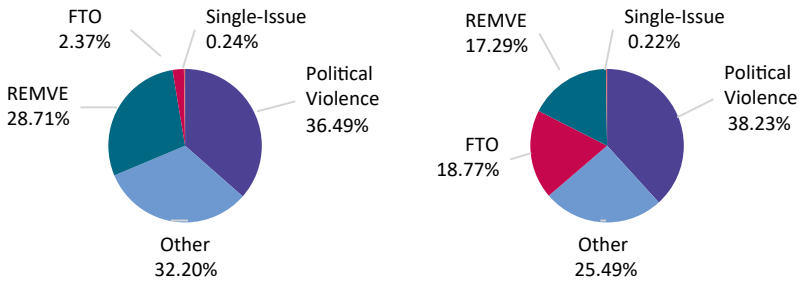


Volume-over-time graph showing total engagement per platform

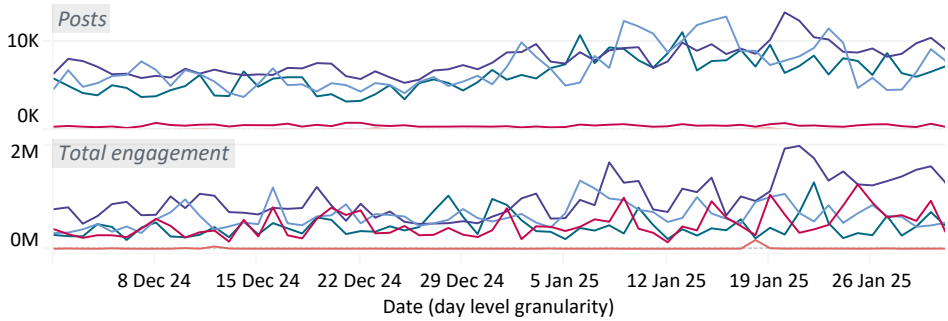
Threat categories

Overview

Drawing on terminology from DHS and other US government agencies as well as ISD's own categorizations, ISD manually classified violent extremist accounts and channels into five threat categories: (1) Foreign Terrorist Organization Support; (2) Racially or Ethnically Motivated Violent Extremism; (3) Violent Extremism Targeting Political Opponents; (4) Single-Issue Violent Extremism; and (5) Other Violent Extremism.



Circle charts showing the % of posts (left), and total engagement (right) by threat category



Volume-over-time graphs showing the number of posts (top) and total engagement (bottom), by threat category

The snapshots in this section exclude analysis from violent extremism-linked forums, where anonymous posting and more ephemeral account identity make actor-based analysis more challenging. It also excludes YouTube and Reddit comment data, as accounts were categorized on a channel and subreddit-level respectively, rather than by individual commenters.

32,373 *posts*
20 *active accounts*
29.36M *total engagement*

Foreign Terrorist Organization (FTO) support

This category encompasses US-based and US-focused accounts that express support for FTOs and endorse their violent tactics. In December and January, accounts within this category represented only 3% of total posting activity and ranked fourth for number of active accounts yet received the second highest amount of engagement (after 'Other' accounts). Within this category, 89% of activity was associated with Iran's 'Axis of Resistance', while 27% of accounts were supportive of Hamas specifically.

440,059 *posts*
36 *active accounts*
39.89M *total engagement*

Other Domestic Violent Extremism (Other)

This category encompasses threats involving the potentially unlawful use or threat of force or violence in furtherance of violent extremist ideological agendas which are not primarily motivated by one of the other domestic threat categories. This was the most active threat category during the period, with activity increasing by over 33%, despite having a relatively low number of active accounts. A new trend within this category saw 13% of activity come from accounts affiliated with the 'Com' network, a community of violent extremists motivated by a blend of nihilism, misanthropy, and an obsession with violence.

498,720 *posts*
140 *active accounts*
59.82M *total engagement*

Violent Extremism Targeting Political Opponents (Political Violence)

This category includes the potentially unlawful threat of violence against political opponents, as well as in furtherance of violent ideologies derived from anti-government or anti-authority sentiment. During this period, this was the second-most active threat categories based on messages produced (down by 5% from the last report) and number of active accounts. Despite this, content produced by these accounts comprised over 86% of the views across all threat categories, showing considerable engagement.

392,392 *posts*
252 *active accounts*
27.05M *total engagement*

Racially or Ethnically Motivated Violent Extremism (REMVE)

REMVE encompasses the potentially unlawful use or threat of force or violence in furtherance of supremacist ideological agendas based on race or ethnicity. During the monitoring period, REMVE was the third-most active threat category in terms of messages produced but had the highest number of active accounts. Compared to the last report, posts by REMVE actors increased by over 15%. Of note, almost all REMVE accounts active in this period espoused white supremacist or neo-Nazi views.

3,286 *posts*
26 *active accounts*
0.35M *total engagement*

Single-Issue Violent Extremism (Single-Issue)

Single-Issue Violent Extremism, which produced the fewest messages, is divided into the following sub-categories: Animal Rights-Related Violent Extremism; Environment-Related Violent Extremism; Abortion-Related Violent Extremism; and Israel/Palestine-Related Violent Extremism, which ISD defines as violent extremists who are singularly motivated by the ongoing Israel-Hamas conflict. While this latter group were the most active - followed by those motivated by abortion and the environment – Israel/Palestine-related violent extremist activity notably declined by over 38% compared to the previous period.

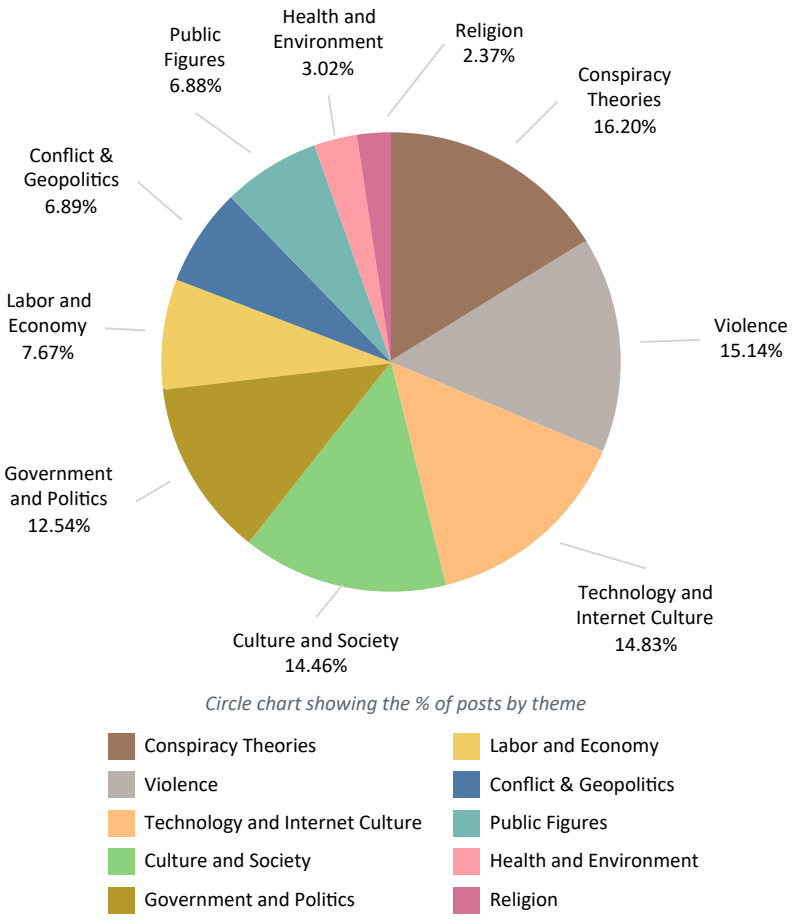
Semantic mapping

Summary

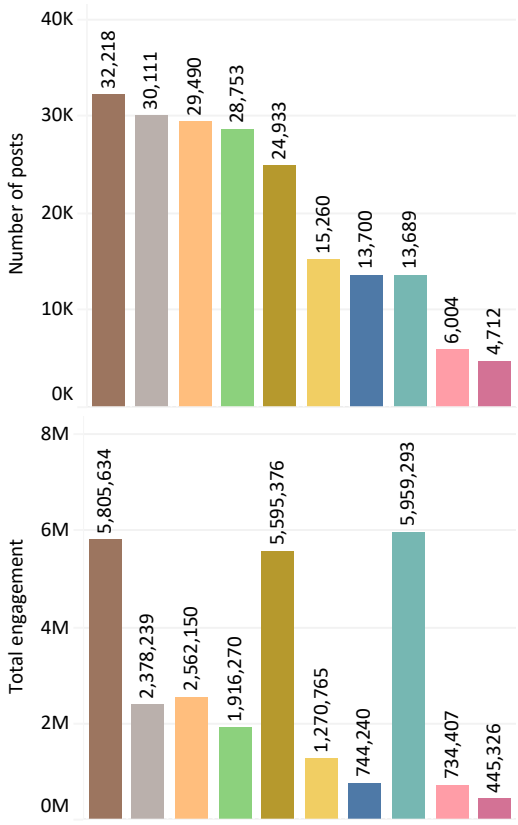
Analysis of the evolving online narratives espoused by violent extremists is crucial for practitioners to understand potential entry points for helping individuals disengage from violence. In this report bespoke Large Language Models (LLMs) were used to group messages into semantically distinct clusters to aid the analysis of key narratives at scale. LLMs were also employed to help analysts characterize clusters of messages into cohesive themes. This process identified 70 sub-themes grouped into 10 overarching themes. For example, the Government & Politics theme was comprised of eleven sub-themes, which included discussions around political figures, parties and ideologies.

Many messages posted by violent extremists were grouped into themes not directly related to hate, extremism, or violence, but were still deemed relevant for understanding the key narratives driving DVE activity. Furthermore, semantic clusters were not treated as discrete categories, with posts often encompassing multiple narratives.

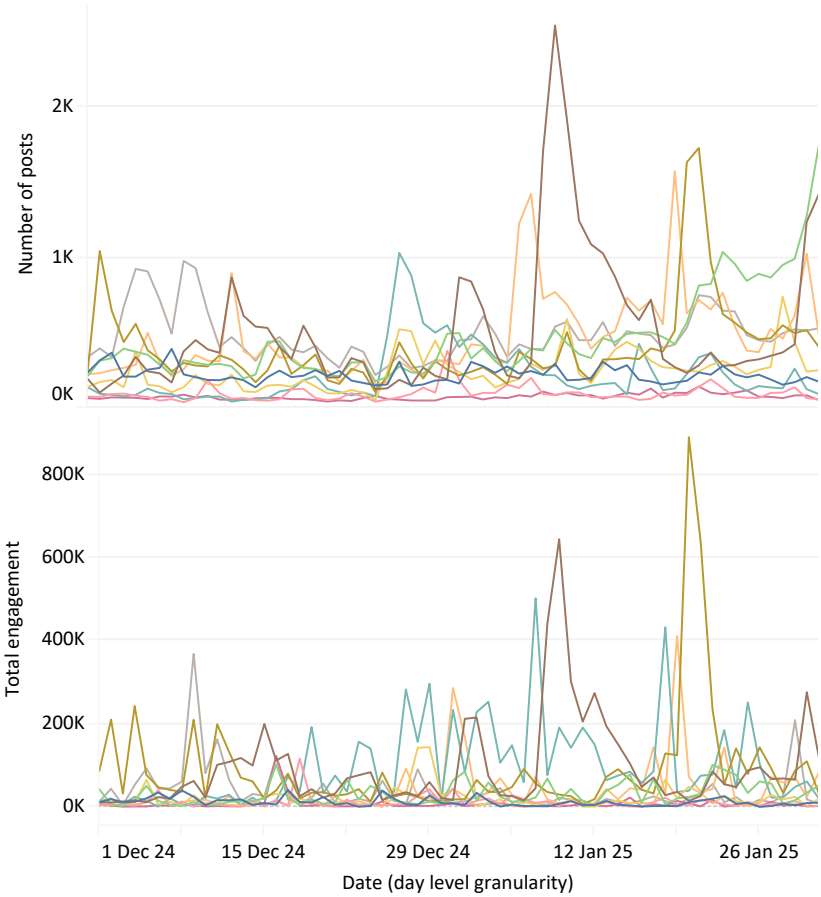
On the next page, we outline the new and notable themes emerging from this semantic mapping exercise for this period, as well as analyze the use of overtly violent language among violent extremist actors. On the following page we drill down on trends in Targeted Hate, utilizing specially trained classifiers to understand trends in targeted hate directed at Jewish, Muslim and LGBTQ+ communities by violent extremists.



Circle chart showing the % of posts by theme



Bar charts showing the number of posts (top) and total engagement (bottom) by theme

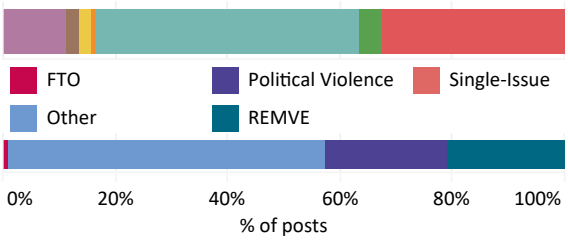


Volume-over-time graphs showing the number of posts (top) and total engagement (bottom) by theme, for the top 5 themes (by post count)

Semantic mapping - key themes

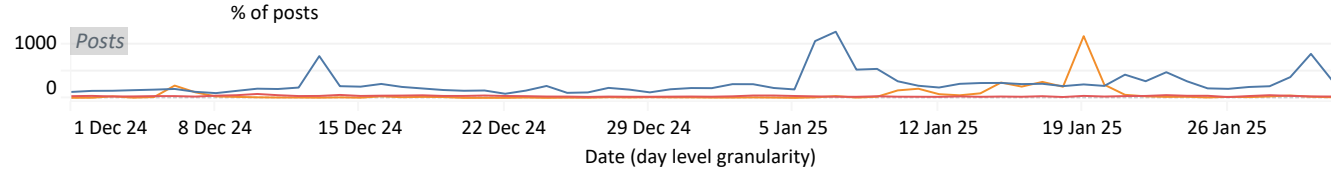
Technology and Internet Culture

Violent extremist conversation around technology and internet culture was dominated by discussions around free speech and censorship (57%), with leading narratives including complaints by violent extremists about being blocked or de-platformed, as well as distrust in social media platforms and communication apps around freedom of speech and privacy. Conversation pertaining to this theme was driven by violent extremist accounts characterized by their targeting of political opponents (33%), as well as Racially or Ethnically Motivated Violent Extremists (35%). Most violent extremist discourse relating to this theme took place on Telegram (47%), followed by YouTube (32%). Relevant messages with the greatest engagement were observed on X, which represented 85% of total engagement across this theme.



Rank	Sub-theme	Posts	Accounts
1	Free Speech And Censorship	16,694	6,258
2	TikTok Ban	4,128	2,597
3	Memes	2,256	1,869

Table showing the top 3 sub-themes by post count for this theme (above); & bar charts showing the % proportion of posts of this theme per threat category (left - bottom) and per platform (left - top)

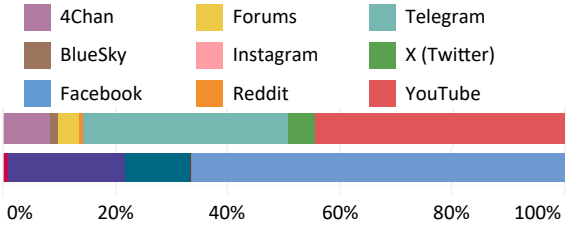


Volume-over-time graph showing the number of posts by sub-theme, for the top 3 sub-themes (by post count) for this theme

Semantic mapping - key themes

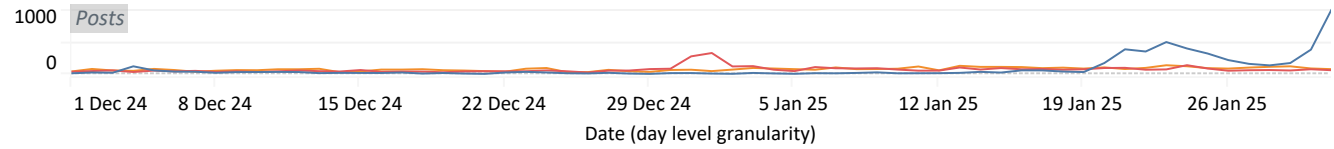
Culture and Society

Conversations among online violent extremists in this theme were dominated by discussion of transgender people (21%) and distrust in online media (19%). Top narratives about transgender people included discussions of diversity, equity, and inclusion (DEI) in the context of recent aviation accidents and declarations that only two genders are valid. Allegations around 'fake news' regarding the Israel/Hamas war and US politics dominated conversation about online media distrust. Relevant thematic conversation was driven by violent extremist accounts with no cohesive ideology (61%), followed by those calling for political violence (19%) and REMVE accounts (11%). Most violent extremist discourse within the culture and society theme took place on YouTube (44%), followed by Telegram (37%). However, the most engagement with relevant messages was observed on X (82% of total engagement).



Rank	Sub-theme	Posts	Accounts
1	Discussions Of Trans People	6,082	4,544
2	Online Distrust	5,428	1,979
3	New Era	5,085	1,410

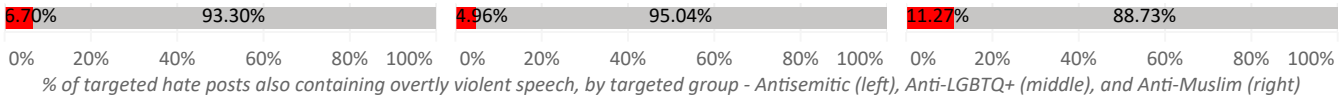
Table showing the top 3 sub-themes by post count for this theme (above); & bar charts showing the % proportion of posts of this theme per threat category (left - bottom) and per platform (left - top)



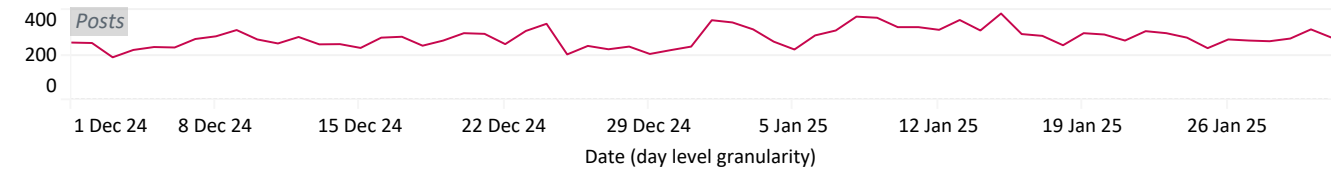
Volume-over-time graph showing the number of posts by sub-theme, for the top 3 sub-themes (by post count) for this theme

Overt Violence

Bespoke classifiers trained to detect violent speech shows that - while all accounts in this research are included based on clear violent extremist behavior - only a small proportion of their messaging (0.8%) can be classed as overtly violent, including online content that threatens, incites, or glorifies acts of violence. The most engaged-with violent content centered on the promotion of Israel's destruction and the veneration of Hamas leader Yahwa Sinwar from a single account with a high following. This reflects a broader pattern where a small number of key actors produce a disproportionately high share of violent content across our dataset. For example, on Telegram, 69% of violent posts originated from just 10 of the 165 included accounts. Antisemitism represented the most notable intersection of violence and targeted hate, accounting for 58% of total violent hateful posts. While our analysis did not reveal any significant shifts in overall violent rhetoric against specific communities across the time period, a significant surge in anti-Muslim violent rhetoric followed the New Orleans truck attack on January 1, 2025.



% of targeted hate posts also containing overtly violent speech, by targeted group - Antisemitic (left), Anti-LGBTQ+ (middle), and Anti-Muslim (right)



Volume-over-time graph showing the total number of posts inciting violence

Targeted hate

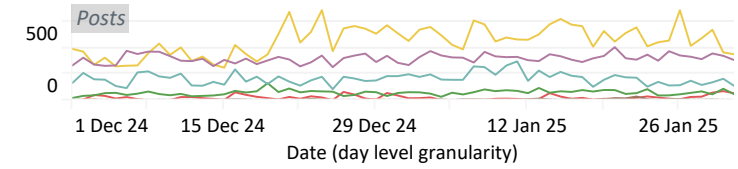
Antisemitism

Antisemitism once again represented the most prominent form of targeted hate in our dataset, comprising 3% of all messages (a 0.4% increase from the last report). The highest number of antisemitic posts came from the four violent extremist-relevant forums (44%), followed by 4chan (31%) and Telegram (16%). Violent extremist accounts without a clear ideology were responsible for more than half of antisemitic posts in the dataset (68%), followed by REMVE (25%) accounts. Prominent antisemitic narratives included conspiratorial messages about Jewish control of media, financial, and government institutions and Israeli influence over US and global politics.

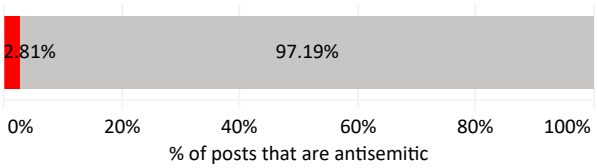
63,618
posts

19,973
active accounts

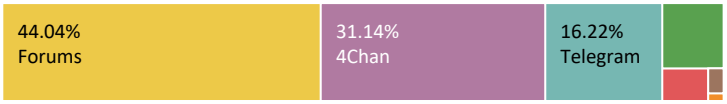
5,337,878
total engagement



Volume-over-time graph showing the number of antisemitic posts by platform

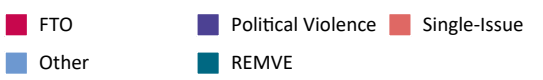


Bar chart showing the % of antisemitic (red) posts



Treemap showing number of antisemitic posts per platform

Treemap showing number of antisemitic posts per threat category



Targeted hate

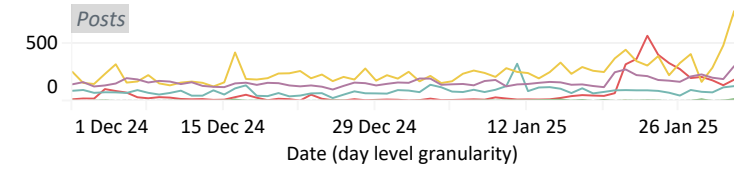
Anti-LGBTQ+ Hate

Anti-LGBTQ+ content comprised 1.65% of messages in our dataset, a 0.1% increase compared to the previous report. Hate targeting LGBTQ+ communities spiked around inauguration day and remained slightly elevated through the remainder of the month. 43% of relevant posts were made on the four violent extremist-dominated forums analyzed, 28% were posted on 4chan, and 15% on Telegram. Violent extremist accounts with no clear ideology were behind 87% of relevant messages, distantly followed by REMVE accounts (11%). Prevalent narratives included the demeaning use of hateful anti-LGBTQ+ slurs and allegations of pedophilia. A handful of high-engagement posts also called for unspecified violence against LGBTQ+ people.

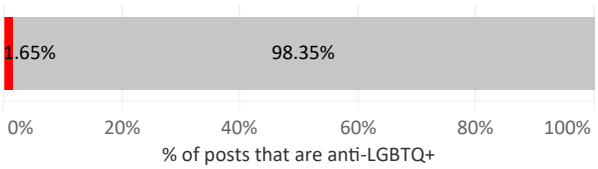
37,241
posts

15,001
active accounts

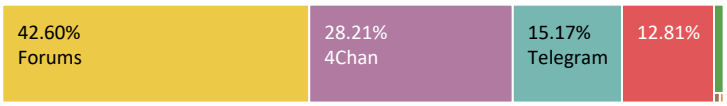
711,900
total engagement



Volume-over-time graph showing the number of anti-LGBTQ+ posts by platform



Bar chart showing the % of anti-LGBTQ+ (red) posts



Treemap showing number of anti-LGBTQ+ posts per platform

Treemap showing number of anti-LGBTQ+ posts per threat category

Targeted hate

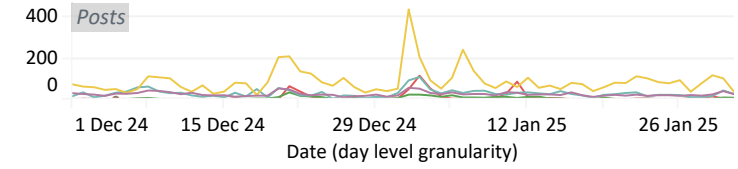
Anti-Muslim Hate

Posts containing anti-Muslim hate made up 0.5% of the messages in our data set, a 0.2% increase from the last monitoring period. 53% of anti-Muslim targeted hate was derived from violent extremist-related forums, while 19% occurred on Telegram. Violent extremist accounts without a clear ideology produced the majority of anti-Muslim content (78%). Targeted hate against Muslims was particularly salient in the wake of the IS-linked New Orleans truck ramming attack early on New Year's Day. Prominent anti-Muslim narratives during this period include calls to expel Muslims from the US, as well as comments on the alleged incompatibility of Muslims and Western values.

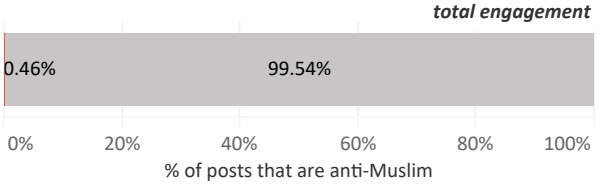
10,389
posts

2,451
active accounts

475,394
total engagement



Volume-over-time graph showing the number of anti-Muslim posts by platform



Bar chart showing the % of anti-Muslim (red) posts

Treemap showing number of anti-Muslim posts per platform

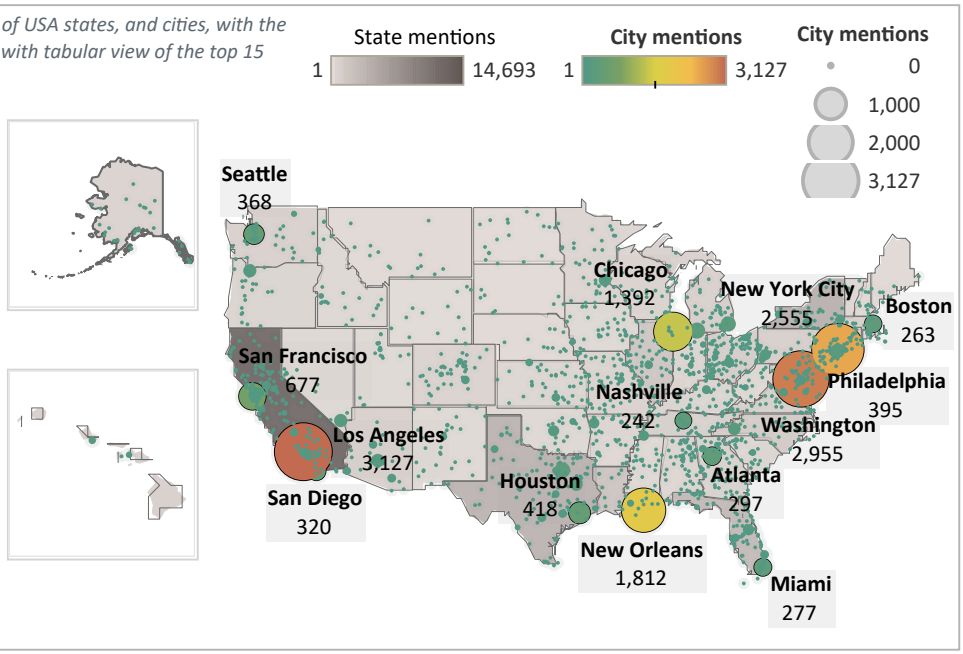
Treemap showing number of anti-Muslim posts per threat category

Domestic lens

Geographic overview

Choropleth showing the number of mentions of USA states, and cities, with the top 15 most-mentioned cities labelled; along with tabular view of the top 15 most-mentioned states

Rank	US state	Mentions
1	California	14,693
2	New York	4,871
3	Texas	4,869
4	District of Columbia	4,834
5	Florida	3,823
6	Louisiana	2,724
7	Illinois	2,383
8	North Carolina	1,600
9	Ohio	1,439
10	Pennsylvania	1,411
11	New Jersey	1,340
12	Michigan	1,255
13	Arizona	1,243
14	Washington	1,227
15	Virginia	1,082

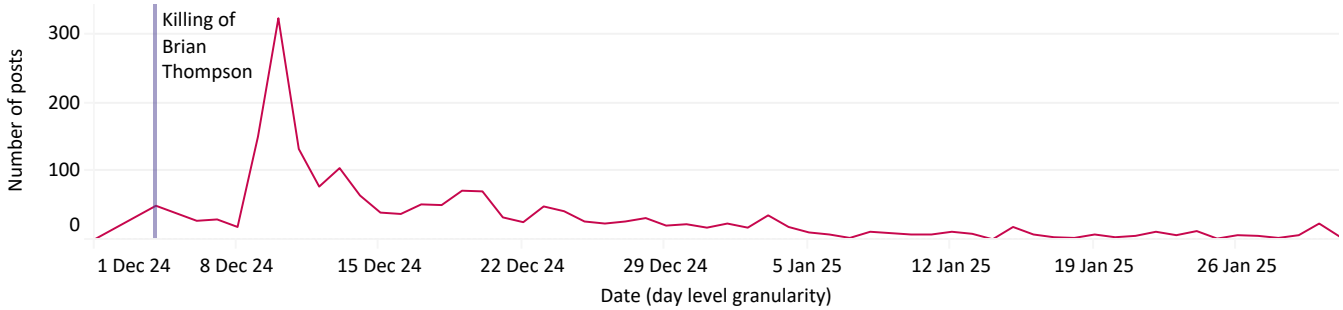


Domestic lens

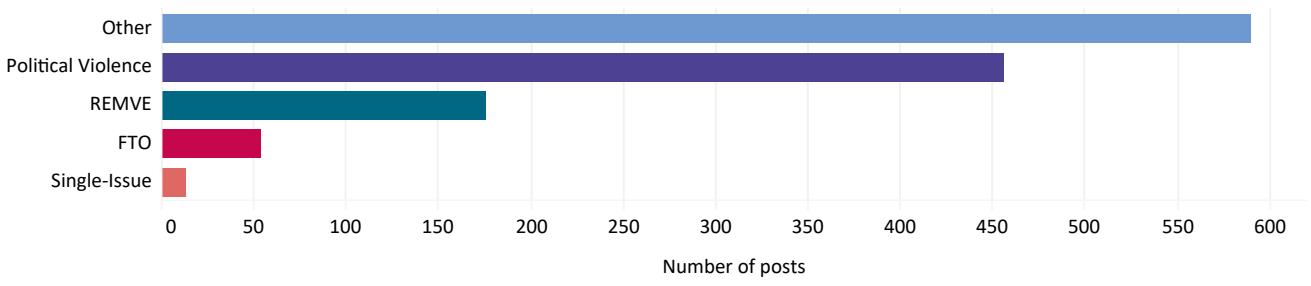
Spotlight - Violent Extremist Responses to Shooting of United Healthcare CEO

On December 4, 26-year-old Luigi Mangione allegedly assassinated United Healthcare CEO Brian Thompson, triggering strong online reactions by violent extremists and mainstream communities alike. Conversations about the attack were led by accounts promoting violence against political opponents (43%), followed closely by those without a clear ideological alignment (38%), and were particularly prominent on Telegram (35%). Collectively, violent extremists produced over 4,500 messages that contained keywords related to the incident. While some extremist constituencies rebuked Mangione's alleged actions, our analysis showed widespread support for the attack across the ideological spectrum, with numerous actors encouraging similar targeted attacks in the future. For practitioners, the broad support for Mangione suggests that violence targeted against individuals can generate more support than mass-casualty attacks, and that Mangione's actions could inspire or serve as a template for future offline violence.

Relevant posts about the incident spiked on December 10, eventually decreasing and stabilizing from December 20 onwards. An analysis of these messages revealed that violent extremists capitalized on the incident to encourage assassinations against other high-profile figures, including CEOs and politicians. Further, some violent extremists valorized Mangione by labelling him a "saint" or "hero." The 50 such posts observed within our dataset garnered 12,586 views and 9,348 likes; top-liked posts included a reference to Mangione as "the patron saint of healthcare" and "a hero of the people." Practitioners should note how anger toward the US healthcare system was used to laud Mangione's actions, in ways that resonated in violent extremist-dominated online spaces.



Volume-over-time graph showing the total number of posts, filtered to just posts mentioning at least one of "Luigi", "Mangione", or "United Healthcare" (in any casing)



Bar graph showing the number of posts per threat-category, filtered to just posts mentioning at least one of "Luigi", "Mangione", or "United Healthcare" (in any casing)

International influence

Geographic Overview

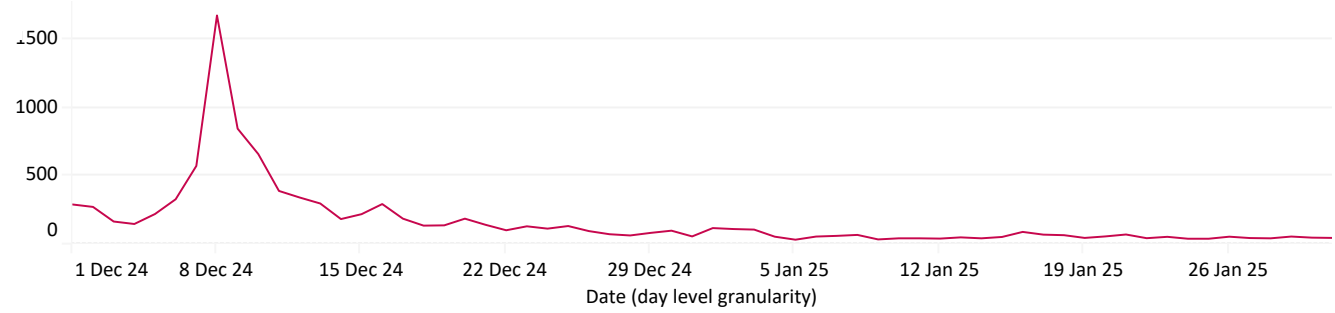
Choropleth showing the number of mentions of international countries (excluding USA), along with tabular view of the top 15 countries mentioned

Rank	Country	Mentions
1	Israel	34,388
2	Russian Federation	16,749
3	United Kingdom	15,142
4	China	13,066
5	Ukraine	12,365
6	Syrian Arab Republic	10,148
7	Canada	8,957
8	Germany	8,172
9	Palestine	6,917
10	Iran	6,142
11	India	5,576
12	Mexico	5,472
13	Australia	4,454
14	Italy	4,094
15	France	3,344

International influence

Spotlight - Violent Extremist Responses to the Fall of the Assad Regime

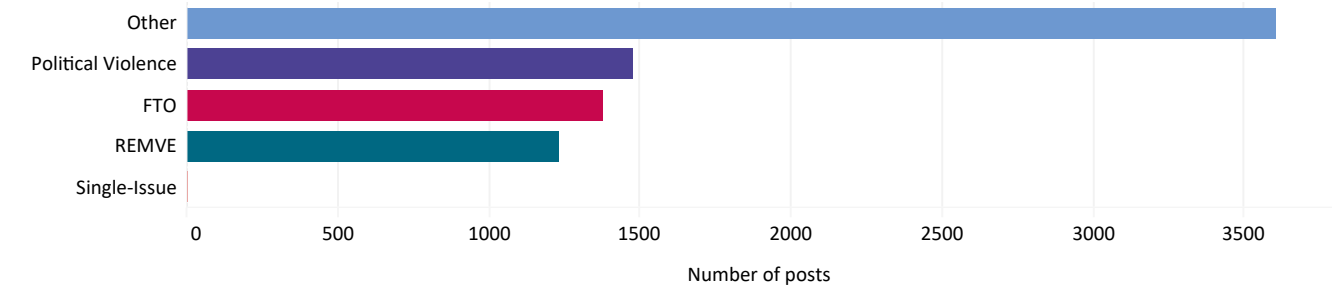
In early December 2024, a lightning offensive by jihadist group Hayat Tahrir al-Sham (HTS) successfully toppled the longstanding Assad regime in Syria. The group, considered a foreign terrorist organization by the US government due to its earlier ties to Al-Qaeda (AQ), has promised to serve as a transitional government. For practitioners, the reaction by violent extremists from across the ideological spectrum demonstrates the importance of understanding how international events and foreign policy could mobilize extremist actors to take offline action against perceived ideological opponents.



Volume-over-time graph showing the total number of posts, filtered to just posts mentioning at least one of "Syria", "Assad", or "Bashar" (in any casing)

Messages by violent extremists mentioned Syria 10,148 times during December and January, an increase of 52% compared to the previous monitoring period. Violent extremist responses to the events included conspiratorial rhetoric around the potential benefit of Assad's fall for Israel and Jews worldwide; that the CIA and "deep state" were behind the takeover; and that HTS was masquerading its continued allegiance to AQ. Pro-FTO accounts expressed concern the events benefited Israel with a weakened neighbor to its east and were a part of US-Israel-engineered regime change.

Discussion around the Assad regime's fall spiked on December 8, the day the regime fell, and returned to marginal levels (under 100 daily posts) in the days and weeks afterwards. Notable messages included one X post accusing the Israeli Mossad and its "traitorous puppets" in the US of engineering the regime change, accruing tens of thousands of likes. Analysis revealed that violent extremist accounts not affiliated with a specific ideology were most active in discussions around Syria (46%), followed by accounts promoting violence against political opponents (19%) as well as pro-FTO accounts (18%).



Bar graph showing the total number of posts, filtered to just posts mentioning at least one of "Syria", "Assad", or "Bashar" (in any casing)

Appendix

Methodology

Account Selection

ISD's and CASM's research drew on data from over 760 violent extremist accounts and forums. To be included in analysis, accounts or forums had to meet the following criteria:

1. Advocating for an extremist ideology or worldview.
2. Promoting terrorism or unlawful violence, are operated by a group or movement with a history of violence, or are supporting designated Foreign Terrorist Organizations (FTOs).
3. Are operated by individuals or groups based in the US or which produced content primarily focused on the US.

All accounts were rigorously reviewed based on these criteria by at least two expert analysts prior to inclusion, and analysts re-reviewed the existing list of accounts to ensure they still met the above threshold. To compile this list of accounts, analysts:

1. Began by using existing lists of accounts maintained by ISD from previous research into domestic violent extremism.
2. Identified additional accounts through targeted keyword searches in posts and user biographies, aimed at capturing violent extremist issues and groups across a broad ideological spectrum.
3. Expanded this existing list of accounts using network analysis to identify additional relevant accounts. Specifically, we analyze interactions between the original accounts and others they engage with, such as those they share, reply to, or mention. By identifying the most frequently linked accounts, we generate a larger pool of candidate accounts. These accounts are then manually reviewed by analysts for relevance, ensuring that only pertinent accounts are added to the updated list.

PII Removal

This work deployed technological approaches for removing personally identifiable information (PII) at the point of data collection, with several robust measures taken so sensitive data was properly anonymized while maintaining the integrity of the dataset. We focused on the removal of locations (to the zip code level and below), names, and other obvious PII like credit card information. This included both metadata and free-text fields.

- For free-text data, we employed Microsoft's Presidio anonymization tool, which is specifically designed to identify and remove PII from text. Presidio allowed us to automate the detection and removal of various PII elements, including personal names, locations, and other sensitive identifiers.
- At the same time, a curated list of over 1,000 public figures—primarily key political figures—was compiled and integrated into the process. These names were not redacted due to their analytical utility.
- Outbound URLs were shortened to preserve the information but not allow potential PII to remain in free text or metadata.

Our approach leaned towards over-removal of content to ensure compliance and protect privacy. While this occasionally resulted in the inadvertent removal of words that were not PII, the overall impact on the research was minimal.

Data Collection

Data was collected from December 1st, 2024 - January 31st, 2025.

- Through official API endpoints for Telegram, Reddit, YouTube, BlueSky and 4Chan.
- Through third-party tool BrandWatch for X, Facebook, Instagram, which, itself, employs the platform's API.
- Through third-party tool BrandWatch for forums.

Our collection and storage of data was compliant with GDPR, the US. Privacy Act, and all platforms' terms of service.

Datasets

Different subsets of data are used in this report:

1. The "wider dataset" refers to all messages collected from accounts within the report time window, and includes comments collected on Reddit posts and YouTube videos. Additionally, this dataset contains posts from forums and US tagged accounts on the 4Chan /pol/ board, messages from these sources are only included if they match hate keywords; N.B., this does not include violence-related keywords.
2. The "hate analysis subset" includes all messages from the "wider dataset", except that a 10% random sample of 4Chan is used for tractability.
3. The "sampled subset" includes a stratified sample of 372,348 messages from the "wider dataset" and is used to build the thematic classifier.

To generate the "sampled subset", data is randomly sampled from the "wider dataset" on a per-platform and message-type basis, with the aim to include a reasonable number messages from the "wider dataset" while making processing tractable.

The sampling process takes:

- 75,000 X/Twitter messages
- 75,000 Telegram messages
- 75,000 Forum posts matching hate and violence* keywords
- 65,614 4Chan messages (a random 10 percent of data matching hate and violence* target keywords)
- 236 (all) YouTube videos
- 41,590 YouTube comments (all data matching hate target keywords)
- 1,612 (all) Reddit posts
- 1,959 Reddit comments (all data matching hate target keywords)
- 565 Facebook (all) messages
- 14 (all) Instagram messages

Named Entity Recognition (NER)

This process extracts mentions of people, locations, and organizations from the text after the PII removal process to identify references to prominent figures and places above the ZIP code level.

- **Organizations**: We used a language model from SpaCy (en_core_web_lg) to automatically find all organization names in the text.
- **Persons**: We compared the text to a pre-approved list of people's names and added both the version of the name found in the text and the official name it matches.
- **Locations**: A combination of Microsoft's Presidio and CLIFF geoparser were used to identify and preserve location text that was considered sufficiently coarse granularity to not be considered PII, such as third-order administrative divisions and above and capitals of political entities. Granularity was determined using Geonames feature codes; one of ADM1, ADM1H, ADM2, ADM2H, ADM3, ADM3H, PCL, PCLD, PCLF, PCLH, PCLI, PCLS, PPLA, PPLA2, PPLA3, PPLC, PPLCH, or PPLG. For further categorising locations into countries, US states, and US cities, we applied the Mordecai3 geoparsing tool on this redacted text, which extracted:
  - o **Countries**: The country code and the original country name mentioned.
  - o **US States**: The state code and the original state name mentioned.
  - o **US Places**: The formal name of any U.S. location found and its original mention in the text.

Semantic Mapping

We followed a topic modelling approach in which semantically related messages are placed into topic clusters which are then manually assessed and broken into sub-themes (e.g. Christianity) and themes (e.g. Religion). Topic modelling represents each message

Appendix

Methodology

numerically such that messages with similar (mathematically close) representations have similar semantics. Such numerical representations are referred to as embeddings.

CASM computed embeddings using "BAAI/bge-m3" due to its open accessibility, widespread adoption and ability to handle longer text length than many other models. We began by applying semantic mapping to a sample dataset of 320,557 messages. This ultimately yielded 241 clusters of messages with semantically similar text, corresponding to varying narratives, claims, themes, tactics, etc.

For assessing topic clusters and deriving the themes and subthemes, we randomly sampled 100 messages per topic cluster. Analysts then identified sub-themes by reading through messages in the sample, and removing clusters of noisy data (e.g. clusters with just one word of text) or those with completely irrelevant discussion (e.g. skateboarding).

Having isolated dozens of analytically useful sub-themes, analysts then discussed and agreed how to group these sub-themes into overarching themes. The most relevant themes were analyzed further in a dashboard allowing analysts to layer on additional analytically useful data, such as Named Entity Recognition, engagement data, volume over time, hate classification and others.

Thematic Classification

A nearest neighbor classifier was used to assign topic clusters, subthemes, and themes to messages in the "wider dataset". In this approach, messages are classified based on their proximity to messages in the "sampled subset", analyzed in semantic mapping. We use scikit-learn's NearestNeighbors implementation with 'cosine' metric and 10 neighbors. We set a minimum cosine similarity threshold of 0.65 for neighbor retrieval. Labels are assigned using the majority label of neighbors.

Hate & Violence Classification

Hate and violence classification was performed on the "hate analysis subset"; described in the "datasets" section. The dataset contains all messages in the "wider dataset", except 10% of 4Chan is further randomly sampled for tractability. The process contained the following phases:

1) Keyword filtering.

- For antisemitic content, 164 exact words/phrases and 176 partial matches were applied to this filtered dataset to identify discussions about Jews or Judaism, resulting in 87,631 flagged messages.
- For anti-Muslim content, 56 exact keywords and 181 partial matches were used to identify mentions of Muslims or Islam, yielding 26,715 flagged messages.
- For anti-LGBTQ+ content, 229 exact keywords were used to identify mentions of the wider community or specific sub-groups (e.g. transgender individuals), yielding 24,800 anti-trans and 68,083 anti-LGBQ+ flagged messages. While analysis grouped trans and the LGBQ+ communities into a wider LGBTQ+ community, two separate pipelines were used for each of these two sub-groups. This is due to wide linguistic differences between discussion of LGBQ+ and transgender individuals and associated issues.
- For violent content, 362 exact words/phrases were used to identify candidate mentions of violence, yielding 133,053 flagged messages.

2) Classifier Selection and Evaluation

Samples of messages from each of the keyword-filtered datasets (Antisemitic, anti-Muslim, anti-LGBQ+, anti-trans, and violence) were manually labelled. The labelled samples were then used to evaluate a combination of LLM models, prompts (tailored to each task using keywords and examples), and hyperparameters to optimize the precision and recall for each classification task. The precision and recall for hateful messages for each classifier, as evaluated on an evaluation set is as follows:

- Antisemitic: 0.78 precision, 0.91 recall
- Anti-Muslim: 0.86 precision, 0.84 recall
- Anti-LGBQ+: 0.74 precision, 0.88 recall
- Anti-trans: 0.88 precision, 0.97 recall
- Violence: 0.81 precision, 0.80 recall