ISD | Powering solutions to extremism, hate and disinformation

CASM technology

US Online Domestic Violent Extremism Monitor:

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

August-September 2024

ISD Powering solutions to extremism, hate and disinformation | CASM technology

Bi-monthly **Report #1, August-September 2024**

US Online Domestic Violent Extremism Monitor

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

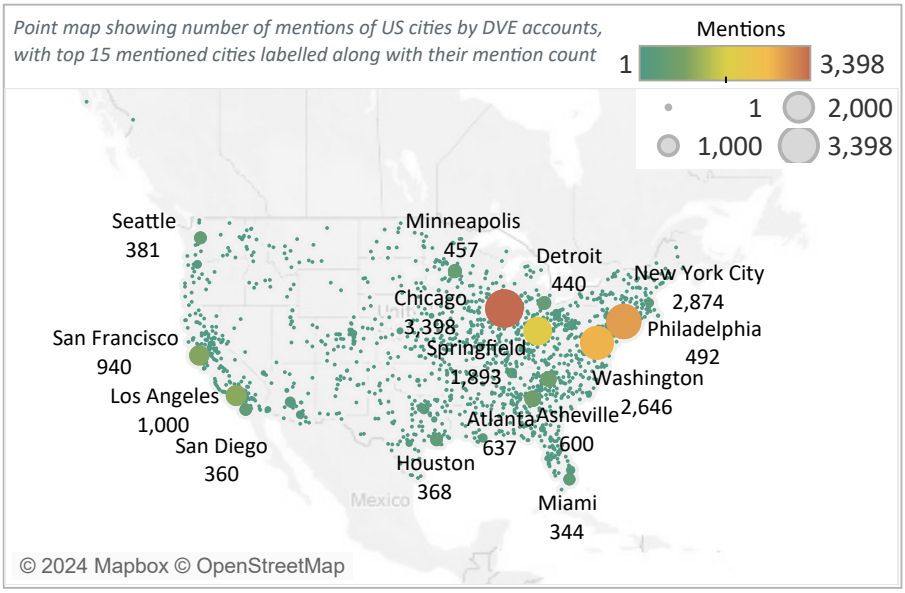
Executive Summary

This bi-monthly data snapshot overviews key trends in US domestic violent extremist (DVE) activity online, providing an up-to-date picture of the evolving digital threat landscape. This report is divided into the following sections: (1) key platforms; (2) threat actors; (3) semantic mapping of key topics; (4) targeted hate; (5) domestic and international events animating activity; and (6) methodology. Analysis is based on a dataset of around 1,000 US-linked accounts and channels across numerous platforms and ideologies that were manually vetted by experts as engaging in clear violent extremist behavior. All insights are anonymized and presented in aggregate, with all personally identifiable information removed at the point of data collection. By identifying key themes and platforms, researchers aim to provide practitioners with data that can help inform prevention efforts. Understanding the relationship between patterns in online DVE activity and real-world developments can help practitioners more efficiently prioritise resources. Analysis of the evolving narratives espoused by DVEs might allow practitioners to better understand potential entry points for helping individuals disengage from violence. Further, cross-platform data can inform practitioners' understanding of the different risks across online spaces, and where prevention efforts might be focused.

Findings

During August and September 2024, our analysis showed the following key trends:

Point map showing number of mentions of US cities by DVE accounts, with top 15 mentioned cities labelled along with their mention count



Platform Activity:

- In a dataset of over 2.4 million messages across nine social media platforms (X, Telegram, YouTube, Reddit, 4chan, Gab, Rumble, Bitchute and Odysee), DVE accounts were most active on Telegram, while the highest number of active DVE accounts was on X. 4chan's notorious /pol/ board received over 500,000 posts using language indicative of extremism or targeted hate from US users during this period.
- DVE-linked accounts were less active on YouTube and Reddit, but generated substantial comment activity, with nearly 350,000 comments on videos from DVE channels and over 16,000 comments in DVE subreddits.

Threat Categories:

- Racially and Ethnically Motivated Violent Extremist (REMVE) accounts were most active, constituting 35% of total posting activity. Anti-Government & Anti-Authority Violent Extremist (AGAAVE) accounts were next most active but generated by far the most engagement.
- 'Other' domestic violent extremists were mainly motivated by a wide range of conspiracy theories, alongside extreme misogyny and anti-LGBTQ+ hate. The Middle East conflict produced spikes in posting activity among Single-Issue Violent Extremist accounts .

Key Narratives:

- Topic modelling analysis based on trained classifiers showed that DVE accounts were most animated by discussions around targeted hate (51% of classified messages), government and politics (20%), and religion (10%).
- Specific narratives that drew significant engagement from DVEs included conversations around elections and voting (including voter fraud narratives), discussions around Christianity and Islam, as well as conspiracy theories around law enforcement.

Targeted Hate:

- Trained hate speech classifiers showed that 4chan was the platform with the highest proportion of antisemitic language, where a third of a filtered sample of messages from US users on the extremist /pol/ board were overtly antisemitic in nature.
- The vast majority of anti-Muslim hate was found on 4chan and Telegram, evenly distributed across the different DVE threat categories. Anti-LGBTQ+ sentiment was the most common form of hate on YouTube and Telegram.

Domestic and International Influences:

- Domestically, locations in California, Illinois, Texas, New York and Florida were most mentioned by violent extremists. Domestic events galvanizing DVEs included falsehoods spread around Haitian migrants in Springfield eating pets, prompting a spike in anti-migrant hate, as well as antisemitic conspiracy myths rooted in the Great Replacement theory.
- Internationally, US DVEs spoke most about events in Israel, Russia, UK, Ukraine and China. The arrest of Telegram founder Pavel Durov in France prompted a significant reaction among US DVEs, including claims of government censorship and discussions around potential alternatives to the encrypted messaging platform.

Approach

Data was collected through platforms' public Application Programming Interfaces (APIs) with the exception of Gab, Bitchute, Rumble, and Odysee which were collected via third-party tool Pyrra. Data availability differs considerably by platform, with the shuttering of Meta's CrowdTangle tool preventing meaningful analysis of Facebook and Instagram during the period of study. Data analysis (explained in full in the methodological annex) incorporates 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 hate speech targeting minority communities. Going forward, this bi-monthly snapshot will more systematically compare domestic and international trends in violent extremist activity, widen platform coverage, and extend hate speech analysis. 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

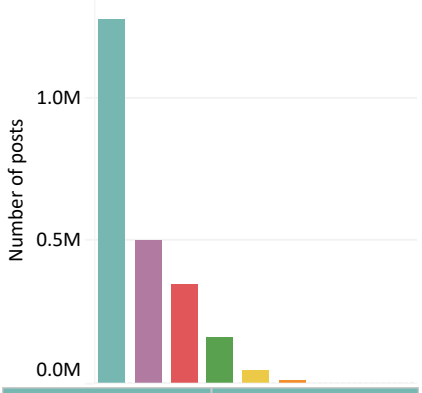
Overview

With over a million messages from DVE accounts, Telegram was the most active platform in the dataset. Its loose moderation and reputation for secure communications enables channels and group discussions where extremist networks disseminate messages to dedicated followers. However, the arrest of Telegram CEO Pavel Durov has sparked concerns over platform security, provoking discussions among DVEs about migrating to more secure applications.

4chan's /pol/ board has played a longstanding role as a key forum for anonymous and unmoderated discourse among online extremists. Rather than collecting from the entire board, for this analysis we collected data from US-based users, using keywords related to hate and extremism or associated with groups regularly targeted on the forum. Based on this criteria, /pol/ was the second most active platform in our analysis, although the anonymous and amorphous nature of 4chan precluded analysts from classifying users into specific threat categories, as with other platforms.

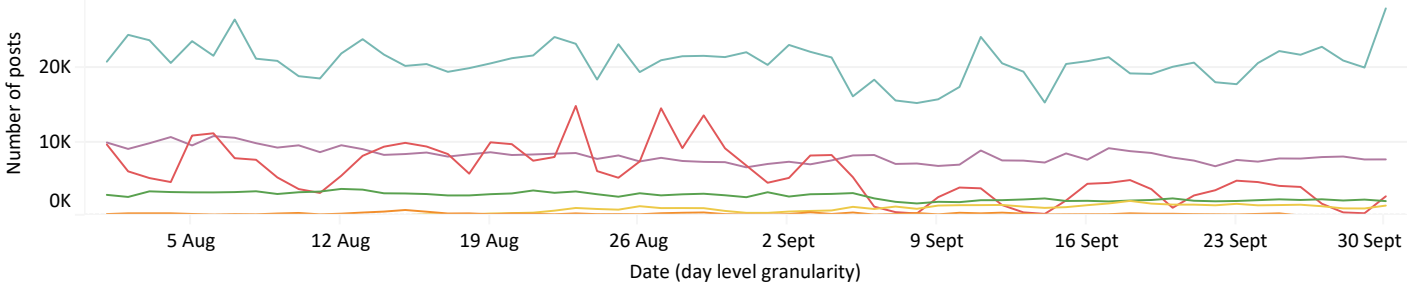
With DVE-relevant posts in the hundreds of thousands, X (formerly Twitter) has become a significant hub for domestic violent extremist activity, with notable spikes in activity occurring around political events such as the August 12 interview between former president Donald Trump and X CEO Elon Musk and the September 10 presidential debate between Trump and Vice President Kamala Harris.

Our analysis included tens of thousands of messages from Reddit and Gab, as well as thousands from video sharing platforms Rumble, Bitchute as well as slightly less than a thousand posts from fellow alternative video streaming service, Odysee. YouTube saw less activity, whether due to stronger moderation or the increased effort required to create videos compared to the shorter message-based data of other platforms. However, hateful content still proliferated in the comment sections of videos from DVE users.



Telegram	1,278,104
4Chan	502,162
YouTube	348,820
X (Twitter)	161,446
Gab	48,806
Reddit	16,349
Rumble	2,762
Bitchute	2,291
Odysee	859

Bar graph & corresponding tabular view showing total posts per platform

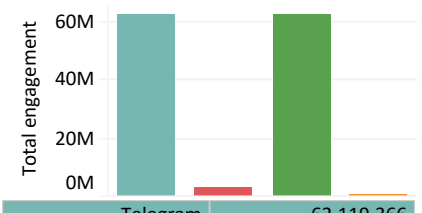


Volume-over-time graph showing number of posts per platform

Platforms

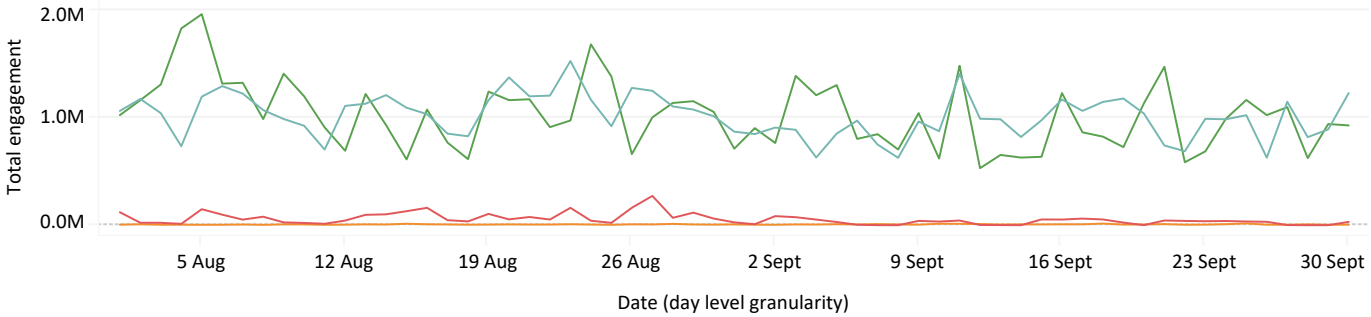
Engagement

The types and availability of engagement data (e.g. likes, comments, shares) varies dramatically by platform but can provide insight into the traction gained by DVE content. X had the most significant engagement during the reporting period, with DVE accounts commanding nearly 14 million followers and receiving over 40 million likes. Mirroring Telegram's continued centrality in violent extremist mobilization internationally, content from the platform's DVE channels was liked over 12.6 million times and received a cumulative view count of over 5.3 billion. YouTube videos from DVE-linked channels received over 108,000 comments and nearly 2.4 million likes, while posts in DVE-relevant subreddits were liked almost 350,000 times and received over 17,000 comments. Gab, 4chan, Rumble, Bitchute, and Odysee did not provide user engagement data.



Telegram	62,119,366
YouTube	3,339,617
X (Twitter)	62,479,375
Reddit	446,757

Bar graph & corresponding tabular view showing total engagement per platform

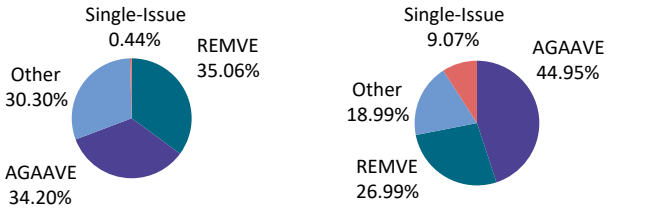


Volume-over-time graph showing total engagement per platform

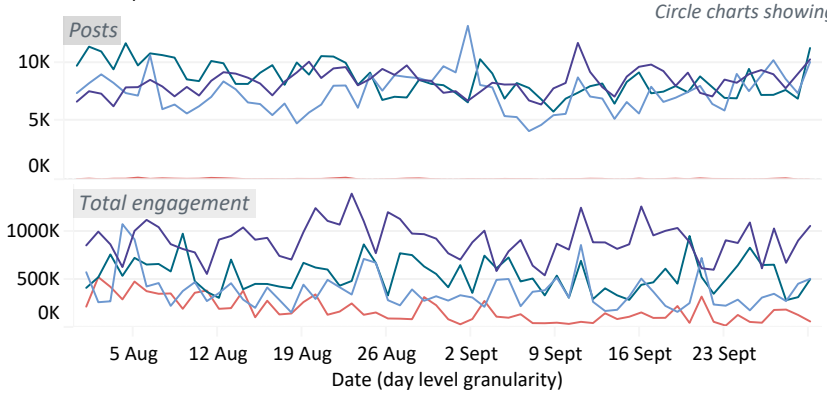
Threat categories

Overview

Drawing on terminology outlined by DHS and other US government agencies, ISD manually classified DVE accounts and channels into four threat categories: (1) Racially or Ethnically Motivated Violent Extremism; (2) Anti-Government or Anti-Authority Violent Extremism; (3) Single-Issue Violent Extremism (sub-categorized into Animal Rights, Environmental, Abortion, and Israel/Palestine-related Violent Extremism); and (4) Other Domestic Violent Extremism (including violent conspiracy movements).



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

Because of how data was collected, the snapshots in this section exclude analysis from 4chan's /pol/ board, 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 sub-Reddit-level respectively, rather than by individual commenters.

Anti-Government or Anti-Authority Violent Extremism (AGAAVE)

511,226 *posts*
154 *active accounts*
56.01M *total engagement*

AGAAVE encompasses the potentially unlawful threat of violence in furtherance of ideological agendas derived from anti-government or anti-authority sentiment, including opposition to perceived economic, social, or racial hierarchies, or perceived government overreach, negligence, or illegitimacy. During the review period, AGAAVE accounts were the second-most active among the threat categories based on messages produced and number of active accounts. Despite this, content produced by these accounts constituted nearly 74% of the views across all threat categories, showing considerable engagement. The overwhelming majority of AGAAVE content and active accounts was comprised of actors motivated by grievances against their perceived political enemies, including US political parties. In many instances, these same actors appeared to have been motivated by the QAnon movement.

Racially or Ethnically Motivated Violent Extremism (REMVE)

523,953 *posts*
373 *active accounts*
33.62M *total engagement*

REMVE encompasses the potentially unlawful use or threat of force or violence in furtherance of ideological agendas derived from bias, often related to race or ethnicity, held by the actor against others or a given population group. During the monitoring period, REMVE was the most active threat category in both messages produced and number of active accounts. The majority of the content produced within this threat category was posted by actors who espouse racial or ethnic grievances but who do not identify as belonging to a particular group. However, neo-Nazi accelerationist groups produced nearly 11% of REMVE messages, making them the most active discernable grouping.

Single-Issue Violent Extremism

6,503 *posts*
31 *active accounts*
11.30M *total engagement*

This threat type 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, a new category which ISD defines as domestic violent extremist actors who are singularly motivated by the ongoing conflict between Israel and Hamas. In August and September, actors motivated by Single-Issue Extremism were the least active among the four threat categories and produced the fewest posts by an order of magnitude. The most active accounts in this category were motivated by Israel/Palestine, followed by those motivated by abortion and environmental issues. Accounts motivated by animal rights comprised a small subset of this category.

Other Domestic Violent Extremism

452,931 *posts*
55 *active accounts*
23.66M *total engagement*

This category encompasses threats involving the potentially unlawful use or threat of force or violence in furtherance of ideological agendas which are not otherwise defined under or primarily motivated by one of the other domestic threat categories. In August and September, actors motivated by Other Domestic Violent Extremism were the third most active threat category in messages produced and number of active accounts. Within this category, 17% of accounts were primarily driven by manosphere or misogyny-related content, including anti-LGBTQ+ hate , and 57% of the accounts were motivated by a wide range of conspiracy theories. The remainder were driven by niche sub-communities or a complex mixture of grievances that defied clear categorization (such as the blending of neo-Nazi and Salafi-jihadist ideologies).

Semantic mapping

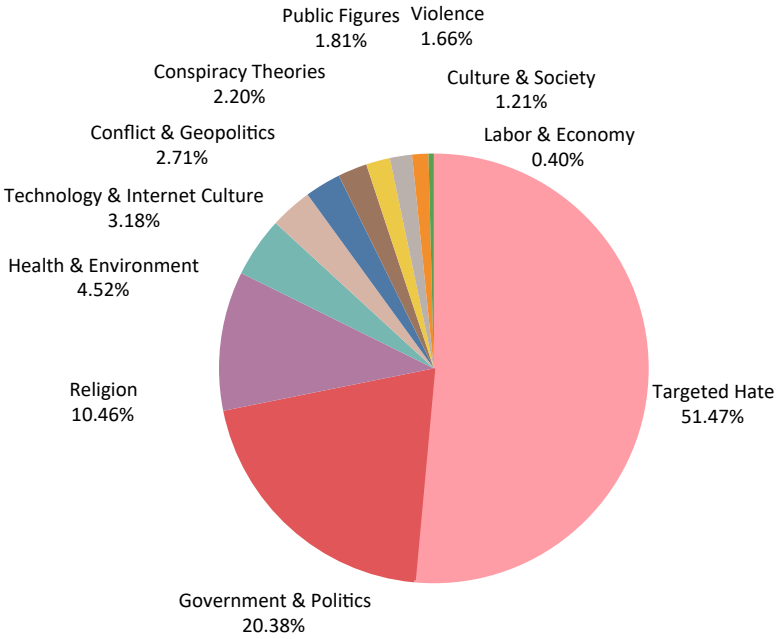
Summary

Bespoke Large Language Models (LLMs) were used to group messages into semantically distinct clusters to streamline the analysis of key narratives. Analysis identified 11 overarching themes, encompassing 72 sub-themes. For example, within the Health & Environment theme, the LLM identified seven sub-themes that generated significant conversation, including discussions around abortion, vaccines and climate.

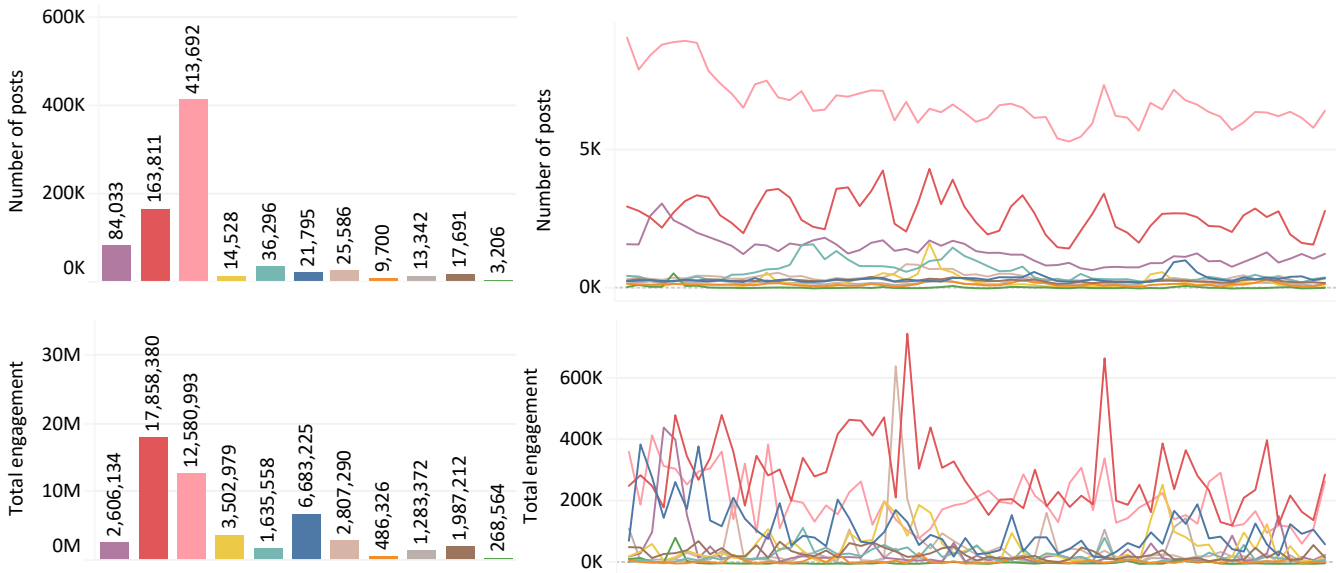
While many messages posted by DVEs were grouped into themes that were not directly related to hate, extremism, or violence, they were still deemed relevant for understanding the key narratives underpinning online DVE content. Further, semantic clusters were not treated as discrete categories—for example, antisemitic posts falling under the Targeted Hate theme may be imbued with conspiratorial rhetoric or may include attacks on public figures.

Of the 11 master themes, discussions relating to Targeted Hate were the most common theme identified within the data, followed by conversations around Government & Politics and then Religion. DVE messages about Government and Politics gained the most engagement, followed by Targeted Hate, Conflict and Geopolitics then Technology & Internet Culture.

Analysis of the most prominent sub-themes showed that messages relating to antisemitism and homophobia (both examples of it and discussions about it) were the most frequent in the dataset and are analyzed in more granularity in the following section. Beyond Targeted Hate-related topics, conversations around Elections and Voting predominated, and represented one of the most engaged with topics. Discussions of Christianity were prominent in the dataset, but received less engagement than content relating to Law Enforcement.



Circle chart showing the % of posts by theme



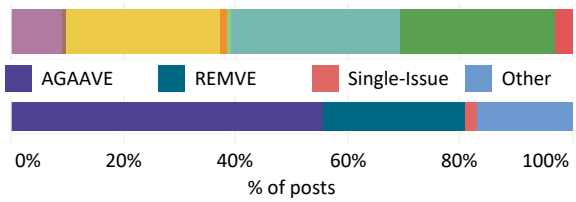
Bar charts showing the number of posts (top) and total engagement (bottom) by theme

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

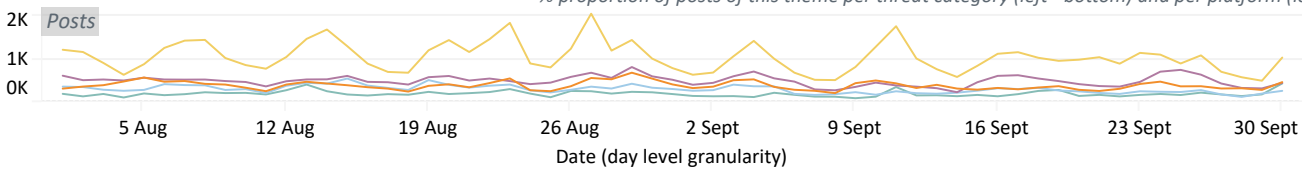
Government & Politics

With the US presidential election nearing, discussions around Elections and Voting constituted just under 10% of all DVE conversations classified in our dataset. The majority of this conversation was driven by AGAAVE accounts, including high engagement posts spreading narratives about voter fraud and election interference. This was followed by discussion about Law Enforcement, where salient narratives included dehumanizing representations of police, as well as conspiracy theories about the Secret Service. Most DVE-related discussion around Government & Politics was found in YouTube comments as well as Telegram, where such conversations gained the most engagement by some margin.



Rank	Sub-theme	Posts	Accounts
1	Elections and voting	64,292	27,745
2	Law Enforcement	30,635	10,374
3	Immigration	23,924	10,531
4	Political Ideology	18,589	9,120
5	Trump	12,050	2,347

Table showing the top 5 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)

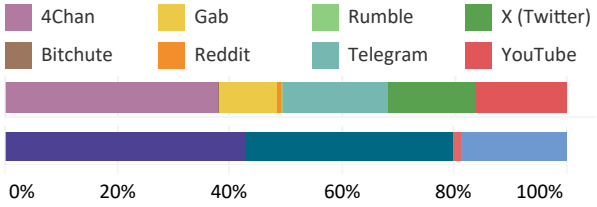


Volume-over-time graph showing the number of posts by sub-theme, for the top 5 sub-themes (by post count) for this theme

Semantic mapping - key themes

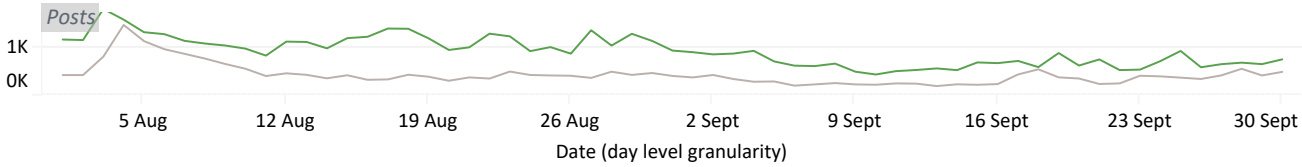
Religion

Discussions of religion comprised around 1 in 10 messages with an identifiable theme from DVE accounts. The majority of religion-themed messages were focused on Christianity, mostly from AGAAVE accounts engaging in general religious discussion. Around half as many posts were focused on Islam, with many high traction messages presenting Muslims as regressive or dangerous. Notably DVE discussions around Islam generated more than twice as much total engagement as content focused on Christianity, with most of this generated by REMVE and 'Other' violent extremist actors. Discussions around both faiths peaked during the UK riots at the beginning of August, where numerous DVE actors sought to present Muslim immigration as a threat to Christian identity.



Rank	Sub-theme	Posts	Accounts
1	Christianity	57,019	27,007
2	Islam	27,014	18,103

Table showing the top 2 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)

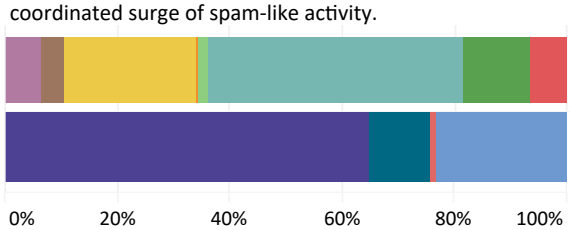


Volume-over-time graph showing the number of posts by sub-theme, for the top 5 sub-themes (by post count) for this theme

Semantic mapping - key themes

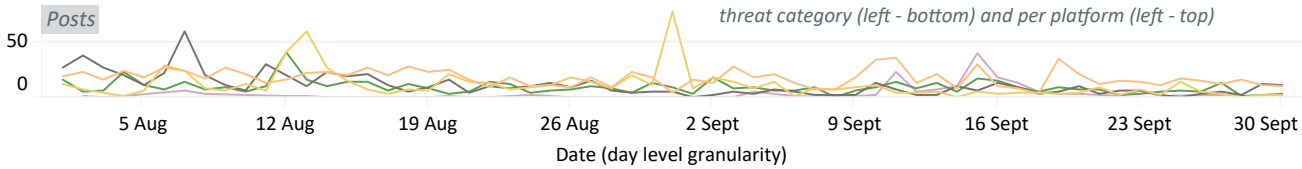
Public Figures

Drawing on Named Entity Recognition approaches, our analysis showed prominent political figures dominated mentions of public figures during August and September. References to Vice President Kamala Harris among DVE accounts peaked during her presidential debate with Donald Trump. In contrast, the highest number of mentions of Donald Trump occurred on the day the Former President returned to X after a year off the platform, ahead of an interview with owner Elon Musk. This included conspiracy theory-based DVE accounts using QAnon language to claim that the "The Storm is Upon Us." Mentions of Elon Musk peaked on the day of this interview, with a later spike largely attributed to a coordinated surge of spam-like activity.



Rank	NER-entity	Posts	Accounts
1	Kamala Harris	1,132	259
2	Elon Musk	746	368
3	J. D. Vance	711	407
4	Donald Trump	613	280
5	Taylor Swift	208	103

Table showing the top 5 Named Entity Recognition (NER) person entities (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 NER-person-entity, for the top 5 NER-person-entities (by post count) for this theme

Targeted hate

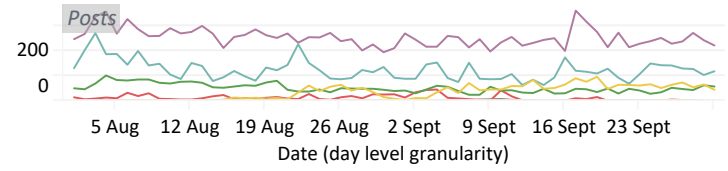
Antisemitism

ISD analysts employed bespoke Large Language Models (LLMs) to analyze the scale and nature of explicit antisemitic, anti-Muslim and anti-LGBTQ+ hate speech within the dataset. Overall, 3.1% of message data included at least one of these forms of overt speech hate. 4chan hosted most of the antisemitism in our dataset, with a third of a sample of /pol/ posts from US users containing antisemitism (although these messages were already filtered by a set of keywords likely to surface extremism or targeted hate). This was followed by X, Bitchute, Gab and Odysee, which all hosted a similar proportion of antisemitism (2-4% of messages). Almost 60% of antisemitic messages came from REMVE actors, while a quarter came from 'Other' violent extremists (largely violent conspiracy theorists). Antisemitic posts with the highest engagement include classical antisemitism claiming Jewish religious texts permit child abuse, calls for deportations of Jews, and conspiracy theories around Jewish control of media, financial and government institutions.

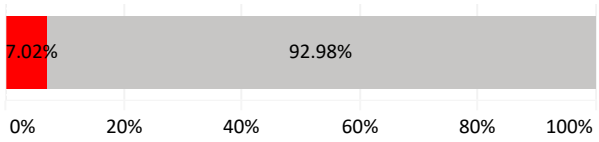
29,266
posts

15,499
active accounts

2,411,066
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

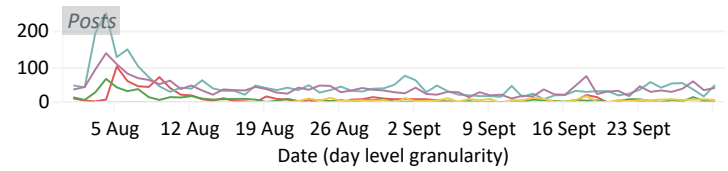
Anti-Muslim Hate

4chan was the platform with the highest proportion of anti-Muslim hate content (which was present in around 1 in 16 messages), and along with Telegram comprised 75% of anti-Muslim hate messages across the dataset. Anti-Muslim hate was evenly spread across DVE actors associated with REMVE, AGAAVE and 'Other' violent extremism threats, and was significantly less visible in the online activity of Single-Issue violent extremists. High engagement Islamophobic narratives include anti-Muslim racist slurs, the associations of Muslims in general with sexual violence and terrorism, and calls to ban Islam in the US.

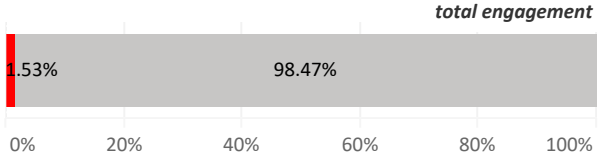
7,652
posts

3,597
active accounts

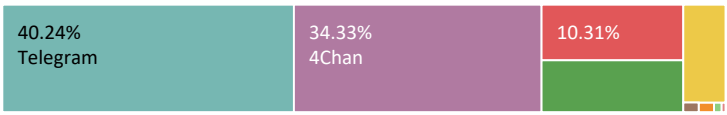
517,365
total engagement



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

Targeted hate

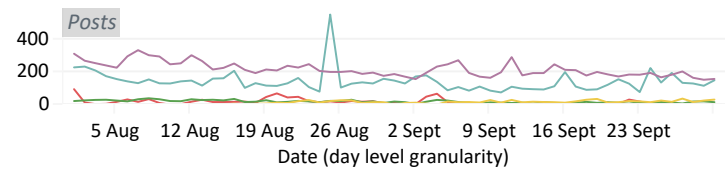
Anti-LGBTQ+ Hate

Almost a quarter of our 4chan filtered sampled included anti-LGBTQ+ hate, showing its highly mainstreamed nature on the /pol/ imageboard. While containing a lower proportion, anti-LGBTQ+ sentiment was the most common form of hate speech detected among DVE accounts on Telegram and YouTube. Around half of this content originated from REMVE actors, with just under a third from 'Other' violent extremists, which includes accounts motivated to violence by misogynistic, homophobic and transphobic grievances. Hateful narratives with the highest engagement include associations of LGBTQ+ public figures with grooming, degeneracy or mental illness, as well as conspiracy theories presenting LGBTQ+ communities as part of a global plot to undermine the traditional social order.

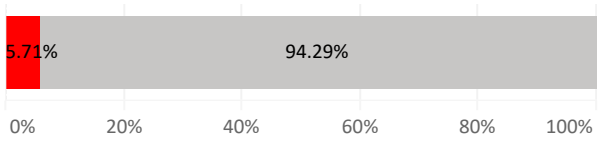
25,507
posts

14,496
active accounts

1,238,296
total engagement

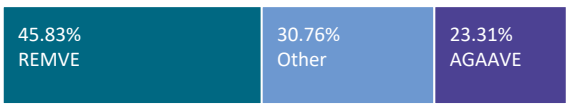


Volume-over-time graph showing the number of anti-LGBTQ+ posts by platform



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

Treemap showing number of anti-LGBTQ+ posts per platform

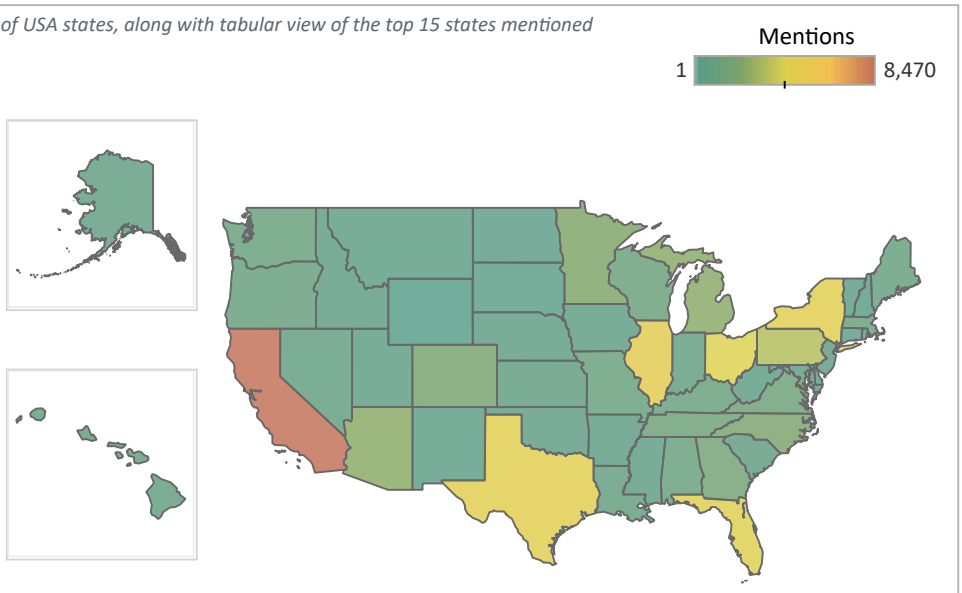


Treemap showing number of anti-LGBTQ+ posts per threat category

Domestic lens

Geographic overview

Choropleth showing the number of mentions of USA states, along with tabular view of the top 15 states mentioned



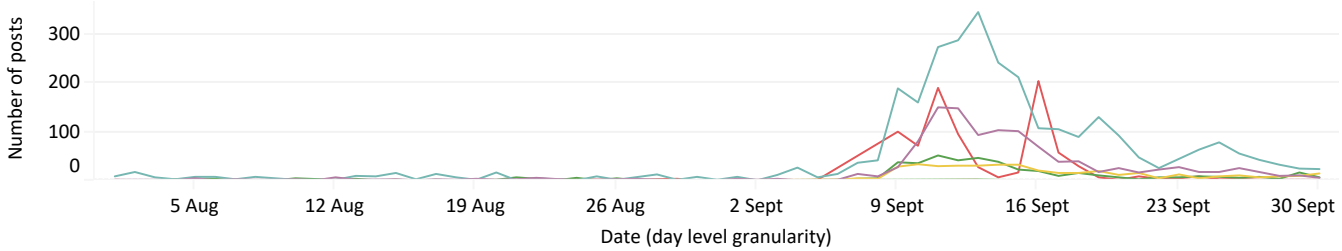
Rank	US state	Mentions
1	California	8,470
2	Illinois	4,910
3	Texas	4,757
4	New York	4,731
5	Florida	4,455
6	Ohio	4,259
7	District of Columbia	3,392
8	Pennsylvania	3,321
9	Arizona	2,395
10	Michigan	2,390
11	Minnesota	2,122
12	North Carolina	1,983
13	Colorado	1,823
14	Georgia	1,361
15	Tennessee	1,219

Domestic lens

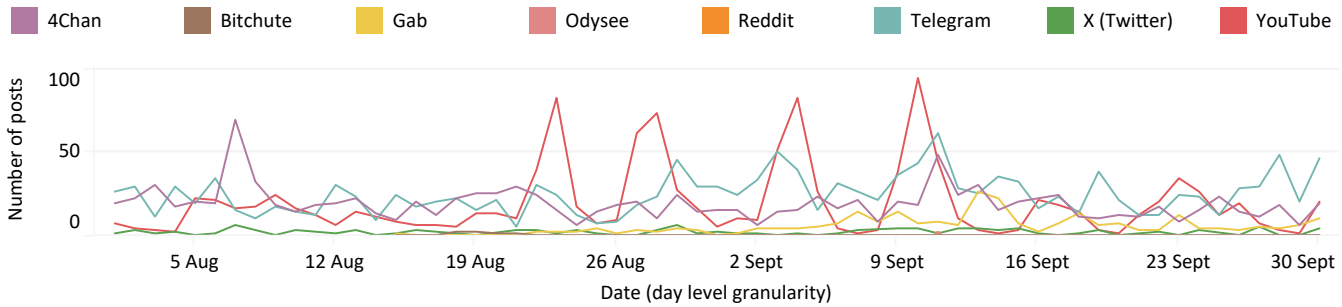
Spotlight - Anti-Migrant Targeted Hate in Springfield, Ohio

In early-mid September, there was a sharp increase in online discussions among DVEs around Haitian migrants in Springfield, Ohio. Springfield came into the spotlight in early September when online posts circulated claiming that Haitian migrants were eating residents' pets. The September 10 presidential debate sparked even further discussion on this topic, which was largely driven by AGAAVE accounts on Telegram. Analysis showed how this spike in discussion was likely in response to misinformation concerning the alleged disappearance of pets in Springfield, a narrative promoted by leading political figures both online and offline.

Anti-migrant rhetoric jumped after September 10, with many posts describing migrants as "invaders." References to "illegals" peaked on Telegram and YouTube for the two-month period on September 11, and saw the second highest number of mentions during this period on 4chan (with the highest during the UK riots in August) . Several posts featured violent rhetoric targeting migrants, alongside allegations that noncitizens would be voting in upcoming elections. On September 11, there was a significant increase in DVE activity on 4chan, with numerous posts mentioning "Haitians", including narratives dehumanizing Haitians, often using racist slurs, alongside antisemitic narratives claiming Jews were "flying Haitian cannibals into white American towns". On X, the most engaged post came from a violent conspiratorial account with a large following which speculated that politicians and law enforcement had been "paid off" by Jewish NGOs to "destroy" Springfield with Haitians. YouTube comments that generated substantial engagement mentioned black gangs who eat and sacrifice pets as part of "voodoo" rituals. Activity from DVE actors concerning Springfield started declining in the second half of the month, although it continued to feature as a major topic on Telegram, with one post at the end of September celebrating protests featuring the slogan "Haitians Have no Home Here", which received over 500 likes.



Volume-over-time graph showing the number of posts by platform, filtered to just posts mentioning at least one of "Springfield" or "Haiti" (in any casing)



Volume-over-time graph showing the number of posts by platform, filtered to just posts mentioning "illegals" (in any casing)

International influence

Geographic Overview

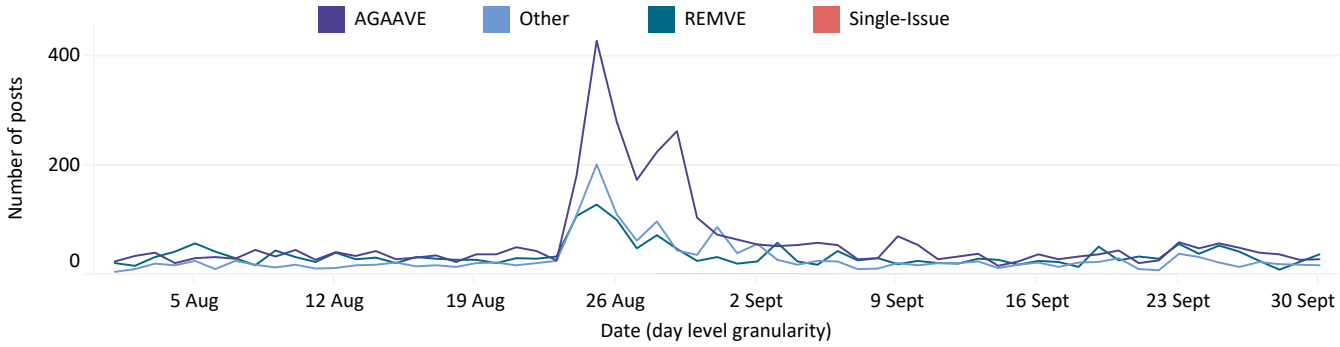
Choropleth showing the number of mentions of international countries (excluding USA), along with tabular view of the top 15 countries mentioned

Rank	Country	Mentions
1	Israel	37,463
2	Russian Federation	20,169
3	United Kingdom	18,032
4	Ukraine	13,628
5	China	10,315
6	Iran	9,916
7	Germany	6,144
8	France	3,960
9	Canada	3,747
10	Italy	3,532
11	India	3,515
12	Palestine	3,396
13	Lebanon	3,247
14	Mexico	2,903
15	Australia	2,703

International influence

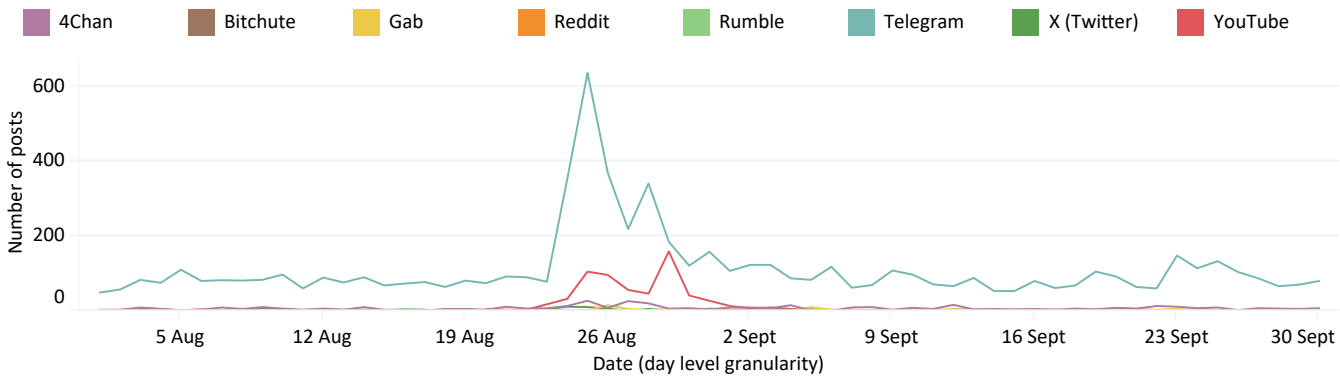
Spotlight - Arrest of Telegram's Founder and CEO

The arrest of Telegram founder and CEO Pavel Durov by French authorities on August 24 triggered a strong reaction among US DVEs, which was exacerbated by his announcement on September 23 that Telegram would begin providing certain user data in response to law enforcement requests. In the period following the arrest and the policy change announcement, there was a considerable increase in messages from US DVE accounts mentioning keywords related to these developments, with the majority coming from Telegram itself.



Volume-over-time graph showing the number of posts by threat category, filtered to just posts mentioning at least one of "Pavel", "Durov", or "Telegram" (in any casing)

Analysis revealed that US DVEs cited the arrest as evidence that free speech was under attack globally and that Americans needed to fight back against widening government censorship. Posts that accrued significant engagement included a Telegram message posted by a channel with tens of thousands of subscribers claiming that the arrest was part of a war being carried out by the US Department of State against its political enemies abroad. Further, a QAnon-related Gab account with nearly 8,500 subscribers indicated that the arrest was conducted further to orders made by the US government to censor criticism of Ukraine. Meanwhile, ISD witnessed a flurry of posts by DVEs suggesting new platforms where they could safely conduct their activities, with several neo-Nazi accelerationist accounts recommending the messaging platform SimpleX as a suitably secure alternative.



Volume-over-time graph showing the number of posts by platform, filtered to just posts mentioning at least one of "Pavel", "Durov", or "Telegram" (in any casing)

Appendix

Methodology

Account Selection

ISD's and CASM's research drew on data from around 1000 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. 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 August 1st – September 30th, 2024:

- Through official API endpoints for Telegram, Reddit, YouTube and 4Chan.
- Through third-party tool BrandWatch for X, which, itself, employs the platform's API.
- Through third-party tool Pyrra for Gab, BitChute, Rumble, and Odysee, which do not provide API access and does not preclude such data collection.

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" includes all messages collected throughout the report time window. 4Chan is included along actor-based collections, this dataset is composed of messages posted to the /pol/ board, posted by US tagged accounts, and matching hate target keywords.
2. The "hate analysis subset" includes all messages collected throughout the report time window, except that a 10% random sample of 4Chan is used for tractability.
3. The "sampled subset" includes 305,106 messages taken 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 of messages from the "wider dataset" while making processing tractable. The sampling process takes:

- 51,468 4Chan messages (a random 10 percent of data matching hate target keywords)
- 100,000 X/Twitter messages
- 100,000 Telegram messages
- 360 (All) YouTube videos
- 11,827 YouTube comments (all data matching hate target keywords)
- 1,447 (All) Reddit posts
- 649 (All) Reddit comments (all data matching hate target keywords)
- 33,122 (All) Gab posts
- 2,924 (All) Rumble posts
- 949 (All) Odysee posts
- 2,340 (All) BitChute posts

Named Entity Recognition (NER)

This process is integrated with PII removal process to extract mentions of people, locations, and organizations from the text, identifying 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 (i) countries: the country code and the original country name mentioned; (ii) US States: the state code and the original state name mentioned; (iii) US Cities: the formal name of any U.S. city combined with its state code, 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 numerically such that messages with similar (mathematically close) representations have similar semantics. Such numerical

Appendix

Methodology

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 the "sampled subset" of 305,106 messages. This ultimately yielded 271 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 11 overarching themes. The "wider dataset" was then classified with thematic annotations for analysis, described below. The most relevant themes – Government & Politics, Public Figures, Violence, and Conspiracy Theories – 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 neighbour 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", analysed in semantic mapping. We use scikit-learn's NearestNeighbors implementation with 'cosine' metric and 10 neighbours. We set a minimum cosine similarity threshold of 0.65 for neighbour retrieval. Labels are assigned using the majority label of neighbours.

Hate Classification

Hate classification was performed on the "hate analysis subset" (all messages collected throughout the report time window, except 10% of 4Chan is further randomly sampled for tractability). The process contained the following phases:

1. For antisemitic content, 164 exact words/phrases and 176 terms for partial matching were applied to this dataset to identify discussions about Jews or Judaism, resulting in 12,887 flagged messages with at least one match. For anti-Muslim content, 56 exact words/phrases and 181 terms for partial matching were used to identify mentions of Muslims or Islam, yielding 3,139 flagged messages. For anti-LGBTQ+ content, a total of 245 exact words/phrases were applied resulting in 56,717 total messages containing mentions of people contained within the LGBTQ+ community or the LGBTQ+ community itself. The LGBTQ+ messages were broken down into different target categories;

LGBTQ+ community: 38 exact words/phrases
Gay: 49 exact words/phrases
Lesbian: 22 exact words/phrases
Bisexual: 5 exact words/phrases
Queer: 48 exact words/phrases
Trans: 83 exact words/phrases.

Moving forward it was decided that messages with trans as the target would be treated separately, as the hate directed towards them tends to be phrased differently than others within the LGBTQ+ community. This resulted in 13,853 messages directed towards trans people and 45,387 messages directed towards the rest of the LGBTQ+ community.

All messages filtered with keywords are non-exclusive, meaning a message can be targeting multiple groups.

2. The filtered messages were further processed using a generative language model, Llama 3.1 ("grimjim/Llama-3.1-8B-Instruct-abliterated_via_adapter") from Huggingface, using a separate prompt designed for each target group.

The following prompt was used for antisemitic hate speech classification:

You are an AI language model trained to identify and classify comments that SPECIFICALLY contain CLEAR antisemitic hate speech.

Antisemitic hate speech means CLEAR and EXPLICIT hate speech SPECIFICALLY directed towards Jewish people, if there is doubt then it is not antisemitic hate speech.

Comments are NOT antisemitic hate speech if they are merely rude, offensive or distasteful. Likewise just stating the words 'jew' or 'jews' is not antisemitic hate speech.

Respond with 1 if the comment contains EXPLICIT antisemitic hate speech.
Respond with 0 if the messages do not meet the criteria of the 1 classification.

For each of the ^ comments in the provided text you will return a breakdown of your classification in a json format with the classification, followed by a confidence score ranging between 0 to 1 based on how likely you think the classification is correct, followed by an explanation for the classification:
{
"Comment Number":,
"Classification":,
"Confidence score":,
"Explanation":
}

You MUST respond to EXACTLY ^ comments only in this format and provide no additional text outside this format,

Please classify the following ^ comments:

Nearly identical prompts were employed for anti-Muslim hate speech, LGB hate speech and trans hate speech. Any message that was classified with a "1" by the LLM was regarded as hateful for that target group.

In terms of performance, the four generative LLM-based classifiers typically classify around 75 to 80% of messages correctly. In cases where messages are incorrectly classified, it is typically the case that the message is judged as borderline hateful.