ISD | Powering solutions
to extremism
and polarisation

Tangled Web

The interconnected online landscape of hate speech, extremism, terrorism and harmful conspiracy movements in the UK

About this report

This report provides a snapshot of the online landscape of terrorism, extremism and hate speech related to the United Kingdom, and is the result of digital analysis conducted by the Institute for Strategic Dialogue (ISD) for Ofcom. It illustrates trends relating to a range of actors promoting terrorism, extremism, hate speech and harmful conspiracy theories across several relevant social media platforms with broadly comparable data access. These are Facebook, YouTube, Twitter, Instagram, Reddit, Telegram and 4chan. ISD researchers employed both qualitative and quantitative methodologies to produce this research, which is underpinned by a unique data analysis system called Beam, which uses machine learning and natural language processing for innovative cross-platform social media research.

Authors

Milo Comerford, Jacob Davey, Jakob Guhl and Carl Miller

Acknowledgements

Analysis support from Francesca Arcostanzo, Will Baldet, Julia Ebner, Lea Gerster, Lukasz Janulewicz, Richard Kuchta, David Leenstra, Zoe Manzi, Julia Smirnova and Francesca Visser.

This research was commissioned by Ofcom.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

Copyright © Institute for Strategic Dialogue (2023). Institute for Strategic Dialogue (ISD) is a company limited by guarantee, registered office address PO Box 75769, London, SW1P 9ER. ISD is registered in England with company registration number 06581421 and registered charity number 1141069. All Rights Reserved.

www.isdglobal.org

Contents

Executive Summary

The increasing importance of social media platforms has shaped society in profound ways. Platforms have provided users with the ability to freely express themselves, build communities and engage with a broad range of viewpoints. At the same time, the landscape of online harms is in flux, with social media being abused in an ever more sophisticated fashion to foment hatred, promulgate conspiracy theories and incite real world violence. While research by Ofcom shows that most users do not regularly encounter content that is hateful, violent or that contains misinformation, these threats nonetheless need to be taken seriously for their potential for real world harm.[1]

In the UK, social media has played an increasing role in violent extremist mobilisation across the ideological spectrum.[2] Meanwhile, in the wake of the Covid-19 pandemic, social media platforms have seen an upsurge in hate targeting vulnerable communities as well as abuse and harassment against public figures, including health workers, journalists and elected officials.[3]

Such harms are driven by a continuum of organised extremist actors and looser networks associated with anti-minority hatred and the promotion of conspiracy theories.[i] Individuals in these movements are defined less by their membership of proscribed terrorist organisations than by the online subcultures they inhabit and often fluidly move between. Disinformation, conspiracy theories, hate speech, harassment and violent extremism often interlink in ways that are extremely difficult to separate and isolate.

Understanding the blurring of these lines will be an important factor in designing effective regulatory and policy responses to illegal and harmful activity since these rely on clear legal definitions. This moment of flux in the online landscape is coinciding with a moment when policy makers in the United Kingdom, Europe and around the world are actively grappling with the challenge of addressing the psychological and physical harms associated with social media activity. As regulatory frameworks are developed, there is a pressing need for improved evidence around the nature of illegal and harmful social media activity across platforms, and the actors and groups responsible for it.

This research was commissioned by Ofcom to provide evidence around the cross-platform manifestations of this broad landscape of online harms, namely the interrelated issues of terrorist, extremist and hate speech content. It should not be taken as a reflection of how Ofcom will be approaching regulation, but instead a broad snapshot of the current state of play of these threats in the UK. It is worth highlighting at the outset that researchers and regulators alike face major challenges when seeking to understand these issues at scale. Platforms impose limits on access to data and do so in ways that are sufficiently different per platform to make it hard to compare the prevalence of harmful content across different platforms. Given these limitations, it is highly challenging for independent researchers or regulators to provide a definitive assessment of the scale of such online harms.

Approach

Focusing on seven social media platforms – Facebook, YouTube, Twitter, Reddit, Instagram, Telegram and 4chan – this report analyses a range of interconnected online communities engaging in potentially illegal and harmful activity targeting UK audiences, specifically those promoting terrorism, extremism, hate speech and harmful conspiracy theories. This research illustrates broad trends across a subset of accounts associated with these phenomena – it does not purport to be a comprehensive mapping of all UK accounts and channels promoting hate and extremism across these platforms.

These platforms have been included due to the availability (albeit of very different scales and natures) of data through public application programming interfaces (APIs), previous research identifying them as relevant venues where extremists have sought to mobilise,[4] as well as their prominence within the UK. All these services

[i] Defined in full on page 6, ISD defines extremism as the use of violence, politics or societal change to further a supremacist ideology, which frames the survival of an identity-based 'in-group' in terms of the destruction of an 'out-group'. In this report we focus specifically on extremist content associated with incitement, violent threats or harassment, which directs hate against a protected group, or perpetuates harmful disinformation. Conspiracy theories explain events in terms of a small group of powerful persons acting in secret for their own benefit against the common good. This report focuses on conspiracy movements associated with real-world harm, including incitement of violent threats and harassment, or hate directed against a protected group.

are among the top 10 platforms for UK users,[5] except for the imageboard 4chan, which was included because of the important role of the platform's /pol/ board within hateful far-right extremist subcultures.[6]

Subject matter experts at ISD identified (through both manual and semi-automated discovery) 768 accounts, groups and channels which met our working definitions of terrorism, extremism, hate speech or harmful conspiracy theories - outlined in detail below - and which were deemed to be UK relevant. These included accounts associated with a known harmful group or actor in the UK , a social media account engaging in harmful activity expressly focussed on the UK, or a harmful online community where there is substantive evidence of engagement by UK individuals.

Ranging from far-right white supremacists to Islamist extremists, antisemitic conspiracy theorists to anti-Muslim Hindutva groups, these accounts and channels were coded according to their support for terrorism, extremist activity, hateful targeting of a protected group, or spreading of conspiracies associated with real world harm.

Gathering data from these accounts between 1 October 2021 and 31 March 2022, we collected just over 2.5 million messages. While not all the content shared by these accounts was overtly illegal or harmful, researchers used a range of innovative methodological approaches to establish a multi-platform snapshot of broader account behaviour, as well as a narrower focus on hateful content.

In the first chapter of this report, we provide a high-level platform-by-platform quantitative overview of the landscape of UK-relevant accounts that ISD have identified as engaging in online behaviour associated with extremism, terrorism, hate speech or harmful conspiracies. In the second chapter, we produce a network map of this interconnected landscape of online actors, using natural language processing to analyse the overall messaging of these accounts. In the final chapter we deploy an 'ensemble' of hate speech classifiers to understand the specific hateful narratives these communities seek to advance.

To guide the analysis outlined in this report, ISD has built on established understandings within the academic and policy domains to develop working definitions (outlined below the key findings) of key concepts – including 'terrorism', 'extremism', 'hate speech', and 'harmful conspiracies' – aimed at relating online activity to concrete harms. These definitions were in part informed by priority offences specified in the Online Safety Bill such as terrorism, hate crime, harassment, threats and incitements to violence.[ii] However, they go beyond this, factoring in corresponding terms and conditions of some services and ISD's subject matter expertise around these evolving threats.

Key Findings

- **Accounts associated with hate speech and extremist content are much more easily discoverable than ones associated with terrorism on the platforms studied for this report.**
    - Of the 768 UK-relevant accounts and channels identified in our study, only 55 (18 on Instagram, 13 on YouTube, 10 on Facebook, 8 on Telegram and 4 on Twitter) met the project's definition of terrorism, suggesting that such activity may be taking place in more opaque areas of the internet. The majority of these accounts (42) were supportive of proscribed groups linked to Northern Ireland related terrorism.

- **Most accounts in scope of the study were found to be defined less by an association with specific terrorist, extremist or hate groups, than the broader hateful and conspiratorial online environments they inhabit.**
    - Our analysis shows significant overlap between a broad spectrum of harmful conspiracists and overtly white nationalist communities online. Disinformation, conspiracy theories, targeted hate, harassment and extremism often interlink online in ways that are extremely difficult to separate and isolate. An innovative cross-platform mapping of messages from these accounts using natural language processing approaches detected nine interlinked linguistic 'communities', characterised by a shared focus on topics such as Covid-19 conspiracies, anti-immigration narratives and opposition to the LGBTQ+ community.

[ii] At the time of writing, these are set out in Schedules 5, 6 and 7 of the Bill. https://bills.parliament.uk/bills/3137

- **Large social media platforms host accounts coded as hateful and extremist which can attract hundreds of thousands, or even millions, of UK users.**
  - Accounts identified as associated with Islamist extremists have a large following on Facebook, with the four largest accounts averaging over 568,000 followers.
  - We identified accounts associated with the UK extreme right in particular on Telegram (with some of the largest far-right channels having over 150,000 subscribers) and YouTube (with the biggest channel having close to 2 million subscribers). This finding demonstrates that major social media platforms remain venues for hate and extremist content, despite growing evidence from researchers that this sort of activity is increasingly manifesting on fringe platforms.[7]

- **More links to YouTube were shared by accounts in our study than to any other platform.**
  - In total, accounts in our study associated with promoting hate speech, extremism, terrorism or harmful conspiracy content shared links to YouTube over 50,000 times, accounting for 78% of links to external platforms identified in this study. It was not in the scope of this specific study to explore whether or not the content of these links was harmful.
  - Extremist-associated accounts on various platforms such as 4chan, Twitter, Instagram and Facebook also regularly directed their followers to Telegram, hinting towards the importance of the platform within these communities.

- **Notwithstanding their presence on larger platforms, our data offers indications that UK-relevant actors associated with hate speech and extremism are potentially interested in smaller sites.**
  - Emerging platforms such as Bitchute, Odysee, Gettr and Rumble were each linked to within our data more often than Facebook, Instagram or Reddit (though less often than YouTube and Telegram). This indicates that such platforms might potentially be of interest for accounts that spread content associated with terrorism, extremism and hate.

- **Content from accounts in the study reached significant audiences and garnered high levels of engagement across platforms.**
  - Posts from UK users on 4chan's hateful /pol/ board garnered 1,891,328 comments. Accounts identified in our study generated 526,398 replies on Twitter, 462,009 Facebook comments, 321,830 YouTube comments, 179,140 comments on Instagram, and 4,864 on Reddit during the period of study.
  - Where this can be measured, Telegram channels received 95,388,986 cumulative views, whilst videos from YouTube accounts associated with hate speech, extremism and terrorism were viewed 37,429,616 times. Posts from these accounts received 6,520,902 likes on Twitter, 3,874,941 on Instagram and 1,569,893 on Facebook during the period of research.

- **A great deal of content posted by hateful and extremist actors wasn't considered explicitly hateful by a bespoke ensemble of hate speech models.**
  - Our innovative research approach created an 'ensemble' of algorithms to identify hate speech, including 24 open source, commercial and bespoke models, and 25 bespoke lexicons.
  - This approach (which sets a high bar for inclusion, outlined below) found 2,260 messages meeting our definition of hate speech, and 5,371 messages constituting offensive speech.
  - 47% of the hate speech gathered from accounts in this study targeted individuals based on national origin, followed by anti-Muslim hate speech (24%), antisemitism (15%) and anti-black hate speech (7%).
  - Notably, explicit hate speech represents a very small proportion - 0.35% - of overall messages sent by accounts included in our study. Challenges and limitations of such algorithmic approaches are outlined in the section below.

Methodology

The aim of this project was to provide Ofcom with insights on the cross-platform online ecosystem of accounts, channels and digital spaces explicitly spreading terrorist, extremist, hateful, or harmful conspiracist content (defined below), as well as being related to the UK. Our approach has brought together different research methods, ranging from qualitative analysis to conventional data science to machine learning-driven natural language processing. Here we lay out the key research steps:

Step 1: Developing working definitions

The research process began by establishing working definitions for the terms 'terrorism', 'extremism', 'hate speech' and 'harmful conspiracies', to guide data collection. This exercise was rooted in a review of relevant priority offences in the Online Safety Bill, platform terms of service, and consideration of academic definitional frameworks. This exercise was intended to ensure relevant insights could be produced for Ofcom, although these ISD working definitions are not intended to correspond with Ofcom's own forthcoming remit under the Online Safety Bill. Thus, the content and actors in this report may not fall within Ofcom's regulatory scope, and we make no assessment of whether they likely will or should do so.

These high-level working definitions – whose rationale is outlined in further detail in Annex A – are as follows:

Terrorism – The research draws on the UK Terrorism Act (2000)'s definition, which describes terrorism as the use or threat of action, designed to influence any international government organisation or to intimidate the public, for the purpose of advancing a political, religious, racial or ideological cause. This was interpreted as encompassing support for groups or organisations proscribed under the Terrorism Act, including Islamist terrorist groups such as ISIS, far right terrorist groups like National Action, and proscribed groups linked to Northern Ireland-related terrorism such as the Irish Republican Army and Ulster Volunteer Force. It also encompasses the broader online behaviours by which terrorists and their supporters build community, disseminate content and communicate online for terrorist purposes, in line with the 2020 Interim Codes of Practice published by the Home Office.[8] While most of the accounts identified expressed support for proscribed groups, a small number of accounts were judged to fall under wider behaviours outlined in the UK Terrorism Act's definition of terrorism.

Extremism – ISD uses a social identity definition of extremism, which describes the use of violence, politics or societal change to further a supremacist ideology, which frames the survival of an identity-based 'in-group' in terms of the destruction of an 'out-group'. Extremism is therefore distinct from terrorism in describing a supremacist ideological framework, rather than a specific set of illegal activities (including by proscribed groups). In light of the harms within scope of the Online Safety Bill, this broader definition was narrowed to focus on specific illegal and harmful behaviours in scope of platform policies and potential regulation at the time of writing this report. Our research therefore focuses specifically on extremist content associated with incitement, violent threats or harassment, which directs hate against a protected group, or which perpetuates harmful disinformation (understood as false, misleading or manipulated content presented as fact, intended to deceive or harm).

Specific manifestations of extremism referenced in this report include (but are not limited to):

Far right extremism: A form of nationalism that is characterised by its reference to racial, ethnic or cultural supremacy. Right-wing extremism is the advocacy for a system of belief in inequality based on an alleged difference between racial/ethnic/cultural groups. Extremism expert Cas Mudde characterises the far right as commonly exhibiting these features: nationalism, racism, xenophobia, anti-democracy and strong state advocacy.[9]

Islamist extremism: The advocacy of a system of belief that promotes the creation of an exclusionary and totalitarian Islamic state, within which those who do not subscribe to this vision are portrayed as an inferior 'out-group' and are subjected to implicit, explicit or violent means of subjugation and prejudice. This supremacist ideological goal might be pursued through violent action, political activism or systematic societal change.

Harmful conspiracies – Conspiracy theories explain events in terms of a small group of powerful persons acting in secret for their own benefit and against the common good. An increasingly prominent subset of harmful conspiracy movements such as QAnon have been linked to violent radicalisation and are prompting responses from platforms, such as Meta's policy on violence-inducing conspiracy networks.[iii] For the purposes of this report, we focus on conspiracy movements associated with real-world harm, including the incitement of violent threats and harassment, or hate directed against a protected group. We have not used a broader conception of harm that might include, for example, potential threats to public health from Covid-19 conspiracy theorists, unless these actors were found to incite violence, make threats, engage in harassment or direct hate against a protected group.

Hate speech – In this report hate speech is defined as activity that seeks to dehumanise, demonise, express contempt or disgust for, exclude, harass, threaten, or incite violence against an individual or community based on a protected characteristic. We have defined protected characteristics as race, national origin, disability, religious affiliation, sexual orientation, sex, and gender identity. The rationale for this working definition - based on a review of relevant legal frameworks and various platform terms of service - is outlined in detail in the accompanying hate speech-focused paper.

Greater detail on the process for establishing these working definitions can be found in Annex A.

Step 2. Account discovery and appraisal

ISD analysts first triaged against our working definitions existing seed lists of UK-relevant accounts previously identified by ISD analysts as relevant to terrorism, extremism, hate and harmful conspiracies to develop a 'high certainty' group of social media channels, accounts and groups. This would serve as a basis for further account discovery, based on a 'snowballing' method supplemented by qualitative approaches, including exploring other accounts shared and recommended within these spaces, as well as keyword searches based on a systematic review of relevant resources on the landscape of terrorism, extremism, hate and harmful conspiracies in the UK. This initial process yielded 597 accounts/channels for collection.

To ensure a consistent standard of evidencing across diverse platform contexts, to qualify for inclusion, an account, channel or group needed to have posted at least three (but most often considerably more) pieces of content that clearly met the working definitions above, and have a clear connection to UK audiences, either through self-identification by the account holder or the nature of the content posted. Each identified account was qualitatively appraised by two expert researchers, with disagreements resolved between coders, and evidence systematically collected by analysts to substantiate coding decisions.

Analysts next undertook a process of semi-automated 'account discovery' to identify new candidate accounts for manual appraisal. This approach differed per platform, with researchers identifying Twitter accounts through analysis of the followership of UK-relevant accounts identified as promoting terrorist, extremist, hateful or harmful conspiracist content, while - across other platforms - using links spread by this account set to identify relevant channels, spaces or groups. This process yielded an additional 171 candidate accounts, of which 155 were deemed relevant to our categories (130 Twitter, 15 YouTube, 1 Facebook, 2 Instagram, 7 Telegram), resulting in 768 actors in total.

All qualifying accounts were manually coded to determine if they had expressed 1. any overt hate speech directed at a protected group; 2. any specific extremist ideological conviction (for example far-right or Islamist extremism); 3. support for terrorism; or 4. support for harmful conspiracy theories, to establish the high-level 'account types' analysed comparatively throughout the report (namely those accounts and channels identified as sharing terrorist-related content, far right extremist related content, Islamist extremist related content, hate speech content, or harmful conspiracy theorist content).

[iii] 'Pages, Communities, Events and Profiles or other Facebook entities that are, or claim to be – maintained by, or on behalf of, militarised social movements and violence-inducing conspiracy networks are prohibited. Admins of these pages, communities and events will also be removed.' https://transparency.fb.com/en-gb/policies/ad-standards/unacceptable-content/militarized-social-movements/

Step 3. Data collection and per-platform analysis

Researchers used the respective platform-specific application programming interface (API) to gather posts from public pages and groups (Facebook), posts from pages (Instagram), posts and comments on subreddits (Reddit), tweets and follower information from accounts (Twitter), video titles, descriptions and comments from channels (YouTube), and posts and comments in channels and groups (Telegram). Where available, data on engagement metrics and links was also gathered.

On 4chan's /pol/ board, all comments from users with a UK flag were downloaded. It should be noted that this flag can be altered through the use of a virtual private network (VPN), or the manual selection of a number of non-country-specific flags (such as a swastika) and thus it is not possible to conclusively determine all of these comments were posted by actual British users, or users living in the UK.

All publicly available behaviour (including non-harmful behaviour) from these accounts between October 2021 and March 2022 was back-collected in a one-off historic data collection in May 2022 and subject to a mixed methods analytical approach. This included:

- Analysis of volumes of messages sent over time to understand overall behaviour;
- Qualitative and ethnographic appraisal of messages to establish harmful narratives;
- Quantitative analysis of reach and engagement behaviour to understand audiences;
- Consideration of visual/non-textual content from relevant channels;
- Link sharing analysis to understand how activity on the mainstream social media platforms analysed here might be relating to an array of emergent platforms;

Step 4. Multi-platform semantic clustering

The next stage in the methodology deployed a novel methodology to compare accounts across platforms in a unitary way, based on their language use. This used natural language processing approaches to map the location of accounts in our study on the basis of their language, allowing for comparison of behaviour across platforms going beyond a focus on platform-specific behaviours.

This approach contrasts with more traditional network mappings often produced using social media data that are based on friend-follower relationships and other forms of engagement behaviour (such as the use of hashtags).

Data Collection

All messages from nominated accounts were collected across 2021 Q4 and 2022 Q1. Excluding messaging from YouTube and 4chan,[10] a total of 317,932 messages were obtained from 422 different actors. After removing reposts and accounts only involved in reposting, 201,366 original messages from 417 different actors remained.

Platform	Messages (Exl. Reposts)	Actors (Exl. Reposts)
Twitter	143,736	210
Facebook	24,759	59
Telegram	23,402	73
Instagram	8,220	68
Reddit	1,249	7

Table 1. Total number of messages and actors used for the network analysis

Account Representation

This method used an established technique that measures how semantically similar any given messages are. This involves mapping messages into a common (vector) space in which the similarity of texts can be compared numerically. To do this, we used what is known as a pre-trained sentence encoder. A pre-trained sentence encoder is an example of a pre-trained language model that is specifically optimised for measuring similarity between sentences. In the field of Natural Language Processing, pre-trained language models are currently viewed as the most effective way of capturing certain aspects of language meaning, and act as building blocks that can be adapted to suit a broad range of language processing tasks.

We encoded each of the 201,366 messages using a sentence encoder called all-mpnet-base-v1.[11] This encoding process places each message into a 768-dimensional space, where a pair of messages with encodings (vectors) that are closely located in this space are taken to have similar meaning. To compute the account-level representation for a given account, we aggregated (averaged) across the message-level representations (vectors) associated with that account in the dataset. This was performed for each of the 417 accounts.

Account Network Construction

Since it is not possible to visualise this 768-dimensional landscape directly, we next constructed a network in which the nodes represent the 422 accounts, and (un-directed) weighted edges represent the similarity between two accounts. UMAP is an efficient algorithm for constructing such a network, which aims to preserve the structure (or account positionings) in the original space such that neighbouring accounts in the original space will be neighbouring accounts in the network. We used UMAP (with the cosine-similarity metric) to compute the connectivity between account representations, which resulted in 4,802 edges (links between nodes).

We graphed the network using Gephi. For positioning nodes, we applied the ForceAtlas2 layout algorithm, a widely used approach to spatialise a weighted undirected network in two dimensions. For community detection, we applied a modularity-based algorithm, which assigns a single community to each node (the Louvain community detection algorithm).

Community Characterisation

The next step in the process was to characterise each of the communities: how were they similar to other communities, and what set them apart? To do this, we combined three different forms of cluster-specific inquiry:

- Manual appraisal of accounts randomly sampled from each cluster, with an attempt to draw out themes in their behaviour and identity;
- The identification of keywords and phrases that most distinguished the cluster from any other;
- Cluster-specific statistics around activity, followership and so on.

Manual appraisal included a study of the profile picture of the account/channel (especially the presence of motifs, tropes, regional or ideological identifiers); the profile description of the account (interests, hobbies, political or ideological attachments) and its reach. Expert analysts read 100 of the most recent messages sent by each account to create an impressionistic narrative summary of the content sent by the account. This included wherever possible any thematic, regional, temporal or ideological features of the messages themselves. Once this exercise was completed for the sampled accounts for each cluster, analysts sought to identify the key attributes that were held most in common by accounts grouped within each cluster.

Keywords and phrases were extracted from the messaging of each cluster using TF-IDF, a method that aims to define how important a word or phrase (term) is to a passage of text (document). TF-IDF scores each term based on its frequency within a document, and rescales this score based on the term's frequency across all documents in the dataset. In essence, terms that occur in a message that are generally rare in messaging overall are scored highly and extracted as keywords. We use messaging from all clusters to compute background term statistics, and then for each cluster we concatenate a random sample of 1,000 messages to form the cluster's document, from which we extract the 100 highest scoring keywords. Potential limitations to this experimental approach are outlined in the following section.

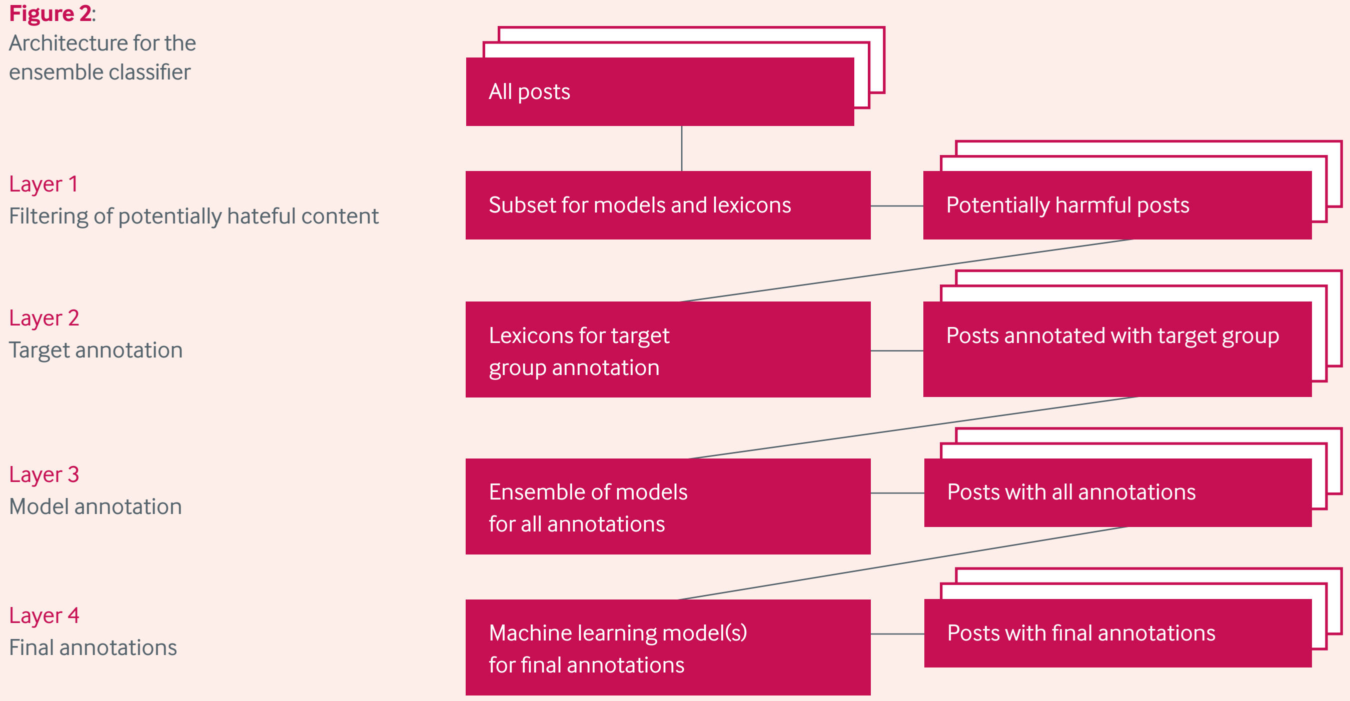
Step 5. Ensemble classification of hate speech from harmful accounts

Overview

One of the key research aims of this project was to achieve a way of classifying any social media post from any of the in-scope platforms as hate speech or not (as defined above). This methodology must be automated in order to cope with any scale of data required, and to conform to the definitions of hate and targets of hate outline above. It also must be explainable and understandable to any end-user, and its performance must be robustly measured and clearly communicated.

To build on the great deal of work that has already been done to classify hateful (and related) speech, the team pursued an 'ensemble' classification approach. This involves the combination of a number of different models together to collectively decide whether any given post is hateful or not. The principle underlying this approach is that any model will have inherent strengths and biases depending on the training data and classification architecture, and combining them

Figure 2:
Architecture for the ensemble classifier


All posts
Layer 1
Filtering of potentially hateful content
Subset for models and lexicons
Potentially harmful posts
Layer 2
Target annotation
Lexicons for target group annotation
Posts annotated with target group
Layer 3
Model annotation
Ensemble of models for all annotations
Posts with all annotations
Layer 4
Final annotations
Machine learning model(s) for final annotations
Posts with final annotations

allows us to radically enlarge both the training datasets which can be leveraged and the underlying approaches to machine learning that can be tested. The other goal of this methodology is for it to be expandable: new models can be added as they become available, and new sources of data can be added as they become relevant.

The Ensemble Architecture

Practically, each social media document (post, tweet, comment etc.) is passed through four layers of annotation. As we detail below, the first is a broad filtering to remove content unlikely to be hateful. The second identifies any possible target of the hate, the third passes each document through every model included in the ensemble, and the fourth is our bespoke 'model of models', where the best patterns of annotation are identified that correlate with actually hateful messages.

Layer 1. Filtering potentially hateful content. We limit the total amount of data being processed for reasons of computational burden and delay. To do this, a subset of the models was used to filter all the collected data for those likely to contain hate speech, before being passed on to the full ensemble of models.[iv]

Layer 2. Target annotation. The second layer annotated each document for the likely protected characteristic targeted by the hate speech, where this exists. This covered speech targeting individuals on the basis of: race; religion; disability; gender; sexuality, gender identity or national origin.

Layer 3. Model annotation. All posts that passed the Layer 1 filtering were run through all available models and the annotations saved.

Layer 4. Final annotation. Each post was run through a machine learning algorithm (XGBoost) that was fine-tuned on 6,496 messages collected across all the platforms and randomly sampled on the collection. This

[iv] These models needed to be both fast (to handle the large volumes of data) and together have a high recall. To evaluate based on this criteria a large random sample of our dataset, 10,019 posts from these actors, were manually inspected and combinations of filters were used to produce the best result. The final combination removed 21% of the dataset, while only losing 3% of the potentially hateful posts.

algorithm took into account all the annotations provided in layer 3 for each post and ultimately determined - on the basis of these annotations - whether any document is likely to contain hate speech. For this dataset the model was determined to have an accuracy of 70%.[v]

The full list of models and lexicons included in this ensemble are included in Annex B.

Caveats and limitations

Attempts to compare platforms can be fraught with difficulties, ranging from differences in data affordances, collectable behaviours, actor discoverability and historic data availability. These issues, alongside a series of caveats and research limitations are laid out in detail below.

Difficulty of Platform Comparisons

The kind of research conducted for this report can naturally lead towards platform comparisons, where the overall scales of certain forms of behaviour, especially posting illegal and harmful content, are viewed side-by-side. Given different data affordances quantitative inter-platform comparisons are not possible, and therefore this report should be read as a platform by platform summary.

First, the amounts of measurable behaviour on each platform may reflect differences in what data each platform allows researchers to acquire, rather than the underlying phenomena. This, itself, is a multi-level issue that is determined by:

Collectable spaces vs. non-collectable spaces.

Perhaps the most important factor is how much of the overall platform is collectible by the researcher in the first place.

- At the time of writing in early 2023, on **Twitter** virtually all visible activity was collectible directly using the official developer tools/API provided by Twitter, with the result that Twitter was the most comprehensively covered platform in our study and the volume of hateful content identified on Twitter appeared significantly greater than other platforms.[12]
- On **Facebook** and **Instagram**, the Meta-owned tool, CrowdTangle, provides access to parts of the platforms.[13] On Facebook, CrowdTangle reports to index data from all public pages with at least 25,000 page likes or followers, all public groups with at least 95,000 members, all US-based public groups with at least 2,000 members, and all verified profiles. Similarly on Instagram, CrowdTangle reports to index data from all public Instagram accounts with at least 50,000 followers and all verified accounts. On both Facebook and Instagram access is provided only to top-level posts, and thus other data, such as comments, is not accessible. Accordingly, it is certain that our approach to hate speech analysis on Meta owned platforms was significantly limited, and that messages which are publicly viewable on the platform (such as comments on posts, and posts from less popular public pages) are not covered in our study.[14]
- On **Telegram**, the official developer tools/API allows for messages to be collected from all groups accessible to the researcher.[15] However, while messages within specified channels are collectable, searching for message content is not possible through the API and thus finding channels that participate/produce relevant content is a challenge.
- On **Reddit**, the official API provides access to all top-level posts and comments that are accessible by the researcher, given the identifiers are already known.[16] Searching for top-level posts and comments is possible through the API; however, there are limits on how many results are returned per query which, while the Reddit API is not explicit about, is often reported in related documentation to be at most 1,000 items for a given query.[17]
- On **YouTube**, the official API provides access to all videos and comments that are visible on the platform. However, the content of comments cannot be directly searched, and thus the researcher must first provide the video to which comments are required before collecting and analysing comments.[18]
- On **4chan**, the official API provides access to all posts, both top-level and replies, currently visible on the platform. While this allows the full platform at the current state in time to be explored, the number of visible posts in each board is limited, and no

[v] This accuracy score is derived from an average of a 68% recall rate (the proportion of hateful content the model detected) and a 73% precision rate (how often these annotations were correct).

functionality exists to obtain posts that surpass this limit.

Collectible vs. non-collectible behaviours
Another crucial factor is exactly what kind of behaviour is collectible. On Twitter, both the Tweet and Retweets are collectible, as are comments on any Tweet. By contrast, on Facebook and Instagram, the CrowdTangle tool makes only top-level posts available, and comments are excluded. On Reddit, both top-level posts and comments are collectable. On Telegram, messages and replies made within a given channel are collectable. On 4Chan, both the top-level posts and comments can be collected. On YouTube, channels, videos, and video comments can be collected.

Historic data availability
Some platforms allow researchers to collect data much further back in time than others. Twitter allows full access to public historic Tweets (using Twitter Academic API or paid-for API) but a limited 7-day search window for those without privileged/paid access. Through CrowdTangle, Facebook and Instagram also allows for the collection of historic posts. Reddit affords limited data collection in a way that is not very transparent, such that a limit of 1,000 items can be returned for a given search, this means that high-volume searches will often be very close in time to the point of collection, and lower-volume searches will go back further in time. Telegram provides access to the complete message history for a given channel. 4Chan data is largely ephemeral, with a limited archive per board capped up to the past 3,000 threads posted within the past 3 days. For popular boards such as /pol/ this often results in data being available for less than 24 hours before it is inaccessible. YouTube provides access to the full history.

Discoverability
Each platform (via its APIs) allows researchers to discover data in different ways, and some are much more encompassing than others. Some platforms, such as Twitter, Facebook and Instagram (via CrowdTangle), YouTube and Reddit allow researchers to find data that contains specific words or phrases. In the case of YouTube, this is available for the discovery of videos and does not relate to searching comments. This can return very broad collections of data back to researchers, where lots of messages contain certain phrases with different meanings and in different contexts. Others, however, such as Telegram and 4Chan only allow data to be returned that are sent within a certain channel or board, and thus channels must be first identified and collected to explore message content.

Account/Channel/Space Discovery

The second factor - especially important for this report - is how these affordances combine to make it easier or more difficult to discover new accounts. As the methodology explains, this report was based on the identification of extremist and terrorist accounts, channels or spaces. These are discovered by researchers in different ways that are specific to each platform, and some platforms are easier to discover accounts on than others. On Twitter for instance, accounts were discovered by inspecting significant followership overlap with already identified extremist accounts, whereas this follower-network functionality is not available on Instagram, Facebook, Reddit, and Telegram, thus relationships from already identified extremist accounts could only be discovered by inspecting the links that they had posted.

Applying definitions of hate speech to social media data

It was challenging to consistently apply our definition of hate speech to the social media data we collected. We observed many posts to fall within a 'grey' area where different coders could take them as hateful, offensive, or indeed neither. This causes an issue when making a binary classification of hateful or not, as both training and evaluation data can represent a high degree of analyst bias. Our response was to blind code data, measure inter-annotator agreement, and work through edge-cases as a team to develop our shared understandings of hate speech through practical examples.

A large amount of hate was expressed in terms of derogatory slurs targeting people on the basis of their protected characteristics. In most of these cases, the hateful nature of posts was evident. At the same time, across most categories, – and on most platforms except for 4chan – we also encountered posts of a more ambiguous character. In these edge case posts, hate sometimes took more subtle forms, or messages didn't cross our threshold for hate. For example, in posts about migration, distinctions between critique on immigration policies on the one hand and anti-migrant hate can be ambiguous. Dehumanising posts referring to migrants were classified as hateful, for example, but when posts advocated for "migration stops" or "keeping illegal immigrants out of the country" they did not meet our definition of hate. Similarly, posts about Islam sometimes demonstrated the blurry boundary between anti-Muslim hate and atheistic critique of religion. In the anti-Jewish category, posts that referred to conspiracy theories with antisemitic features proved not always to be hateful, however, when blame was attributed to Jewish people, or if it was alleged they were responsible for secretly conspiring, content was more likely to be classified as hateful. Because hate based on sex, gender identity, sexual orientation and disability primarily manifested through the use of slur terms, edge cases were less of an issue here.

Different norms and meanings of language across platforms

While the groups targeted by hate were similar across the social media platforms analysed, the language and terminology used to do so varied from platform to platform. Most significantly, it was observed from the data collected that 4chan's user community has a distinct and characteristic vocabulary, that includes the wide-spread use of derogatory slurs to refer to one another in a way which could be interpreted as hate-speech by a reader coming from a targeted community, but not necessarily interpreted as hateful by the recipient of the message. A similar type of posting language was found on Reddit, although it should be noted that the amount of such language was significantly lower.

Hateful language on Facebook, Instagram and Twitter also looked different. Posts on these platforms were, overall, less aggressive in nature. Compared to 4chan and Reddit, hate was less overt on other platforms; however derogatory slurs were still the primary means through which hate was expressed.

For some protected categories on Facebook, however, the plausibly hateful content identified involved fewer overt slurs. Anti-Muslim hate was often framed as criticism of Islam as a religion and hate against Jews involved anti-Jewish conspiracy theories more than specific slurs or attacks.

Comprehensiveness of discovery

The results presented in this paper are not representative of the entirety of the platforms they relate to. The identification of in-scope accounts is necessarily a qualitative and manual undertaking rather than a systematic one, impacted by a number of effects. These include discoverability of data, the knowledge and biases of researchers, or the nature of platforms where account discovery is taking place.

Semantic similarity mapping

As with any methodology, the specific approach developed for the semantic mapping contained in the report carries with it a series of strengths and weaknesses. When interpreting the data, the following caveats should be regarded:

- **Cluster descriptions are impressionistic**, and characterised by expert appraisal of account activity. Other analysts may have drawn distinct conclusions or emphasis.
- **Cluster descriptions don't capture every account that's a member**. Characterisation of clusters inevitably involves generalisations and each cluster will contain 'noise' (meaningless information).
- **Accounts-based collections will miss relevant activity**. One obvious limitation is that this research was confined to pre-selected accounts and that will mean that other relevant behaviour may be missed. This is offset, to some extent, by the keyword-based collections detailed in the second report in this series.

Ethics and Privacy Considerations

Given the public interest and sensitive nature of discussions around hate, extremism and terrorism on online platforms, it was essential that the research conducted for this report met the highest research ethics standards. This report is additionally based on the use of technologies and analytical methods which may be unfamiliar to many readers, both in how they work and the nature of what they produce. It is therefore crucial to identify and explain ethical challenges as transparently as possible.

This ethical framework attempts to balance two different public goods: that of privacy and autonomy online; and that of public security and safety. These may be threatened by hate, extremism (associated with incitement, violent threats or harassment) and terrorism, as well as their digital manifestations. The aim of the framework was to help shape a project that could provide the clearest, most accurate picture possible for Ofcom about the nature of hate speech, extremism and terrorism online as it relates to the United Kingdom, whilst doing so in a way that would not constitute in any way an intrusion on or risk to any individual. Here we outline the key elements and principles of the ethical framework used for this research.

Focus on Public Data

Issues of privacy online are complex. In some cases, online spaces might be said to be clearly public, such as Twitter's timeline, or clearly private, such as direct messages on Facebook. In some cases, the privacy of some spaces may be more ambiguous, as with open groups on Facebook or very large fora where membership is required.[vi]

In addition to this, public perceptions of what social media spaces are public and which are private can vary significantly, as it may be confusing for users to read and understand platform rules around privacy.[19] Because these discrepancies between reality and perception relate to issues of autonomy, where information that is not public or information that might reasonably be perceived as private is sought within a research project, the acquisition and recording of this data must be well considered, justified and documented.

In many cases, study of such 'private' spaces is also technically impossible, because data from them are not made available by the platforms themselves. But regardless of any technical possibility, no spaces were studied where we thought it likely users would have a reasonable expectation of privacy of any kind, e.g. online spaces that were password protected, that require special membership, that require personal authentication via video-calls or ideological questionnaires; or non-public parts of social media platforms (such as private messaging).

Anonymity of Research Subjects

The research was exclusively focussed on the understanding of broad, strategic trends and patterns over time and across platforms, and this meant that no individual was named in any research output, and no individual-level behaviour was described, unless the individual in question could be assumed to be highly visible publicly already.

The project complied with all relevant UK data regulation and GDPR requirements. To preserve privacy, all outputs from the project are presented at an aggregate level, with no row-level data, usernames or other identifying data related to individuals shared outside the project team. Furthermore, research took place on the basis of anonymity, whereby the anonymity of all research subjects will be guaranteed through our research methodology (including the use of permanent de-identification where possible, the maintenance of a separate and secured coded name register where this is required by the research, and the limitation of access to identifiable data).

In some cases, quotations were used in order to illustrate a particular point, however these were bowdlerised to prevent the retroactive identification of the original post through, for example, an online search. Similarly, account names or other information that could lead to the identification of individuals were blurred.

Criminal Behaviour

This work was undertaken as a piece of research to identify broad trends and patterns. It was not intended, nor designed, to guide or inform any law enforcement investigation or organisation. The research did not reveal the identity of individual involved, except for the caveats outlined above. No part of the research architecture (and especially the automated ensemble classifier) was designed to identify behaviour that passed any kind of criminal or legal threshold. The research was not trying to find or measure criminality.

A Clear Referral Process

While the project was not designed or directed to identifying criminality, it was important that researchers clearly understood what to do if they encountered behaviour online that implied the presence of a credible real and immediate threat to a loss of life, threat to cause serious harm or threat of injury to another. This includes serious sexual assault or rape, specifically targeted towards individuals, groups, events or places. ISD researchers adhered to an institutional referral process whereby researchers would report to relevant authorities

[vi] ISD has outlined some of its own considerations about the difficult distinctions between public and private spaces online in its 2019 submitted response to the UK government's Online Harms White Paper. https://www.isdglobal.org/wp-content/uploads/2019/12/Online-Harms-White-Paper-ISD-Consultation-Response.pdf

any documents encountered during this research, which might be identified as representing a real and present threat as described above.

The Ethics of Inaction

As stated above, the ethical framework was predicated on balancing privacy and social cohesion and public safety. As such, researchers needed to also give consideration to the argument that in some cases a failure to conduct online research might itself be ethically unsound, particularly where such research might inform activities to improve social cohesion and public safety, and reduce hate crime and violence.

As such, the principles above were used to shape work that, we believe, minimises the individual harms that research can impose whilst maximising the capacity of the research to contribute to public goods. We regarded this to be the most ethical course of action to take.

Safeguarding

Safeguarding must be a core consideration of work in this field, where the particular sensitivity of hateful, extremist or violent subject matter and the potential harms that researchers face create a special onus to ensure wellbeing. Potential harms of such research include exposure to upsetting or even potentially traumatic content, degraded mental health, or in extreme cases risks to personal safety. ISD gives constant consideration to potential impacts on researcher welfare, and implements appropriate mitigating actions, including time limits on exposure to harmful content, mental health support or further training.

All efforts were made to minimise direct exposure to extreme or harmful content to only instances when such engagement is absolutely necessary. Managers needed to produce a rationale justifying the necessity of exposure to extreme content against the project's objectives. Strict limits were imposed on the amount of time researchers spent engaged with such material. All researchers are provided access to counselling sessions with an external expert specialised in PTSD and workplace stress, which are promoted to team members, with researchers encouraged to make use of them on a monthly basis if not more regularly as needed.

Part 1: Understanding the Cross-Platform Landscape of Online Terrorism, Extremism, Hate Speech and Harmful Conspiracy Theories in the UK

Harmful activity online is spread across a wide ecosystem of platforms. Movements spreading hateful, extremist and violent content tailor their formats, narratives and approaches to the platforms they use. Understanding where such harmful content manifests online, and the scale and nature of the challenge on different platforms thus remains crucial.

This section therefore focuses on mapping the spaces in which UK-relevant accounts sharing hateful, extremist and terrorist and harmful conspiracy content operate, and provides deep-dives into the dynamics on each of the platforms studied for this report. We outline both cross-platform trends during the period of study, as well as the nature of online terrorist, extremist and hate speech activity identified within platforms through dedicated snapshots of these phenomena on Facebook, Instagram, Reddit, 4chan, Telegram, Twitter and YouTube. These platforms were selected based on previous analysis by ISD and other experts which has identified them as important venues for harmful activity (4chan and Telegram),[20] their large user bases in the UK (Facebook, Instagram, Reddit, Twitter and YouTube),[21] and our ability to gather data from them through platform APIs.

Summary Overview

The following section provides a cross-platform overview of the digital ecosystem relevant to online extremist, terrorist and hateful content in the UK, and presents snapshots of activity from Facebook, Twitter, Instagram, YouTube, Reddit, 4chan and Telegram. These overviews are based on data from a series of accounts, groups and channels identified during a thorough account discovery process led by expert analysts at ISD, which sought to establish a balanced sample by using a 'snowballing' methodology to triage and build out further candidates for inclusion based on UK terrorist, extremist and hate speech-linked hateful accounts identified via previous ISD research. Researchers also conducted keyword-based searches around key terms established through literature review of UK-relevant online threat actors, as well as semi-automated network expansion (a process described in further detail in the methodology).

In total we gathered data from 768 accounts, channels, groups and pages which met our definitional criteria (outlined above), 499 of which were generally active between 1 October 2021 and 31 March 2022. This includes 215 Twitter accounts, 77 YouTube channels, 73 Telegram channels and groups, 68 Instagram accounts, 59 public Facebook pages and groups, and 7 subreddits (a specific online community on Reddit). We also collated UK-relevant comments from 4chan's /pol/ board (these were not distinguished from posts), an anonymous political discussion imageboard – short for 'Politically Incorrect' – on 4chan.

Comparisons between platforms are not possible or meaningful in any straightforward way: one YouTube video with a duration of one hour is not equivalent in content to tweets with a maximum of 280 characters. Furthermore, diverging data access across platforms further complicates comparisons. Recognising these vital caveats around data availability, our data suggests that during the data collection period the terrorist, extremist and hate speech-linked accounts identified for this study generated over 520,000 comments on Twitter, over 460,000 comments on Facebook groups and pages, over 290,000 comments on YouTube channels, and over 170,000 comments on Instagram, and over 4,000 comments on Reddit. As table 3 shows, between 1 October 2021 and 31 March 2022, our data showed that identified terrorist, extremist and hate speech-linked accounts published over 259,000 posts on Twitter, over 25,000 posts on Facebook, over 23,000 posts on Telegram, over 8,000 posts on Instagram, over 2,000 posts on YouTube and over 1,000 posts on Reddit. Our approach did not disaggregate posts and comments on 4chan. Posts on 4chan's /pol/ board represented more than the total posts and comments across all other platforms included in the study combined..

To assess the reach that the most influential extremist, terrorist and hateful accounts achieve on these platforms, we compared the ten accounts in our dataset with the highest number of followers or subscribers on each platform. Our findings indicate that Facebook and YouTube are the platforms with the highest average following of these accounts, with Twitter, Instagram, Telegram and Reddit considerably lower.

Across most of the sub-categories used for analyst coding (outlined in the methodology), Twitter and

Platform	4chan	Facebook	Instagram	Reddit	Telegram	Twitter	YouTube
Accounts, Pages and Channels	NA	59	68	7	73	215	77
Posts	NA	25,613	8,220	1,259	23,402	259,448	2,313
Comments	1,891,328	462,009	179,140	4,864	-	526,398	297,338
Top-10 accounts average following	NA	285,986	55,040	2,757	54,654	84,470	394,860

Table 3. Total number of accounts, posts and comments analysed for this report, as well as the average reach of the ten accounts sharing terrorist, extremist, hate speech or harmful conspiracy content identified on each platform with the largest number of subscribers. While presented in this way for accessibility, such numbers are difficult to compare directly, with a YouTube video not being equivalent in content to a tweet.

Platform	Far-right extremist	Islamist extremist	Hateful	Harmful Conspiracy Theorist	Terrorist
Twitter	126	29	200	36	4
Telegram	111	15	123	30	8
YouTube	98	14	123	2	13
Instagram	48	13	50	15	17
Facebook	46	12	53	4	10
Reddit	6	0	10	3	0
Total	435	83	559	90	52

Table 4. Total number of accounts identified per platform and sub-group (including accounts that were inactive during October 2021 – March 2022). Note that accounts can be coded as multiple categories.

Telegram were generally the two platforms where the most accounts associated with terrorism, extremism and hate speech were discovered. As explained in the caveat above, this may be a reflection of data availability rather than overall prevalence. Instagram, YouTube and Telegram were the platforms on which ISD researchers identified most entities supportive of terrorism or groups proscribed by the UK government,[22] although these were notably few and far between – with a skew towards Northern Ireland-related terrorism (outlined in further detail in case study below on UK-related terrorist content).

Moving beyond platform comparisons, the following sections go into greater detail on platform-specific dynamics among the terrorist, extremist and hate speech-associated accounts in our dataset, and the key events that led to spikes in activity during the period of study.[23]

Case Study

UK-related Terrorist Content on Social Media[vii]

Across platforms, researchers identified 52 UK-relevant entities expressing support for proscribed terrorist groups or advocating for the use of terrorist tactics (17 on Instagram, 13 on YouTube, 10 on Facebook, 8 on Telegram and 4 on Twitter). In this case study we provide a qualitative overview of this landscape.

Northern Ireland-Related Content

The majority of these (39) were supportive of groups linked to Northern Ireland related terrorism, especially on larger platforms such as Instagram and Facebook, with support for the proscribed Irish Republican Army (IRA) and Ulster Volunteer Force (UVF) dominating. Analysts found the majority of terrorist content was propaganda relating to proscribed Republican groups, such as the Irish Republican Army. These accounts are primarily used to organise marches, protests and raise money, and while these activities can be legitimate, they were promoted by accounts which were also supportive of proscribed groups. Very few actual threats of violence are evident, although efforts to intimidate rivals, for example Republicans and Loyalists posturing against each other, or calls to doxx (publish an individual's personally identifiable details that could be used for harmful purposes) alleged members of MI5 and the Police Service of Northern Ireland are common.

During the period of data collection, content from the Republican entities that expressed support for proscribed terrorist groups was often focused on commemorating deaths of Republicans. This genre of commemoration can imply sympathy for terror groups such as the Irish Republican Army, Irish National Liberation Army and Provisional Irish Republican Army, with deceased members of these groups being eulogised for their sacrifice. Separately, while some accounts were found to be sympathetic to proscribed Loyalist groups that promote terrorism, qualitative assessment indicated that content primarily featured benign posts expressing support for the Union, Protestantism, marching bands and sports teams.

Islamist Content

10 UK-relevant entities (6 of them on Telegram) expressed support for Islamist terrorist groups such as the Taliban, al-Muhajiroun, al-Qaeda and Hamas. The latter three are on the UK list of proscribed terrorist groups, and while the Taliban is not proscribed in the UK it remains on the list of groups designated by the United Nations. Its seizure of power in Afghanistan in August 2021 led to a wide range of reactions among UK Islamist extremists which are often difficult to categorise within binary schemes of support vs. opposition. Instead, UK Islamist extremists frequently expressed joy over the "liberation" of Afghanistan, advocated for the Taliban being allowed to implement Sharia law in line with cultural norms, or glowingly reported on the alleged popularity of the group on the ground.

On Telegram, there remain legacy channels of al-Muhajiroun, one of the UK's most notorious Islamist extremist organisations, though these are not particularly active. ISD researchers also identified UK content supportive of al-Qaeda, including celebrations of the 9/11 terrorist attacks. Hamas, which was proscribed as a terrorist group by the UK government in 2021 (previously only the supposed "military wing" had been banned) which operates an English-language digital presence online, has also found some support among Islamist extremists in the UK.

Far-Right Content

Three UK-relevant entities on Telegram expressed support for proscribed extreme right wing terrorist groups National Action and Feuerkrieg Division, which were proscribed by the UK Government in 2016 and 2020 respectively.

[vii] Ibid.; United Nations Security Council Consolidated List, accessed at: https://www.un.org/securitycouncil/content/un-sc-consolidated-listeulogised

Platform snapshots

Outlined overleaf are seven summaries providing overviews of the platforms identified for this study: 4chan, Facebook, Instagram, Reddit, Telegram, Twitter and YouTube. These platforms were selected based on previous analysis which has identified them as relevant venues for harmful activity, their large user bases in the UK (with the exception of 4chan, whose /pol/ board was included because of its outsized role in far-right online mobilisation), and our ability to gather data from them through platform APIs.

For the purpose of this analysis, researchers have applied the coding framework set out in the methodology above, with findings reflecting data gathered from accounts, channels and groups - and the content posted by such accounts - identified as associated with terrorism, extremism, hate speech or harmful conspiracy theories. However, for readability in this section these are referred to through shorthand such as 'extremist actors' or 'hateful accounts'.

Platform
4chan

The following key trends were identified on 4chan's /pol/ board:

- The volume of posts by UK users on 4chan's /pol/ board spiked around the Russian invasion of Ukraine and the conclusion of the Kyle Rittenhouse trial;
- UK /pol/ users shared antisemitic and racist conspiracy theories and justifications for support of both Russian and Ukrainian forces;
- UK /pol/ users were generally supportive of Kyle Rittenhouse and devoted to glorifying his actions, with some interpreting the court decision around his actions as legitimising political violence.

4chan is an imageboard founded in 2003 where anyone can anonymously post comments and share images. One of its most active boards is /pol/ (politically incorrect), which is well-known for its meme culture and subversive humour and which has become a key hub for the extreme right wing, conspiracy theories and hateful content targeting women and minority communities.[24]

As individuals post anonymously it is difficult to provide an ideological breakdown of /pol/ users, who occupy a spectrum of positions. However, the prominence of the board as a hub for global extreme right activism[25] is such that we consider the entire channel to be an extremist space, despite the presence of some counter-speech within the channel. In total, 1,891,328 posts by UK users were collected between 1 Octobear 2021 and 31 March 2022.

The largest spike of posts coincided with the Russian invasion of Ukraine. On the 24 February 2022, when Russia invaded Ukraine there were 23,079 posts by users posting from UK IP addresses (compared to 8,000-11,000 over the previous months). Over the coming weeks, British 4chan users came up with a variety of interpretations for the invasion, often linking it to new or existing conspiracy theories. This frequently included claims alleging that the invasion was a hoax or false flag operation, set up by globalist (often used as a coded term for Jewish[26]) elites. Antisemitic posts on /pol/ accused Jews of being responsible for controlling either one or both of Ukraine or Russia.

4Chan posts per day overall

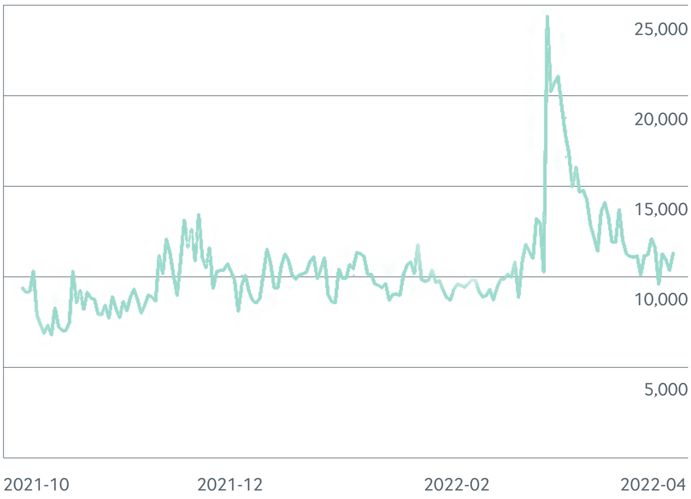


Figure 5: Volume of posts by UK users of /pol/

There was diverging support for different sides in the conflict. Regardless of which country users supported, the justifications they gave were often explicitly antisemitic or racist. Anti-Russian posts condemned Putin as a "Jewish dictator" who has turned Russia into "a Jewish-controlled multi-ethnic degenerate power which specialises in white genocide." Supporters of the Russian invasion on the other hand argued that Putin was standing up to the Jewish-controlled agenda of the West, showing the intersectional targets of hate within the data.

Another peak in activity occurred during the third week of November 2021, likely triggered by the conclusion of the Kyle Rittenhouse trial, peaking with 13,005 posts on 19 November when the teenager was acquitted of all charges. Rittenhouse had shot dead two people during the civil unrest in Kenosha, Wisconsin in August 2020 in the wake of the death of Jacob Blake – a black man shot by a police officer.[27] Comments by British 4chan users were generally supportive of Rittenhouse, often glorifying his actions, for example characterising him as an "honorary Viking". Others interpreted the ruling of the court as a mandate to shoot BLM activists.

Platform
Facebook

The following key trends were identified on Facebook:

- Facebook is one of the platforms where accounts associated with UK extremists and hate actors have the largest reach.
- Among extremist-linked accounts, far-right-associated pages and groups were the most active in terms of producing posts and received most comments, however accounts associated with Islamist extremists attracted more engagement per post on the platform.
- Islamist extremist-linked groups and pages have built up a large following on Facebook, with the top four averaging over 500,000 followers. Private profiles were also found to be used as public channels, but were ineligible for data collection via API.

Reach and Engagement

Our data shows that Facebook continues to be one of the platforms where UK extremists and hate actors have the biggest reach (see table 1), using public pages and groups (and at times private profiles, as outlined below) to communicate with their followers or engage in discussions with other like-minded users in public or private groups.

Among the UK relevant accounts identified for this report, most pages and groups were classified as hateful as well as far-right.

Facebook	Groups and pages
Far-right extremist	46
Islamist extremist	12
Hateful	53
Harmful conspiracy theorist	4
Terrorism	10

Table 6: Facebook groups and pages in study, categorised by type

Pages and groups coded as perpetuating hate speech produced both the most posts and received the most comments. Among extremist accounts, far-right pages and groups were the most active in terms of producing posts and received most comments. However, while Islamist extremists produced considerably fewer total reactions (e.g. 'likes') than far-right and hate speech-linked accounts, their average number of reactions was much greater.

Among the ten accounts in our dataset with the highest number of followers on the platform, four were coded as Islamist extremists (averaging over 568,000 followers) while five were classified as far right (averaging over 103,000 followers). This could be a result of UK-relevant accounts linked to Islamist extremism having a greater international following than far-right-linked accounts, which are more likely to focus on specifically British issues.

Harmful conspiracy theorists, Northern Ireland-related groups and accounts supporting terrorism each produced only about 10% (between 261 to 332 posts) of the Islamist pages' volume of posts and about 1% of that of far-right extremists.

Notably, ISD researchers also noted during the course of the study that UK-related Islamist extremist ideologues were utilising personal Facebook accounts with followings in the tens of thousands. These accounts acted in much the same way as the groups and pages of this study to engage wider audiences, showing the blurred lines between 'public and private'. However, such accounts were outside the scope of the study as they could not be analysed through systematic API access, which is limited to groups and pages on Facebook.

Facebook	Far-right extremist	Islamist extremist	Hateful	Harmful conspiracy theorist	Terrorism
Posts	19,972	2,672	23,989	332	261
Comments	289,527	69,235	449,176	2,716	715
Total Reactions	1,074,023	622,095	2,100,178	18,543	10,998
Avg. Reactions	54	233	88	56	42

Table 7: Reach and engagement of Facebook groups and pages, categorised by type

Platform
Instagram

The following key trends were identified on Instagram:

- Accounts in our dataset have a lower reach on Instagram than on other mainstream social media platforms. This is especially true for accounts which promote hate or harmful conspiracy theories, but did not meet our threshold of extremism.
- Hateful content on Instagram often targeted the LGBTQ+ community, migrants and Jews.
- Posts by terrorist, extremist and hate speech-linked accounts in our data set spiked around the Russian invasion of Ukraine.

In total, 8,220 posts by 68 UK hateful and extremist accounts were collected from the predominantly visual media platform Instagram, between 1 October 2021 and 31 March 2022.

The breakdown shows that posts by accounts classified as hateful, harmful conspiracy theorists and far-right extremists greatly outnumbered the posts by accounts supportive of proscribed terrorist organisations, linked to Northern Ireland or propagating Islamist extremism.

Reach and Engagement

While far-right extremist accounts represented the majority of those in our Instagram dataset, their reach is relatively small compared to other groups. None of the ten accounts with the highest number of followers were classified as far-right extremist. Instead, seven of those ten were classified as hateful, three as harmful conspiracy theorists and three as linked to Islamist extremism (categories may overlap). Similarly, hateful accounts, harmful conspiracy theorists and Islamist extremists had a much bigger average reach and received more likes and comments than the average far-right account. The same trends are visible in terms of comments and likes per post.

Instagram	Accounts	Posts
Far-right extremist	48	3,120
Islamist extremist	13	322
Hateful	50	4,612
Harmful conspiracy theorist	15	3,387
Terrorism	17	478

Table 8: Instagram accounts and total posts during the period of study, categorised by type

Instagram posts per day overall

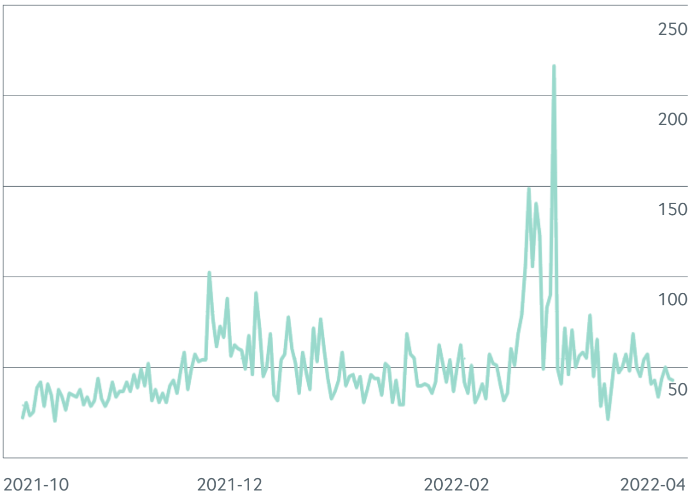


Figure 8: Volume of posts by UK actors linked to far-right extremism, hate speech or harmful conspiracy theories on Instagram

Only two accounts (both classified as hateful) had more than 100,000 followers, and only three more than 50,000. This suggests that accounts included in this study have a lower reach on Instagram than on other mainstream social media platforms.

Hateful but non-extremist accounts seem to be enjoying much greater success on Instagram than extremists. Hateful content found on Instagram was directed at a range of groups, but primarily targeted the LGBTQ+ community, migrants and Jews. Some UK-linked

Instagram	Far-right extremist	Islamist extremist	Hateful	Harmful conspiracy theorist	Terrorism
Avg. followers	510	4,480	12,571	6,057	470
Avg. likes	38	691	210	206	43
Avg. comments	3	6	34	7	2

Table 9: Reach and engagement of Instagram accounts in study, categorised by type

accounts spreading hate speech have a large following, for example, two accounts found to be posting hateful content targeting Jews and the LGBTQ+ community have a combined reach of around 336,000.

Posts by UK relevant accounts spiked multiple times during the period of study, especially around the lead-up to and immediate aftermath of the Russian invasion of Ukraine, when far-right extremist accounts posted memes about the coverage of mainstream media outlets and how the war could supposedly lead to the end of Western civilisation.

Platform
Reddit

The following key trends were identified on Reddit:

- Our research does not suggest Reddit is a central platform for UK extremism.
- Harmful conspiracy theorists on Reddit propagate were found to promote historically revisionist comparisons between Nazi policies towards the Jews and contemporary vaccination programs, a form of holocaust distortion. Hateful content on Reddit primarily targeted immigrants, Muslims and LGBTQ+ communities.
- Posts by the accounts identified for this study on Reddit rose significantly in parallel with the Russian invasion of Ukraine, but this appears to be due to increased activity in a hateful subreddit unrelated to discussions of the conflict.

While some research has suggested that the sub-forum-based platform Reddit has been influential in shaping far-right extremist and misogynist subcultures such as the alt-right and the involuntary celibate (incel) movements,[28] our research suggests it plays only a very small role for UK-related actors spreading terrorist, extremist or hate-speech content. Only 6 of the UK-relevant subreddits identified associated with these harm areas had more than 1,000 followers, with the highest being 8,300. Low follower counts and activity in these subreddits indicate the limited importance of Reddit for UK extremists. However, this observation needs to be caveated by the fact that our study narrowly focusses on channels expressly focussed on the UK, and does not contain analysis of broader more transnational spaces associated with hate and extremism, which might include UK users.

In total, we collected 1,265 posts from Reddit between 1 October 2021 and 31 March 2022 from UK-related subreddits. The posts had a total of 4,949 comments. Subreddits that are linked to harmful conspiracy theorists (I.e. those linked to harmful activity including hate speech and harassment) also propagate COVID-19 disinformation, narratives that warn about 'globalist' plots which utilise antisemitic tropes and the 'Great Reset' or draw historically revisionist comparisons between Nazi policies towards the Jews and contemporary vaccination programs.[29]

Instagram	Subreddits	Posts
Far-right extremist	6	16
Islamist extremist	0	0
Hateful	10	426
Harmful conspiracy theorist	3	823
Terrorism	0	0

Table 10: Subreddits and total posts during the period of study, categorised by type

Reddit posts per day overall

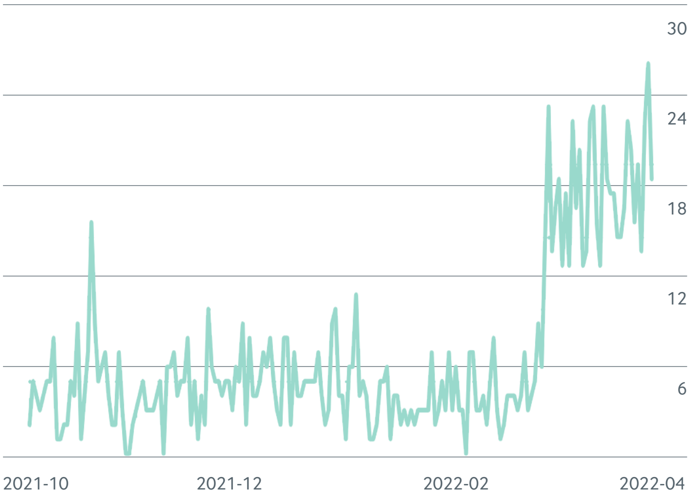


Figure 9: Volume of posts by UK actors linked to far-right extremism, hate speech or harmful conspiracy theories on Reddit

Posts in the subreddits classified as hateful resulted in 2,884 comments, compared to 1,870 comments in reaction to posts by harmful conspiracy theorists, indicating higher levels of activity and more dynamic discussions. Hateful content primarily targeted immigrants and Muslims, with some evidence of anti-LGBTQ+ sentiment.

Only 16 posts were gathered from UK-focused far-right extremist subreddits (receiving 195 comments). As mentioned above, our analysts did not identify any subreddits focussed on Islamist extremism or Northern Ireland related terrorism, or subreddits supportive of terrorist tactics or groups.

Posts on Reddit rose significantly in late February (from 4.4 per day until 27 February to 18.3 afterwards), after which activity levels remained stable. Even though this rise coincides with the Russian invasion of Ukraine, it appears to instead be caused by increased activity in a subreddit supportive of a far-right influencer that was included due to hateful posts identified by analysts.

Platform
Telegram

The following key trends were identified on Telegram:

- The UK far-right appears much more active on Telegram than adherents of other extremist ideologies.
- The UK far-right, hateful and harmful conspiracy channels reached significant audiences on Telegram, with some of the largest channels linked to far-right extremism garnering over 150,000 subscribers. In contrast, no UK Islamist extremist-linked channels on Telegram in our dataset have over 1,000 subscribers.
- Our data indicates Telegram provides a venue for international hateful actors to show solidarity and interact with their global counterparts.

Telegram is a (partially) encrypted messenger app with no terms of service prohibiting hateful content, and has emerged as one of the most popular messaging apps among extremist groups.[30] ISD and other experts have assessed this to be due to the platform's more narrow approach to content moderation compared to major social media platforms, which extends only to banning the incitement of violence on public channels and the sharing of illegal pornographic content on public channels and by bots, as well as prohibiting spam and scams.[31] Users in the UK must also be at least 16 years old to sign up.[32]

Telegram offers its users the option of creating and joining channels and/or groups. Both can have either a public or private setting, with private access granted via an invite link. Channels offer greater relative anonymity; subscribers are unable to view details of other participants and whilst a recent update has enabled comments on posts made by admins, dialogue is still mostly one-sided. Channels also offer an unlimited subscriber number, in contrast with groups which have a cap on membership. This may appeal to channel owners posting hateful content on Telegram as it allows them to reach the widest audience possible.

Instagram	Accounts/channels
Far-right extremist	111
Islamist extremist	15
Hateful	123
Harmful conspiracy theorist	30
Terrorism	8

Table 11: Telegram accounts and channels in study, categorised by type

For this report, we collected posts on UK-relevant Telegram groups and channels associated with terrorist, extremist, harmful conspiracy and hate speech content. Most fell into the far-right extremist category. Eleven channels were deemed pro-terrorist for either recruiting, expressing support for or being official accounts for groups proscribed as terror organisations by the UK government. Most of these entities were Islamist with 6 channels, followed by the far right with 4 and 1 pro-terrorist Northern Ireland-related account.

Reach and Engagement

Our research shows that UK far-right extremist content is easily discoverable on Telegram and garners a significant audience for their content, with some of the largest far-right extremist channels amassing over 150,000 subscribers. Islamist actors are much less prolific, with no UK-focused channel having over 1000 subscribers.

Far-right, hateful and harmful conspiracy channels reached significant audiences, with their view counts reaching tens of millions (notably these are cumulative numbers and do not indicate numbers of unique users). By contrast, Islamist extremists, Northern Ireland-related accounts and pro-terrorist channels are less active and reach much smaller audiences.

Telegram	Far-right extremist	Islamist extremist	Hateful	Harmful conspiracy theorist	Terrorism
Posts	12,453	3,065	14,072	5,642	235
Views	54,264,320	1,447,754	74,352,900	39,180,316	20,266

Table 12: Total Telegram posts and views during period of study, categorised by type

The most viewed post in this dataset is a video posted by a channel associated with one of the UK's most prominent far-right activists who promoted demonstrations against mandatory COVID-19 vaccinations for workers taken place outside of the UK. Viewed over 489K times, it gives an insight into how Telegram functions as a space for far right actors to show solidarity towards and collaborate with their global counterparts. While the content of the post itself is not hateful, it shows how the Covid-19 pandemic and the government's response was used by far-right figures to draw individuals towards their movements, including from non-extremist anti-vaxx communities.

On the same channel, similar videos with an average of 50,000 views detail protests in Germany, Spain, Austria, Stockholm, Canada and the US, showing that transnational actors in this ecosystem align their messaging to amplify wedge issues in an attempt to advance their political agenda. Notably views are calculated approximately on Telegram, and the totals listed in the table above represent cumulative totals, rather than unique viewers, with the counter including times the video was viewed elsewhere after being forwarded, with the context of these views unknown.

Platform
Twitter

The following key trends were identified on Twitter:

- Far-right extremism is more visible than any other category of terrorist, extremist, hateful or harmful conspiracy-related accounts on Twitter;
- Posts by UK Islamist extremists and posts by terrorism supporting accounts were the most likely to be shared;
- Based on the available data, and accounts identified, UK-related accounts engaging in hate speech on Twitter reach a larger total audience than other communities of harm on the platform, such as harmful conspiracy theorists, far-right extremist and Islamist extremist-linked accounts.

Through the project's account discovery methodology, 278 Twitter accounts were identified which met our threshold of geographic and ideological relevancy to hate, extremism and terrorism in the UK. This resulted in the collection of 259,448 posts from 215 accounts that were active between October 2021 and April 2022.

The biggest extremist user community identified were far-right extremists, though an even bigger group of non-extremist accounts were assessed by analysts as posting targeted hateful content towards an outgroup with a protected characteristic. We did not identify Twitter accounts supportive of Northern Ireland-related proscribed terrorist groups.

Twitter	Accounts
Far-right extremist	126
Islamist extremist	29
Hateful	200
Harmful conspiracy theorist	36
Terrorism	4

Table 13: Twitter accounts in study, categorised by type

Reach and Engagement

Over the period of study, most posts were collected from accounts responsible for hate speech towards a broad range of protected characteristics, followed by far-right extremist accounts, harmful conspiracy theorists and Islamist extremists (who received over 1.82 million retweets in total).

Examining the average number of retweets per post, our data suggests that Islamist extremist and terrorism-supporting accounts in our study were the most likely to be shared. The findings show that although hateful and far-right extremist accounts were the most prolific in the dataset analysed, they proportionally received the lowest number of retweets. On the other hand, harmful conspiracist and terrorism-supporting accounts were among the least prolific, but their posts obtained an average number of retweets higher than other groups.

Our data also shows that extremists and hate actors on Twitter obtained a total of 8.6 million likes during the period of study. While hateful accounts received the most likes, far-right and harmful conspiracist profiles were roughly equivalent in their levels of engagement.

Hateful actors saw the most replies to their posts, followed by far-right extremist accounts, harmful conspiracy theorists, Islamist extremists and terrorism supporting accounts. Given the scope of this study it is impossible to ascertain the breakdown of support for, or counter-speech to, hate, extremism and terrorism within these responses.

In terms of overall followers, hateful accounts had the largest number of followers, but also received the lowest number of retweets (together with far-right extremist accounts). Harmful conspiracy theorists, far-right extremist and Islamist extremist

Twitter	Far-right extremist	Islamist extremist	Hateful	Harmful conspiracy theorist	Terrorism
Posts	102,488	19,211	185,991	48,163	4,209
Likes	1,709,050	686,820	4,524,666	1,560,412	145,154
Retweets	441,919	182,138	1,379,131	345,407	39,150
Replies	185,152	26,749	375,008	121,847	6,549

Table 14: Posts, reach and engagement of Twitter accounts in study, categorised by type

accounts all have a comparable number of followers, while terrorist-supporting accounts have far fewer followers collectively.

Twitter	Total followers	Avg. followers
Far-right extremist	469,999	5,108
Islamist extremist	434,017	20,667
Hateful	1,408,282	9,085
Harmful conspiracy theorist	477,187	15,906
Terrorism	15,616	3,904

Table 15: Following of Twitter accounts in study, categorised by type

Despite posting less and being less numerous in our data set, the group with the highest average number of followers are Islamist extremists.

Platform
YouTube

The following key trends were identified on YouTube:

- The majority of UK-relevant YouTube channels spreading extremist or hate speech content were coded as far-right in character.
- These far-right extremist channels were responsible for almost half of the videos collected in our dataset of actors spreading terrorism, extremism and hate speech.
- Far-right extremists and hate actors in our dataset reach considerably larger audiences on YouTube than Islamist extremist channels or Northern Ireland-terrorism related channels.

The account discovery process identified 137 UK-related extremist or hateful channels and accounts on YouTube. The majority were associated with far-right extremism. 11 accounts were classified as linked to Islamist extremism, of which one account was additionally categorised as supporting a terror entity for its support of the Taliban (which while not proscribed in the UK remains sanctioned by the United Nations, and engages in terrorist activity[33]).

During the period of study between October 2021 and April 2022, 77 of these UK-related extremist and hateful channels were active, uploading 2,313 videos on YouTube between October 2021 and March 2022. As the most prevalent extremist actors on the platform, far-right channels were responsible for almost half of this output. Islamist extremist channels were responsible for uploading 637 videos, of which 54 were produced by the channel which directly supported the Taliban (however, not all the videos were related to the Taliban). Northern Ireland-terrorism related accounts were the least active of all categories.

Instagram	Accounts	Videos
Far-right extremist	98	1,004
Islamist extremist	14	637
Hateful	123	1,951
Harmful conspiracy theorist	2	16
Terrorism	14	52

Table 16: YouTube accounts and video uploads during the period of study, categorised by type

Reach and Engagement

Accounts coded as sharing far-right extremist and hateful content reach much larger audiences than any other category of YouTube channel in our data set.

Examining the ten channels in our dataset with the most subscribers, four channels were far-right extremist, and another three were right wing channels which did not meet our threshold of extremism, but nonetheless directed targeted hate towards religious and ethnic minorities, as well as hate on the basis of gender identity. Six of those channels have more than 100,000 followers and one of them has almost two million followers. Far-right and hateful actors who were previously suspended on some other popular social media platforms are able to upload their content on YouTube and in some cases to monetise it.

As part of assessing levels of engagement, researchers also collected and analysed comments associated with videos posted by hateful and extremist accounts. The videos uploaded by the far-right were much more likely to be commented on than Islamist extremist videos. Hateful accounts across all sub-categories received 318,959 comments across almost 2,000 videos, an average of 163 comments per video. As with our analysis of replies on other platforms, given the scope of this study it is impossible to ascertain the breakdown of support or counter-speech to hate, extremism and terrorism within these responses.

YouTube	Far-right extremist	Islamist extremist	Hateful	Harmful conspiracy theorist	Terrorism
Total subscribers	2,631,866	105,027	3,759,831	350	10,626
Total comments	234,623	5,011	318,959	32	1,430
Avg. comments	233	8	163	2	27.5

Table 17: Reach and engagement of YouTube accounts in study, categorised by type

Part 2: Network Dynamics - Understanding Relationships Between Harmful Online Communities

The platform analysis above was based on expert coding of UK-relevant social media entities according to pre-defined categories of harm. However, this section draws on an innovative methodology which subjects these manually-verified accounts to a semantic mapping as another way to understand their behaviour.

By mapping out the relationship between relevant accounts based solely on their language use, we aim to identify online communities that talk about similar things in similar ways, regardless of any prior assessment of the ideological makeup of the account. Going beyond more traditional network mapping ascertained through friend-follower relationships or other forms of engagement behaviour, this approach is unique as it allows for cross-platform comparisons, and can challenge existing knowledge and beliefs around the behavioural patterns of online communities sharing terrorist, extremist, hate speech and harmful conspiracy content. The machine learning approaches on which this process is based are 'unsupervised', and as such they can allow patterns to surface beyond existing hypotheses from researchers.

This approach notably does not seek to apply the working definitions of harm outlined above (i.e. assessments of extremism, terrorism, hate speech and harmful conspiracy content) in analysing content. Instead, the language model provides a broader harm-agnostic assessment of language patterns across these in-scope accounts, used to cluster accounts into different linguistic 'communities'. These algorithmically generated clusters are then qualitatively characterised by expert analysts based on their unique characteristics.

Approach

Data collection and analysis

This analysis is based on a subset of the data gathered from UK-relevant accounts included in this study between 1 October 2021 and 31 March 2022, excluding data from 4chan and YouTube due to data comparability challenges, leaving a dataset of 422 entities (see Methodological Annex for more details on our approach).

This method used a pre-trained language model (a 'sentence encoder') to measure how semantically similar any given messages are. Based on analysis of these linguistic patterns, we then aggregated these relationships at an account level across the entities in our dataset. This allowed for the construction of a network map, in which the nodes represent accounts, and weighted edges represent the similarity between any two accounts. An algorithm was then used to detect 'communities' across these nodes.

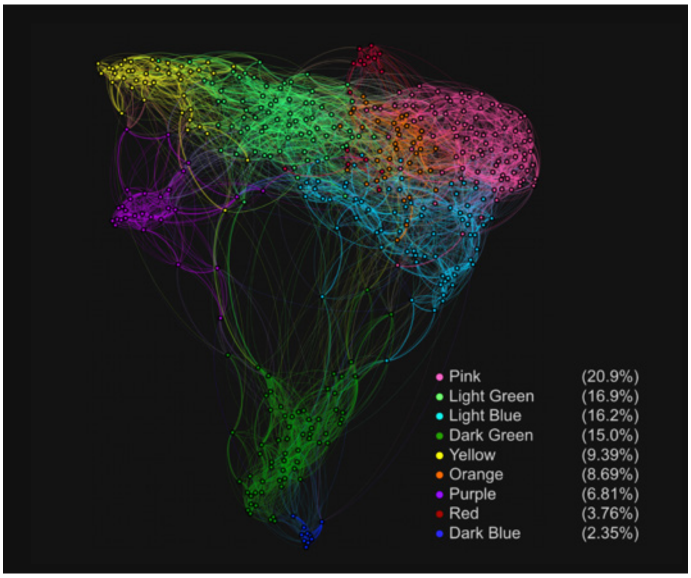


Figure 17. Language map showing the share of 9 different linguistic communities

The network comprises 9 different linguistic communities that can be visualised as an inverted pyramid. Towards the bottom of the visualisation in figure 12, two communities (dark green and dark blue) sit very distinct from any other.[viii] Towards the top, a single red community is also distinct, whereas five

[viii] N.b. The axes of the network are arbitrary, and are intended only to represent the relative semantic similarity between accounts.

Community	Number of accounts	Proportion of total	Messages (Oct–Mar)	Mean number of messages per account
'Anti-immigration nationalists'	89	21%	58,014	652
'Far-right influencers'	37	9%	16,099	435
'White nationalists'	69	16%	15,894	230
'Covid conspiracy theorists'	72	17%	42,356	588
'Extremist conspiracy hybrid'	29	7%	12,485	430
'Islamist extremist promoters'	64	15%	18,442	288
'Culture war networkers'	40	9%	29,238	731
'South Asian diaspore nationalist'	12	2%	1,581	132
'Anti-LGBTQ+ activists'	16	16%	4,583	286

Figure 19: Table of algorithmically generated language-based 'communities'

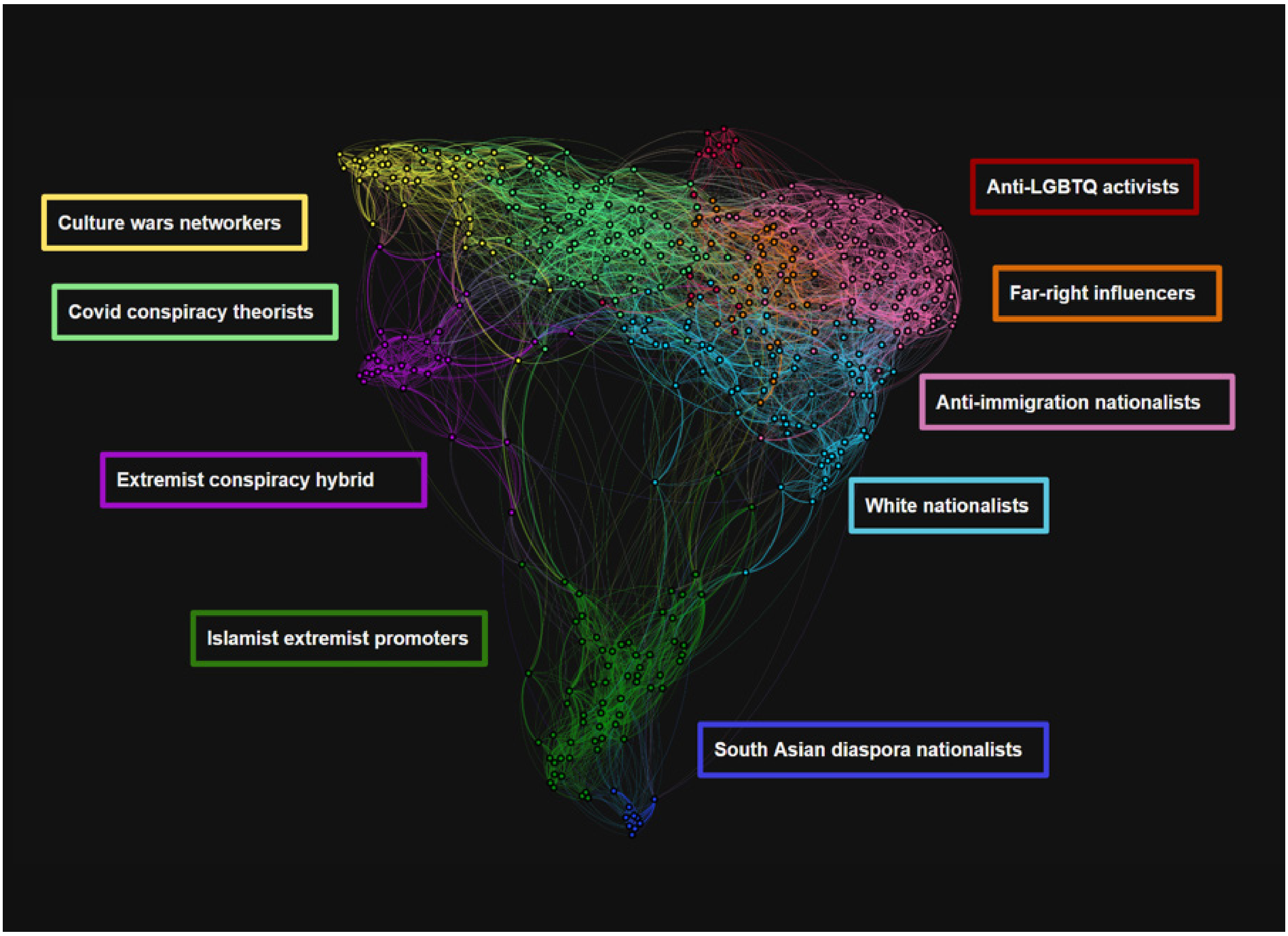

Culture wars networkers
Covid conspiracy theorists
Extremist conspiracy hybrid
Islamist extremist promoters
Anti-LGBTQ activists
Far-right influencers
Anti-immigration nationalists
White nationalists
South Asian diaspora nationalists

Figure 18. An overview of the different linguistic communities identified in this study communities run more contiguously throughout the top of the map, with a final group sitting off to the left edge.

Community characterisation

To illuminate this network, analysts characterised each of these 'community' clusters in terms of their similarities and differences. Analysts combined manual review of a sample of accounts from each cluster, drawing out notable themes in their messaging behaviour and self-identity, with algorithmically-surfaced 'surprising phrases' extracted from the messaging of each cluster to allow for comparative analysis of language use between communities. Once this exercise was completed for the sampled accounts in each cluster, analysts sought to identify the key attributes that were held most in common by accounts grouped within each cluster.

Overall, nine separate communities were identified, whose characteristics are outlined below in greater detail, alongside impressionistic short-hand labels. As noted in the caveats section at the end of this chapter, these characterisations are qualitative summaries drawn from subject matter expert analysts and so are inevitably animated by the judgments, perspectives and biases of the researchers, and other characterisations might have reasonably been reached. Crucially these summaries describe the general linguistic patterns drawing together these groupings, including narratives not necessarily related to terrorism, extremism or hate speech (despite the accounts themselves being coded as engaging in such behaviour):

- **Anti-immigration nationalists**: The largest cluster in the network, responsible for the largest number of messages (over 50,000 during the period of study across 89 accounts). The community also has the second highest mean number of messages per account of any cluster (over 650), with a higher proportion of Facebook accounts, but also a presence of Twitter, Telegram, Reddit and Telegram accounts. Some of the notable language cutting across this group of accounts include sub-cultures of disturbing humour, racist memes, and pervasive 'dog whistles' of white supremacist tropes. In this context, extreme-right wing ideologues with followings in the hundreds of thousands or even millions, conscious of content moderation guidelines, are careful to curate seemingly benign content with sly nods to white nationalism and supremacy.

- **Far-right conspiracy theorists**: This mid-sized cluster connects more overtly nationalist clusters based on the right of the map with communities characterised by more conspiracy theory-based language on the left side of the visualisation. With a proportionally high concentration of Telegram, Facebook, Instagram and Reddit accounts, language use is characterised by discussion of conspiracy theories such as the Great Replacement as well as Holocaust denial, alongside conversations around immigration suggesting a linguistic relationship with far-right extremist identity politics.

- **White nationalists**: One of the clusters on the predominantly nationalistic right side of the map, this larger cluster is predominantly characterised by the presence of more overtly white nationalist accounts. They are comparatively less vocal than accounts from other clusters and have the highest concentration of Telegram accounts of any across the mapping. Alongside white nationalist accounts, the language model has also included in this cluster several accounts expressing support for Northern Ireland-related proscribed terrorist groups, suggesting potential similarity in language across these groupings. Perhaps reflective of the ages of the individuals involved, Facebook remains their preferred platform (2015 research identified 68% of NIRT supporters were between 21-40 years old, with just 6% of supporters under 21[34]). Much of the content shared in this NIRT cluster constitutes propaganda related to proscribed Republican groups.

- **'Pure' Covid conspiracy**: Consisting largely of accounts from Twitter, this cluster is (with an average of over 500 messages per account over the period of study) the third most vocal of any across the network, and is focused on anti-vaccine conspiracy theories, discussions of government overreach and opposition to lockdowns. While much of this discourse is perfectly legitimate speech (although actors themselves have been coded as engaging in hate speech or expressly harmful conspiracy theories associated with incitement or harassment, as outlined in our project working definitions), it is notable that the language associated with this cluster chimes with wider trends analysed by ISD around the massive proliferation of conspiracy networks around Covid-19, with protest movements mobilising against restrictions commonly connecting anti-vaccine conspiracy theorists, anti-government actors, and extremist movements.[35]

- **Extremist conspiracy hybrid**: This is an ideologically diverse group of 29 predominantly Twitter accounts. They contain the broadest-ranging language of any cluster, and fewer obvious distinguishing linguistic features from the clusters around it. However its specific language use nonetheless prompted the language model to classify it as a unique 'community' distinct from its neighbours, albeit with fewer discernible distinguishing ideological features detectable by expert analysts.

- **Islamist extremist promoters**: The most consolidated cluster in the bottom half of the map contains predominantly Islamist extremist-aligned accounts, with a small number of anti-Islamist far-right leaning accounts (who notably use some similar language). The cluster comprises a broad set of platforms, including Twitter, Facebook, Instagram and Telegram. Islamist extremism – defined by ISD in terms of (both violent and non-violent) advocacy for a supremacist theocratic state which mandates the subjugation of other communities – manifests in diverse forms in the UK, from explicit support for proscribed terrorist organisations to activism targeting specific communities with hate, and this is reflected in network activity. The language characterising this cluster therefore ranges from core concepts used to provide an ideological justification for extremist action, to inspirational content intended to provoke hatred towards a particular group. The language also contains broader discussion of political themes, reflecting the fact accounts coded as sharing Islamist extremist content are also engaged in discussing mainstream topics. Previous ISD research has shown how extremists routinely court contemporary trends in their online activity to recruit allies that can legitimise them.[36]

- **Culture wars networkers**: The mid-sized cluster further to the left of the map are the most prolific of any across the network, comprising mostly Twitter accounts. Common themes include conspiracy theory-related and anti-migration language, and some support for the Russian invasion of Ukraine, but also cover a range of social topics including sports and entertainment. Unified more by their online behaviour than any coherent set of harmful narratives, we observed a high proportion of explicit network building activities, including 'follow-for-follow' activity and a very high proportion of amplification as opposed to the creation of original content. It is possible some parts of this cluster are inauthentic in their online behaviour, with some evidence of automation (the exploration of which was beyond the scope of this study).

- **South Asian diaspora nationalists**: At the bottom of the map are a dense, linguistically distinct set of Twitter, Facebook and Instagram accounts whose language use is demarcated from the wider set by a specific focus on Hindu and Sikh issues. They send the fewest messages per account of any in the network, although our analysis focused on English language content and so would not count non-English (messages). As outlined above, while accounts were manually coded for inclusion based on explicit terrorist, extremist or hate speech content, the language model used to form these communities is agnostic to specifically harmful or violent content in how it clusters accounts. While such content was not necessarily harmful, this cluster was characterised by frequent references among these accounts to the 1984 ejection by the Indian army of Sikh militants from the Golden Temple in Amritsar, and references to the movement's leader Jarnail Singh Bhindranwale, events which remain hugely contentious and contested within Sikh communities.[37]

  The language mapping also highlighted discussion around the topic of so-called 'Muslim grooming gangs' characterising this cluster of accounts. Whilst usually associated with far-right agitators, anti-grooming gang activism - which has often veered into anti-Muslim prejudice - has been embraced by some UK Sikh organisations who frame themselves as defending the honour of Sikh women.[38]

Other notable language features characterising this cluster include references to the Rashtriya Swayamsevak Sangh (RSS), a prominent Hindutva organisation which has its ideological roots in European far-right movements.[39] Importantly, a declared association with this group was not a basis for account inclusion in this study, but such references were identified as unique language features amongst a specific cluster of accounts which had been seperately coded as promoting either terrorist, extremist, hate speech or harmful conspiracy content. While posts in the cluster generally try to portray the RSS as a benign, non-sectarian volunteer movement that tries to help Indians independent of religion, RSS supporters have been widely accused of inciting violence against minority groups, especially Muslims.[40] However, the majority of anti-Muslim posts from high profile Hindutva groups avoid direct references to Islam or Muslims, instead using more coded language, a trend also identified in other online contexts.[41]

- **Anti-LGBTQ+ activists**: A small, linguistically distinct cluster of Twitter and Instagram accounts located at the top of the map, it comprises accounts focussed on sexual and gender identity, trans rights and the LGBTQ+ community.[42] This was notably clustered with language supportive of Donald Trump. Anti-LGBTQ+ sentiment represents a growing axis of online hate today. A recent report by anti-bullying charity Ditch the Label analysed 10 million online posts in the US and UK over a period of three and a half years and uncovered 1.5 million transphobic comments amid conversation around trans identities and communities.

While politics and race were the largest themes found within the transphobic comments, parenting and sports are twice as likely to be associated with transphobia in the UK, compared to US expressions.

Overall Network Patterns

Several patterns are clear from analysis of these communities. The most obvious trend is that the mapping suggests a differentiation between the language use of a cluster of more Muslim, Sikh and Hindu-focused accounts on the one hand, and a much more linguistically intermingled blend of white nationalist, anti-migrant and anti-LGBTQ+ hate and harmful conspiracy theory accounts on the other. Islamist extremist accounts are overwhelmingly distributed towards the bottom of the visualisation, indicating that there are several linguistical distinctions from any other kind of account or space in this dataset, although there is linguistic variety within this group too. Another grouping of Northern Ireland-related terrorism accounts is more closely associated with the main network cluster. Accounts coded as engaging in hate speech by expert analysts are distributed across a number of communities throughout the network.

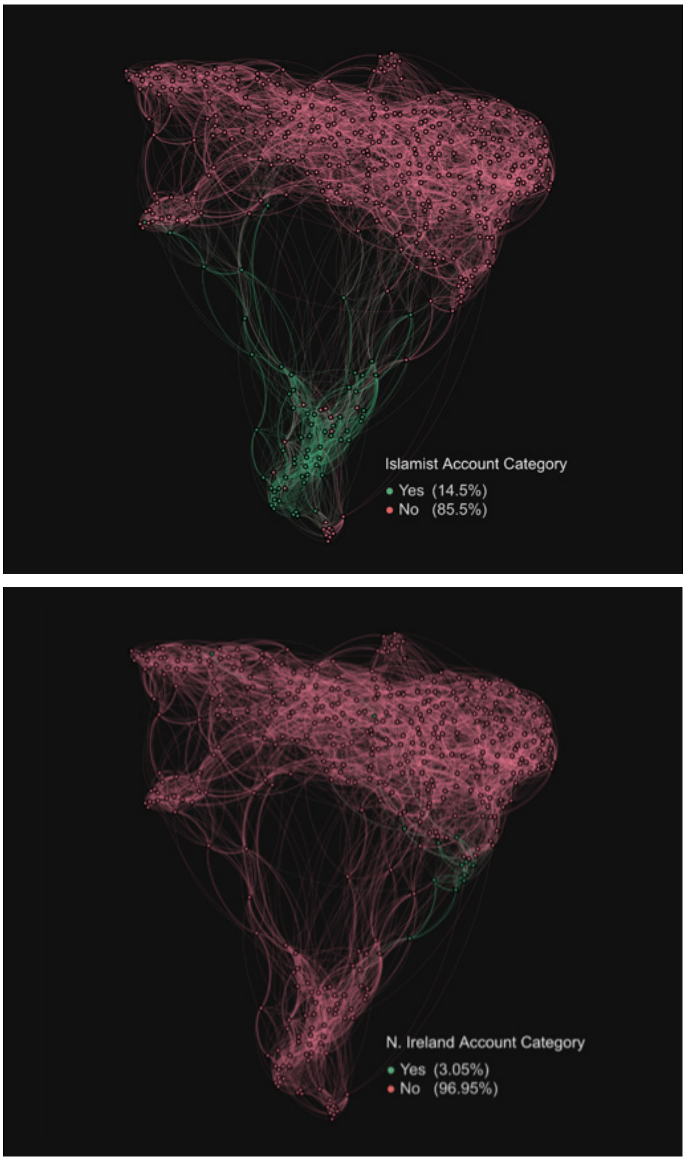


Figure 20: Network map showing the share of Islamist and Northern Ireland clusters

Whilst there is linguistic variety across the clusters at the top of the map, there is also continuity. The various forms of conspiracy theory, anti-lockdown and anti-vaccine activism, concern with immigration, fury at elites and politicians and discussions regarding race and immigration blend into a continuous spectrum reaching from anti-COVID conspiracy on one end to overt white supremacy on the other. This reflects ISD research on the proliferation of 'hybridised' online threats during the Covid-19 pandemic, in which protest movements commonly connected anti-vaccine conspiracy theorists, anti-government actors, and extremist movements, with the boundaries between disinformation, conspiracy theories, targeted hate, harassment and violent extremism becoming ever more blurred[43].

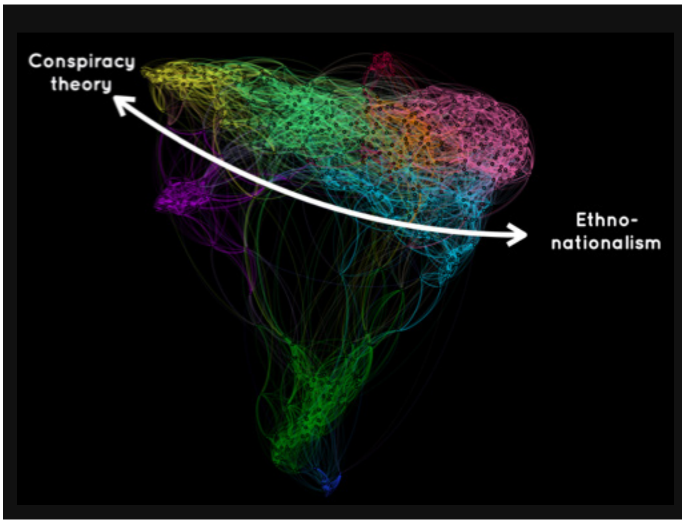


Figure 21: Network maps showing axis dividing actors focussed on conspiracy theories vs. (ethno)nationalism

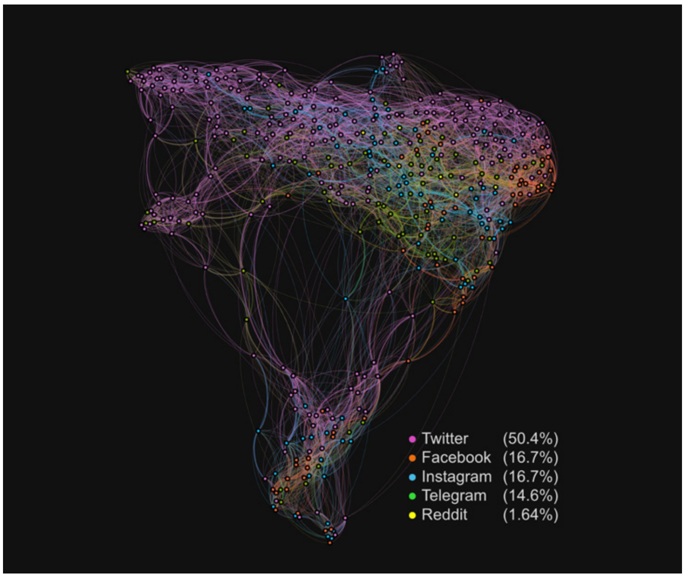


Figure 22: Network maps showing share of platforms

When making quantitative platform comparisons it is important to reemphasise that any analysis says as much about the relative availability of data as it does about the overall picture of the online threat. For example, the below visualisation might imply that that most terrorist, extremist and hate speech content could be found on Twitter, when it could instead reflect the fact that more data from Twitter is more represented in the dataset than from other platforms due to variances in data access.

However, with this caveat made, our analysis suggests a general trend of nationalist leaning accounts in our study being more likely to be on Facebook and Telegram, whilst conspiracy theorist communities tended to be composed of Twitter accounts. Islamist extremist and South Asian diaspora-related clusters are more spread across different platforms.

Dispersal of axes of hate across the network

We can also relate these linguistic groups to another measure: the types of specific hate speech associated with an account, based on the 'ensemble' of algorithms to classify hate speech, described in further detail in the chapter below and in the methodological annex of this report. This allows us to understand how explicitly hateful language maps onto the linguistic-ideological groupings that we have characterised.

When looking at racial hatred, we can observe some concentration in the anti-immigration and white nationalist communities on the right hand side of the map and the culture wars-focused networks on the left hand side. Hate targeted on the basis of religion is widely distributed, with a notable concentration in the Islamist extremist-linked cluster.

Gender-based hate shows a specific concentration within the anti-LGBTQ+ cluster at the top of the network map, whilst misogynistic hate was found to be much more widespread across the network.

Semantic Similarity Mapping: limitations and caveats

As with any methodology, the approach used here carries with it a series of strengths and weaknesses. When interpreting the data, the following caveats should be regarded:

- **Cluster descriptions are impressionistic**, characterised by expert appraisal of account activity. Other analysts may have had distinct conclusions or emphasis.
- **Cluster descriptions don't capture every account that's a member**. Characterisation of clusters inevitably involves generalisations and each cluster will contain 'noise'.
- **Accounts-based collections will miss relevant activity**. One obvious limitation is that this research confined to pre-selected accounts will mean that other relevant behaviour is missed. This is offset, to some extent, by the keyword-based collections detailed in the second report in this series.

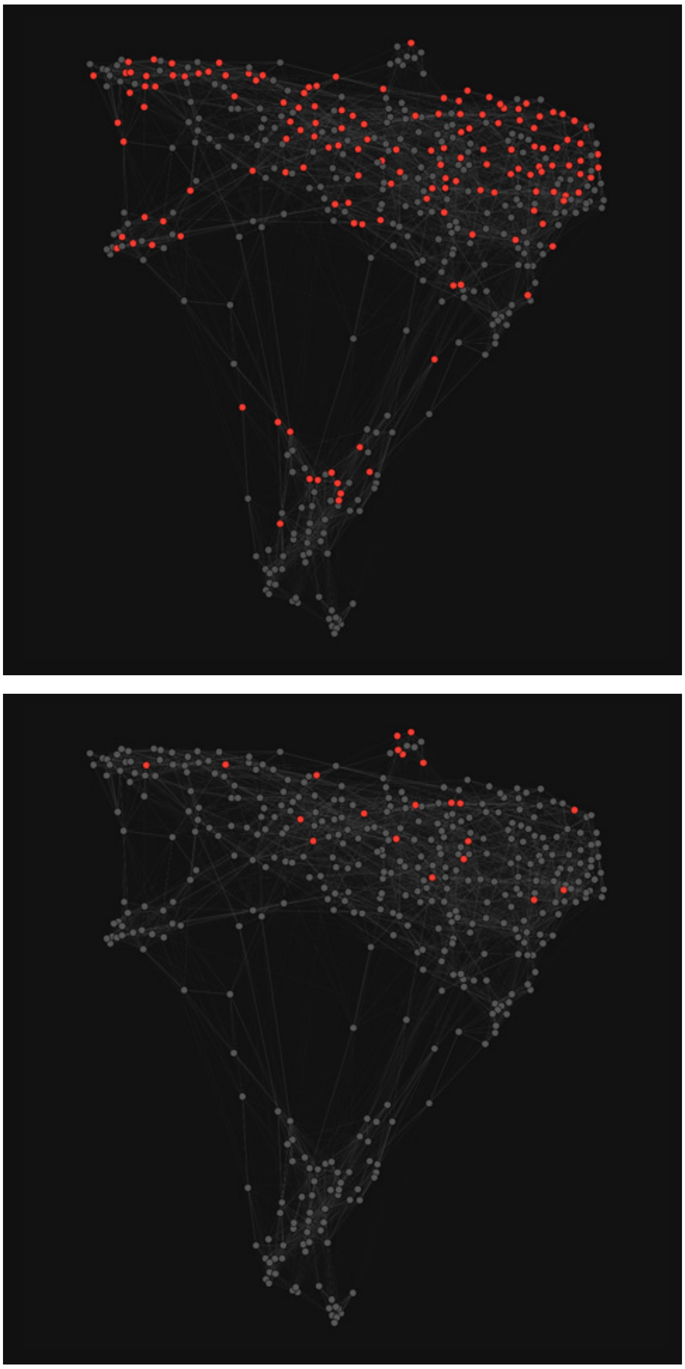


Figure 23: Network maps showing distribution of racial (left) and gender-based hate (right). Highlighted accounts have contributed at least one hateful message to a specific target group.

Part 3: Understanding Explicit Hate Speech

When exploring the concrete impacts of accounts identified as promoting extremist, terrorist and harmful conspiracy theorist content, hate speech is one example of activity which can negatively affect the safety and rights of other social media users, and represents activity which may be violative of platform Terms of Service and potentially legal thresholds. Accordingly, we used a natural language processing algorithm to study the extent to which the range of UK-relevant actors included in this study propagate hate speech.

This technology was subsequently used to inform analysis in the accompanying report in this series which analyses the overall volumes of hate speech across platforms, with data generated from the analysis outlined in this chapter used to train the classifiers used in that wider study. Accordingly, these two chapters utilise closely linked, but distinctive methodologies which each bring with them their own set of distinct caveats and considerations.

Here it should be stressed that these volumes are representative of the speech driven by the 768 social media channels ISD researchers had identified as associated with hate, extremism and dangerous conspiracy theorist content analysed in this project, rather than indicative of broader patterns of hate speech targeting individuals and communities in the UK. Due to the relatively high volume of right-wing extremist channels identified in this project it is possible that certain types of hate speech are over-represented for example. The partner paper in this series explores larger volumes of hate speech targeting UK users produced by a wider range of accounts.

For the purposes of this report, and an accompanying publication exploring the broader prevalence of hate speech online, hate speech is described as:

Activity which seeks to dehumanise, demonise, express contempt or disgust for, exclude, harass, threaten, or incite violence against an individual or community based on a protected characteristic. Protected characteristics are understood to be race, national origin, disability, religious affiliation, sexual orientation, sex, or gender identity.[ix]

Identifying hate speech

Having identified a series of accounts and channels associated with terrorist, extremist and harmful conspiracy theory activity during the process explained above, and determined a working definition of hate speech, we needed to identify a methodology for tracking this type of hateful speech at scale. Due to the scale of messages produced, it would not be feasible to manually label each individual message. Accordingly, we sought to develop a method to facilitate the automated identification of content meeting our working definition of hate speech.

To date there has been extensive effort within academia to develop and train analytical techniques for the identification of hateful speech. We have made use of this body of work in our approach to tracking hate speech.

To achieve this, the project team developed an 'ensemble' approach to classifying hate speech. The structure of our classification approach is the idea of a hierarchical classification pipeline, where outputs from a number of different pre-existing models are used to annotate the data gathered from the accounts, channels and pages outlined above.

In total we identified **22 existing models and 3 lexicons** which were trained to detect hateful content as well as other categories of speech related to hate speech, such as abusive or offensive content. Some of these models have been trained to identify speech targeting specific groups, while some are focused towards narrower problems (such as the hateful use of emojis), others are for general purpose hate detection, and others still are designed to identify counter-speech. The majority of hate detection research on social media to date has been focused on Twitter, but we identified models trained on other platforms too, including Reddit, Facebook, and Instagram.

Additionally, we developed two models with our own training data that were used to classify content, and developed 25 additional lexicons associated with either anti-minority slurs or terminology associated with the groups targeted by hate speech. For a full overview of the different models used, an expanded discussion of this ensemble approach, and discussion of the performance of these models please **see the methodological annex**.

[ix] Please see the Annex of this report for an overview of the process which produced this definition, which is designed to be cognisant of both UK legislation around hate speech and the terms of service adopted by the social media platforms analysed in this study.

Edge Cases and False positives

The core task facing analysts when labelling content was distinguishing between hate speech, and non-hateful speech which contains linguistic markers associated with hate speech.

In some cases, messages contain language which appears hateful, but actually contain counter-speech, or the reclamation of particular slur terms by communities targeted by hate speech. For example:

- Don’t call me a p**i
- My n****s looking good today

Speech in the above categories is often relatively easy for analysts to identify. However, the task becomes more challenging when distinguishing between speech which is offensive (likely to cause someone to be upset), and hateful (targeted speech which dehumanises, demonises, expresses contempt or disgust, excludes, harasses, threatens, or calls for violence). In these cases, there were numerous instances where terminology which could be associated with hate speech was used in an offensive fashion, but where the target group couldn’t be identified, or where there was not language specific to our definition of hate speech. For example: “Shut up you b***h” does not necessarily reach our definition of hate speech without broader contextual information around who that language was targeted at, whilst “women are f***ing subhumans” would be classed as hate speech.

These edge cases proved particularly challenging as it is recognised that the use of certain language is likely to cause distress amongst target groups and is representative of a hateful worldview, and broader societal dynamics of ingrained prejudice. Accordingly, as it is desirable to keep note of this type of speech, analysts coded messages where hate-relevant language was used in an unclear fashion as ‘offensive’, rather than ‘hateful’.

The process for using this ensemble approach to classification followed these steps[x]:

- **Step 1**: Messages produced by accounts or channels associated with extremist, terrorist and harmful conspiracy theorist activity were filtered for those potentially relevant to hate speech.

- **Step 2**: A team of ISD experts reviewed samples of this content to check the accuracy of this stage of analysis, marking up messages as either ‘hateful’ or ‘offensive’ in their targeting of communities included in our definition of hate speech (see box above).

- **Step 3**: Potentially hateful messages were then run through the ensemble of pre-existing classifiers identified through a scope of available models. Messages were annotated based on the judgement of pre-existing models.

- **Step 4**: A team of ISD experts reviewed samples of the content and annotated them as either hateful or not. This process used blind coding to ensure inter-coder reliability and the outputs were compared to the classifications of the machine learning models.

- **Step 5**: Subject matter experts assessed the accuracy of the pre-existing models, and built an additional layer of machine learning. This completed the pipeline, and resulted in an algorithmic solution that is capable of detecting hate speech with 70% accuracy.[xi]

Hate speech classification

In total, for this exercise we gathered **317,932 messages sent by the harmful accounts on Facebook, Twitter, Instagram, Reddit, Telegram,** and **321,830 comments left on extremist YouTube channels** identified for this study. An accompanying report explores hate on 4chan.

[x] Please note that the ensemble methodology employed to detect hate speech across this project is covered in detail in the accompanying report Hate of the Nation which provides analysis of the broader prevalence of hate speech across social media.

[xi] This accuracy is an F1 score derived from the mean of a 68% recall rate (the proportion of hateful content the model detected) and a 73% precision rate (how often these annotations were correct). Both training and evaluation data for this set were taken from the range of platforms included in this study.

Platform	Messages/Comments
Facebook	25,613
Instagram	8,220
Reddit	1,249
YouTube	321,830
Telegram	23,402
Twitter	259,447

Table 24: Total numbers of messages and comments gathered for this study

After being subjected to **Step 1** of the analysis outlined above, these messages were then filtered to remove content which is likely irrelevant to hate speech, and to identify potentially hateful content – which was determined by our classifiers to be 'offensive', or which contained terminology our subject matter experts identified as being relevant to hate speech. This provided us with a set of **160,330** comments and messages which were passed through our ensemble model.

Platform	Messages/Comments
Facebook	3,319
Instagram	1,484
Reddit	168
YouTube	101,723
Telegram	2,775
Twitter	50,861

Table 25: Total number of potentially hateful messages gathered per platform

These **160,330** messages were then run through the ensemble of classifiers outlined above. This left us with a total of **2,260** messages and comments containing content that met our working definition of hate speech. This suggests that **only 0.35% of messages sent by the accounts who expressed support for extremism, terrorism or harmful conspiracy theories analysed in this report contain explicit hate speech (within the meaning of our working definition)**. This is likely reflective of the high threshold for hate speech set by our definition, which sought to distinguish hate speech from broader speech which could be considered offensive or insulting.

This finding seems counter-intuitive, as one could reasonably expect that extremists and hateful actors would focus a large proportion of their conversation on hateful speech. However, this does match findings produced by ISD in other international contexts, which suggested that only a small proportion of extremist conversation is explicitly hateful.[44]

There are several possible explanations for this finding. Perhaps the most obvious is that individuals who have expressed support for extremism, terrorism and harmful conspiracy theories discuss a wide range of topics beyond anti-minority hatred, including both content designed to reinforce an extremist worldview – such as news reports – and innocuous non-political issues such as sports results.

Another possible explanation is that people who have expressed support for extremism, terrorism and harmful conspiracy theories promote supremacist world-views but in a way that avoids overt dehumanisation, demonisation or contempt. This is reinforced when the results of the offensiveness classifier which was used as part of the ensemble are considered, which identified **5,371 offensive messages**.

To explore this concept further we explored samples of the messages produced by accounts which have expressed support for extremism, terrorism and harmful conspiracy theories that were not marked as hate speech by the ensemble classifier. Out of a random sample of 500 messages which were deemed as not-hateful by our classifier we identified 25 messages (5%) which were deemed as edge cases – that is to say, messages which referenced hateful tropes, but which did not cross the threshold of overt hate speech, according to our working definition. Bowdlerised examples of these messages are provided below:[45]

- **Example 1**: The Kalergi Plan is the mass movement of people designed to cause a crisis. This has been in place since the 1940s. Its aim is to replace white people and abolish the notion of nation.

- **Example 2**: There is a total lack of respect for white indigenous people in this country!!!

- **Example 3**: Modern Britain is disgusting! We turn

Platform	Facebook	Instagram	Reddit	Telegram	Twitter	YouTube
Total number of messages gathered per platform	25,613	8,220	1,249	23,402	259,448	321,830
Number of hateful messages	27	33	2	59	1,193	946
Percentage of hateful messages per platform	0.11%	0.42%	0.16%	0.31%	0.45%	0.29%

Table 26: Numbers of hateful messages gathered by platform

> our back on our own people and allow them to be attacked by foreigners.

These examples of messages illustrate how it is possible to promote a worldview in line with the definition of extremism in this study. All three messages reinforce an 'us vs. them' mind-set and prioritise or promote a particular in=group (in this case white British people). However, these messages are not immediately identifiable as extremist when taken out of context, and all avoid overt hate speech which explicitly targets an individual or community.

This delineation makes clear the 'grey area' of content which is produced by online extremists. In the case of the first example given above the user references an antisemitic conspiracy theory which is popular amongst white supremacists and far-right extremists.[46] However, the text of the post does not use readily identifiable facets of hate speech such as slurs, and without an expert understanding of the conspiracy theories used by extremists, would not be identifiable as a white supremacist talking point. This highlights how understanding of the context within which speech is made is important in understanding its meaning. Another potential explanation for the apparently low proportion of content meeting our working definition of hate speech produced by accounts which have expressed extremist views could be that this is reflective of a deliberate tactic to avoid moderation by social media platforms.

Proportion of hate speech by platform

Across the platforms analysed, Reddit had the fewest messages coded as hate speech, whilst Twitter had the most (which again, may be an artefact of more data access on the platform). Only Tweets and comments on YouTube videos contained volumes of hate speech above double figures. Whilst one possible explanation for this could be differing content moderation practices, other factors might also be at play here. For example, proportionally Instagram had four times as much hate speech as Facebook, despite both being Meta platforms with the same community guidelines, although these were both relatively low absolute numbers (33 and 27 posts respectively).

A factor that may have had an impact on our ability to detect hate speech is data access. For example, previous analysis has identified high volumes of hateful visual content on Telegram,[47] which our language-based approach cannot detect, whilst the CrowdTangle API only allows posts sent by the admins of pages and groups to be analysed, with the result that comment threads on these posts were not subject to analysis. Another factor at play includes disparities in the number of messages analysed; due to limited numbers of overtly hateful UK-focused subreddits, we gathered and analysed far fewer Reddit messages than those on other platforms.

In general, the distribution of the volumes of hate-speech matched the overall total of hateful messages per platform, although notably YouTube comments contained the highest volumes of hate speech based on sexual orientation and gender identity.

When the overall volumes of hate speech are broken down, we found that hate speech targeted on the basis of race constituted the greatest proportion of hate speech produced by the users analysed in this study at 43% of messages analysed, followed by hate speech made on the basis of religion, at 37%, with xenophobic hate speech on the basis of national origin the next most prominent in the data.

Additionally, we were able to sub-divide some portions of the hate speech gathered more specifically by the target community in question. This revealed that the

Account Type	Disability	National origin	Race	Sexual orientation	Religion	Sex	Gender identity
Facebook	0	16	10	0	10	5	2
Instagram	0	15	19	2	10	9	2
Reddit	0	1	0	1	1	1	0
Telegram	1	27	36	3	0	14	3
Twitter	13	485	568	9	500	154	22
YouTube	9	294	340	20	309	116	28

Table 27: Breakdown of type of hate speech per platform

Platform	Facebook	Instagram	Reddit	Telegram	Twitter	YouTube	Total volume
Anti Black	1	5	0	9	60	86	151
Anti disability	0	0	0	1	13	9	23
Anti east Asian	0	2	0	0	4	16	23
Antisemitism	1	4	0	13	190	149	357
Anti south Asian	1	3	0	1	24	7	36
Homophobic speech	0	2	0	3	13	23	41
Transphobic	2	2	0	3	25	36	68
Anti Hindu	2	0	0	1	3	6	12
Misogny	2	3	1	3	33	62	104
Anti Muslim	10	6	1	9	337	192	555
Anti National origin	19	16	1	31	665	336	1,068
Intra-Christian Sectarian	0	0	0	6	2	0	8
Anti Sikh	0	0	0	1	0	0	1

Table 28: Breakdown of hate speech by community targeted

Account Type	Far-right extremism	Islamist extremism	Hateful	Harmful conspiracy theorist	Terrorism
Facebook	24	1	27	0	0
Instagram	26	0	31	1	1
Reddit	1	0	1	1	0
Telegram	55	1	54	2	0
Twitter	680	62	1,019	36	7
YouTube	734	36	939	0	3

Table 29: Breakdown of hate speech by account type

highest proportion of hate speech gathered in this study targeted individuals based on their national origin (47% of hate speech gathered), followed by anti-Muslim hate speech (24%), antisemitism (15%) and anti-black hate speech (7%).[xii]

As well as exploring the volumes of hate speech by platform we also explored the volumes of hate speech produced by the different communities of actors identified. Given that a majority of the accounts which were identified as sharing extremist content were also coded as targeting hate against groups with protected characteristics, it is unsurprising that this set of accounts is the most represented in our set. However, this analysis also reveals other interesting trends. Perhaps most noteworthy is the fact that the small number of accounts supportive of Northern Ireland-related terrorism, a subset of accounts sharing terrorist-related content, only produced four explicitly hateful comments. This is perhaps a reflection of the fact that although these accounts are tied to illegal and violent activity, their endorsement of terrorism does not manifest as hatred towards communities with protected characteristics. Conversely, far-right extremists produced a larger volume of content classified as hate speech – which is likely to be a reflection of their exclusionary worldview which overwhelmingly targets minority communities.

Here it should be stressed that these volumes are representative of the speech driven by the social media channels associated with hate, extremism and dangerous conspiracy theorist accounts and channels analysed in this project, rather than indicative of broader patterns of hate speech targeting individuals and communities in the UK. Due to the high volume of right-wing extremist channels analysed in this project it is possible that certain types of hate speech are over-represented.

[xii] For the purposes of this study Intra-Christian sectarian hatred included anti-Catholic and anti-Protestant hatred delivered in the broader context of sectarian online discussion in the UK, and in particular in Northern Ireland.

Part 4: Cross-Platform Link Analysis

Examining the URL links shared by harmful actors online may help researchers to identify other potential venues for harmful activity, and spot emerging spaces which users who share extremist content may migrate to, especially in the face of increasing enforcement of major platforms' terms of service.[48] The following section outlines the findings from our analysis of all links shared to social media platforms within our overall dataset of 2,531,090 messages gathered from UK accounts and channels identified as hate, extremist and terrorist actors between 1 October 2021 and 31 March 2022. The aim of this exercise was to identify how frequently accounts in our dataset share links to other platforms, both to the 7 platforms analysed for this report as well as platforms that are smaller, emerging or have limited data access (e.g. TikTok).

As the specific scope of this project did not allow for systematic analysis of the content of the links themselves, this research cannot draw strong inferences about the potentially harmful substance of such out-links, or the nature of discussions on the platforms that were linked to. However, such data can nonetheless provide a general indication of which platforms and services beyond the scope of this study might potentially be of interest for accounts spreading content associated with terrorism, extremism and hate.

Our findings below show that YouTube was the platform linked to the most, followed by Twitter, perhaps not surprising given the ubiquity of these services within the general social media landscape. Of potentially greater relevance is an examination of the prominence of links from extremism-promoting accounts on various platforms such as 4chan, Twitter, Instagram and Facebook directing their followers to Telegram, hinting at the platform's prominence within these communities. Lastly, smaller and emerging platforms Bitchute, Odysee, Gettr and Rumble were all linked to more often than Facebook, Instagram or Reddit.

Data Gathering

For all messages, links were extracted, expanded, and converted to their root domain.[xiii] Of the just over 2.5 million messages in our dataset, just over 10% contained one or more links, resulting in over 350,000 links extracted in total.

Platform	Messages	Messages containing links	Total links extracted
4Chan	1,891,328	78,103	107,767
Facebook	25,613	24,255	60,163
Instagram	8,220	8,220	9,848
Reddit	1,249	1,165	1,348
Telegram	23,402	7,718	8,449
Twitter	259,448	133,361	175,624
YouTube	321,830	2,653	3,942
Total	2,531,090	255,475	367,141

Table 30: Total number of messages, messages containing links and links extracted per platform

As we can see in table 10, most links were extracted from Twitter, 4chan and Facebook, followed by considerably fewer from YouTube, Instagram, Telegram and Reddit. There are major differences in how frequently links are shared on different platforms: while the average Facebook post contains 2.35 links, only 6 in 100 4chan posts and 1 in 100 YouTube comments contain a link, reflecting differences in platform functionality and ways in which platforms are used for communication by actors we have classified as terrorist, extremist and harmful actors.

Results

Our overall results show that YouTube is by far the most linked-to platform (over 50,000 links, more than three quarters of the total), predominantly from 4chan and Twitter. Twitter follows next, while Telegram, Facebook, Reddit and Instagram are linked to much less frequently.

Facebook is comparatively frequently linked to from Twitter, while Instagram is linked to in almost equal volume from 4chan and Twitter. It is interesting to note that within the digital subculture on 4chan's /pol/ board, the traditional social media giant Facebook is rarely linked to.

Despite the high volume of comments from UK 4chan users in our dataset, the platform is hardly ever linked to by extremist accounts on the other platforms in scope of

[xiii] For example, https://t.co/12345 would be expanded to https://en.m.wikipedia.org/wiki/COVID-19_pandemic and then converted to en.m.wikipedia.org

Links between Hate, Extremist and Terrorist Actors on Platforms investigated

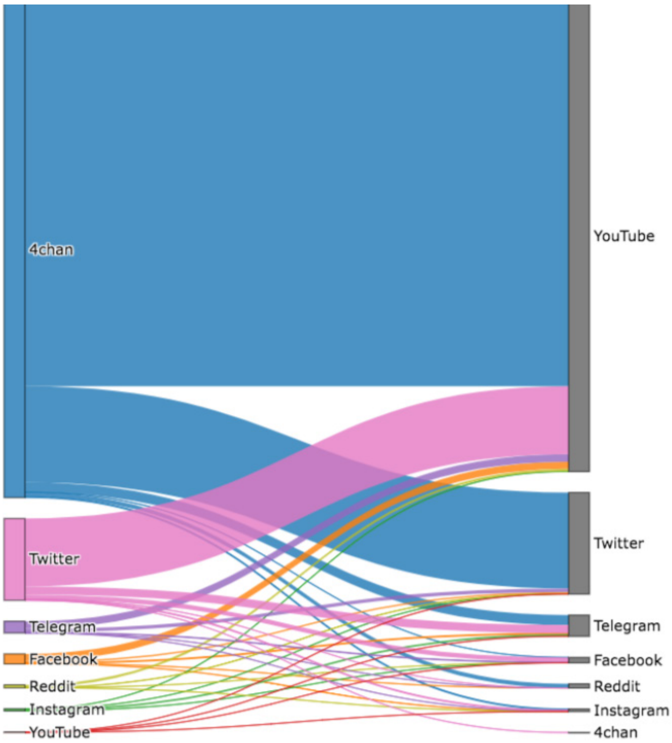


Figure 31: Patterns of cross-platform links to the social media platforms within the study. This shows the proportion of outlinks from the platforms analysed (left axis) and the platforms which are linked to (right axis).

the research. There could be several explanations for this. Perhaps the most likely explanation is that 4chan posts are highly transient and ephemeral – since 4chan posts will disappear within a few days, directly linking to them is less helpful than providing a screenshot of relevant content.[49] Other explanations include the possibility that 4chan users form a community which is distinct from users who share content associated with extremism on other platforms, or the platform so well known for its role for its transgressive humour and meme culture that it generates organic interest and no longer needs to be linked to.[50] Reddit was mostly linked to from 4chan, again reflecting that beyond a specific audience the platform seems to not play a major role for UK hate, extremist and terrorist accounts and channels online.

Beyond the seven platforms analysed for this report, ISD researchers also compared the number of links to other social media platforms. The results show that there is a group of alt-tech[xiv] platforms that are frequently linked to by extremist accounts on other platforms: Bitchute (1,915 links), Odysee (1,915 links), Gettr (818 links) and Rumble (767 links) were each linked to more often than Facebook, Instagram or Reddit.

The most frequently linked to of these platforms is Bitchute, a UK-based video-sharing site that is known for hosting the content of creators whose videos or accounts were previously blocked on larger video-sharing sites such as YouTube. Bitchute, which was mainly linked to from 4chan and Twitter within our data, claims it was created in response to Internet censorship but has been accused by organisations including the UK Community Security Trust of hosting racist, antisemitic and extremist content, including videos of far-right terror attacks and propaganda videos from the proscribed terrorist group National Action.[51]

Odysee, the platform which received the second-most links among platforms not systematically analysed for this report, is a video-hosting platform that creators use to monetise their content. Odysee serves as an alternative service to major video-hosting platforms such as YouTube and it has been reported by some commentators that it is being increasingly favoured by far-right extremist actors.[52] Gettr is a social media app founded by former Trump aide Jason Miller in 2021. There has been criticism of the platform's content moderation after ISIS supporters joined the platform en-masse to share extremist propaganda, according to an investigation by the Institute for Strategic Dialogue.[53]

Another platform within our top 10 links shared is Dlive (385 links), a livestreaming platform with in-built opportunities for monetisation using cryptocurrency. Previous ISD research has indicated that it is being used by British white nationalists to broadcast their ideology, though only rarely in an attempt to radicalise new users.[54] Our data contained relatively few links to TikTok (146

[xiv] ISD uses the term 'alt-tech' to refer to social media platforms used by groups and individuals who believe major social media platforms have become inhospitable to them because of their political views. This includes platforms built to advance specific political purposes; libertarian platforms that tolerate a wide range of political positions, including hateful and extremist ones; and platforms which were built for entirely different, non-political purposes like gaming.

Platform	4chan	Facebook	Instagram	Reddit	Telegram	Twitter	YouTube	Total volume
Bitchute	1,556	18	2	0	42	283	14	1,915
Odysee	627	11	5	85	350	535	15	1,628
Gettr	17	165	0	6	42	579	9	818
Rumble	152	39	0	7	53	509	7	767
DLive	5	3	2	0	131	244	0	385
Brandnewtube	157	7	3	0	6	206	0	379
Gab	31	8	0	34	18	58	16	165
TikTok	84	19	2	1	4	36	0	146
Stormfront	112	0	0	0	0	1	0	113
Soundcloud	47	0	1	0	9	52	0	109

Table 32: Outlinks to other platforms across accounts analysed in this study

links), despite its immense popularity among younger users.

The other platforms in the top 10 included: Brandnewtube (379 links), another alternative video-sharing platform; Gab (165 links), one of the initial alt-tech platforms whose founder has been criticised by some commentators for endorsing far-right extremist positions;[55] and SoundCloud (109), a music streaming app. The long-standing white supremacist forum Stormfront (113 links), which was primarily linked to from 4chan, was the ninth most linked to among these smaller and emerging platforms.

While platforms such as Bitchute, Dlive and Gab have been around for several years, Odysee and Gettr have been created relatively recently. Rumble, a video platform, was launched in 2013, but it has reportedly started increasingly attracting far-right extremist followers (largely from Parler) since 2020.[56]

Links between Hate, Extremist and Terrorist Actors on Platforms investigated

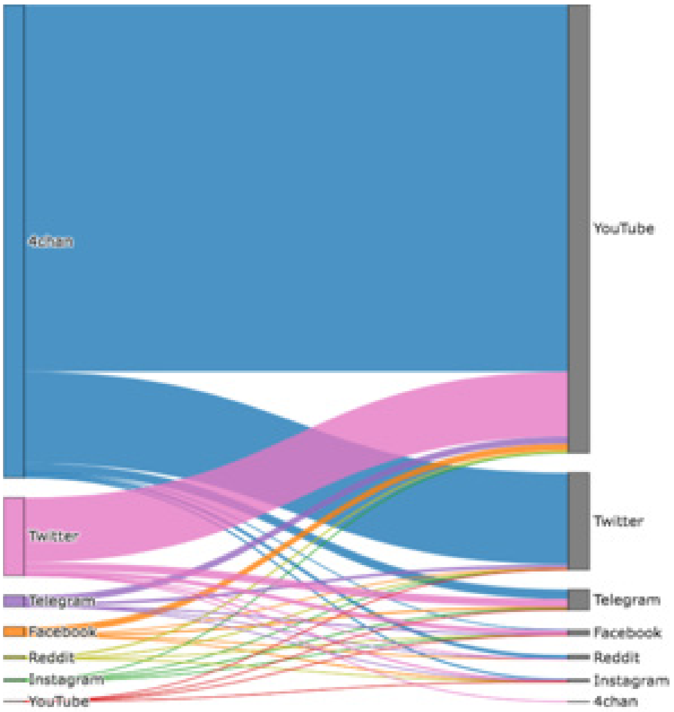


Figure 33: Patterns of cross-platform links from platforms analysed (left axis) to emerging or alternative platforms (right axis) linked to

Conclusion

In this report, we provide a qualitative and quantitative overview of the interconnected online communities of accounts and channels propagating terrorist and extremist content, targeted hate speech and harmful conspiracies to UK audiences. Focusing analysis on a range of relevant social media platforms where data can be collected computationally through application programming interfaces (APIs), this study combines ISD's expertise on these online communities with a unique suite of digital analytics approaches to help understand their composition, behaviour and reach across platforms.

Platform specific snapshots from 4chan, Telegram, Twitter, Facebook, Instagram, YouTube and Reddit draw on a dataset of 768 UK-relevant channels, groups and pages coded by expert analysts as meeting our definitions of terrorism, extremism, hate speech or harmful conspiracy. Our analysis shows there is no single central hub for hateful, extremist and terrorist related content related to the UK on social media, with platforms instead being used in distinct ways by different actors – whether communicating to established online communities or reaching broader audiences.

To help drill down on the level of explicit hate across our dataset, our research approach created an 'ensemble' of algorithms to identify relevant speech, which allowed for much greater granularity of analysis than existing classifier-based approaches. The relatively low proportion of explicit hate speech detected within our dataset (0.35% of all posts), shows the challenge of isolating harmful activity even amongst known communities of concern.

The report uses innovative natural language-based techniques to conduct multi-platform analysis of the relationship between how these diverse online accounts use language, beyond the coding of analysts. Mapping out these relationships allowed us to identify groups of accounts that talk about things in similar ways, revealing nine distinct communities – ranging from anti-immigration nationalists to Islamist extremist promoters. The counter-intuitive clustering of these communities indicates significant continuity between harmful conspiracy theorists, anti-immigration movements and overtly white nationalist communities, with the boundaries between seemingly distinct online communities becoming ever more ambiguous.

Finally, analysis of out-links to platforms beyond those included in our study reveals the growing relevance of smaller and emerging 'alt tech' services such as Bitchute, Odysee, Gettr and Rumble, all of which were linked out to within our data more often than Facebook, Instagram or Reddit.

Annex A: Definitial Discussion of Key Terms

To guide the analysis outlined in this report, ISD built on established understandings within the academic and policy domains to develop working definitions of key concepts – namely 'extremism', 'hate speech', 'terrorism' and 'harmful conspiracies' – aimed at relating online activity to concrete harms. While these definitions were informed by discussion with Ofcom around which harm areas may be in scope, they are not intended to reflect or anticipate UK legal frameworks. These working definitions are outlined in full below:

Hate Speech

There are a range of differing conceptions of hate speech. There are those that are enshrined in legislation, those that are proposed by advocacy groups representing particular communities, and those established by private companies, such as social media platforms, which help determine acceptable behaviour.

Hate speech encapsulates certain hateful activities targeting groups based on protected characteristics, so when distilling a programmatic definition out of these two areas governing hate speech it is also necessary to determine what behaviours constitute hate speech, and what *communities* are being targeted.

In nearly all cases internationally, hate speech is differentiated from offensive speech, based on the understanding that to maintain strong democracies even speech that is seen as offensive must be permitted. However, speech that threatens individual's rights (such as their right to live free from discrimination) or may cause violence against certain groups can be regulated and prevented—often through the frame of illegal hate speech.

In the UK there are a number of different laws which govern hate speech, including the Public Order Act 1986,[57] Criminal Justice and Public Order Act 1994,[58] Racial and Religious Hatred Act 2006[59] and the Crime and Courts Act 2013.[60] In addition to laws around speech, there is also legislation around hate crime, whereby any crime can be prosecuted as a hate crime if the offender has either a) demonstrated hostility based on race, religion, disability, sexual orientation or gender identity, or b) been motivated by hostility based on race, religion, disability, sexual orientation or transgender identity.

Based on a synthesis of existing UK legislation outlined above, as well the relevant sections in the terms of service of the social media platforms (which have adopted voluntary frameworks regarding these activities which often extend beyond the parameters of legislation) studied here, the working definition of hate speech used in this project is:

Activity which seeks to dehumanise, demonise, express contempt or disgust for, exclude, harass, threaten, or incite violence against an individual or community based on a protected characteristic.

Protected characteristics are understood to be race, national origin, disability, religious affiliation, sexual orientation, sex, or gender identity.

Extremism

While there is no legal definition of extremism, the Counter Extremism Strategy 2015 describes extremism as: "the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist." While there were plans to enshrine this conception in law as part of a proposed Extremism Bill in 2015, to date efforts to establish legislative definitions of extremism have been unsuccessful.

This framing of extremism has been criticised as "ambiguous and incoherent" by the Commission for Counter-Extremism in their 2019 report, Challenging Hateful Extremism, which established a working definition for a new term - 'hateful extremism' - constituting behaviours which:

- Incite and amplify hate or engage in persistent hatred or equivocate about and make the moral case for violence.
- Draw on hateful, hostile or supremacist beliefs directed at an out-group who are perceived as a threat to the wellbeing, survival or success of an in-group.
- Cause or are likely to cause harms to individuals, communities or wider society.

ISD believes that it is crucial to properly define terms like 'hateful extremism', to avoid drawing a misleading equivalence between hate and extremism. In particular, it is important to appreciate the distinction between prejudice, hate and extremism in order to ensure the responses to these problems are relevant, proportionate and differentiated. While hate can result from the propagation of extremist ideology, not all hate, or indeed hate crime, is extremist in nature.

ISD seeks to capture this distinction in its in-house definition, which defines extremism as the advocacy of a system of belief that claims the superiority and dominance of one identity-based 'in-group' over all 'out-groups'. Extremism advances a dehumanising 'othering' mind-set incompatible with pluralism and universal human rights.

This 'social identity' based definition seeks to avoid key pitfalls of other definitions of extremism, which define extremism against 'mainstream' values and norms: as such values and norms evolve over time, such approaches necessarily produce relativist and subjective understandings of extremism. At the same time, defining extremism against 'mainstream' values and norms ends up equating radical and extremist movements. It thereby risks conflating dissident movements advocating for human rights and democracy in authoritarian regimes with extremists who advocate for supremacist, exclusionary and anti-democratic worldviews.[61]

This broader definition used by ISD was narrowed to focus on specific illegal and harmful content in scope of platform policies and potential regulation under the Online Safety Bill at the time of writing this report. Our research therefore focuses specifically on extremist content associated with incitement, violent threats or harassment, which directs hate against a protected group, or which perpetuates harmful disinformation (understood as false, misleading or manipulated content presented as fact, intended to deceive or harm).

Specific manifestations of extremism referenced in this report include (but are not limited to):

- **Far right extremism**: A form of nationalism that is characterised by its reference to racial, ethnic or cultural supremacy. Right-wing extremism is the advocacy for a system of belief in inequality based on an alleged difference between racial/ethnic/cultural groups.   Extremism expert Cas Mudde characterises the far right as commonly exhibiting these features: nationalism, racism, xenophobia, anti-democracy and strong state advocacy.[62]

- **Islamist extremism**: The advocacy of a system of belief that promotes the creation of an exclusionary and totalitarian Islamic state, within which those who do not subscribe to this vision are portrayed as an inferior 'out-group' and are subjected to implicit, explicit or violent means of subjugation and prejudice. This ideological goal might be pursued through violent action, political activism or systematic societal change.

Harmful conspiracy theories

Conspiracy theories explain events in terms of a small group of powerful persons acting in secret for their own benefit and against the common good. In the wake of Covid-19 we observed the significant proliferation of harmful conspiracy networks around the world, bringing together loose coalitions around crisis points, shared goals and common objectives.

In this context, mobilisation and threats of violence come from a wide array of actors, not just ostensibly violent groups. An increasingly prominent subset of harmful conspiracy movements such as QAnon have been linked to violent radicalisation and are prompting responses from platforms, such as Meta's policy on violence-inducing conspiracy networks.[63]

For the purposes of this report, we focus on conspiracy movements associated with real-world harm, including the incitement of violent threats and harassment, or hate directed against a protected group. We have not used a broader conception of harm that might include, for example, potential threats to public health posed by Covid-19 conspiracy theorists, unless these actors were found to incite violence, make threats, engage in harassment or direct hate against a protected group (although we do include qualitative analysis of this phenomenon as a unifying narrative among accounts associated with terrorism, extremism and hate speech in the report).

Terrorism

The UK Terrorism Act (2000) defines terrorism as the use or threat of action, designed to influence any international government organisation or to intimidate the public. It must also be for the purpose of advancing a political, religious, racial or ideological cause.

The research has additionally been guided by the framing of terrorist content and activity online expressed in the 2020 Interim Codes of Practice from DCMS and the Home Office:

Online terrorist content is any content which, by uploading it or otherwise making it available to others online, a person is committing an offence under UK terrorism laws. Terrorist content online can take many forms, including but not limited to statements, imagery (including still images and others such as GIFs), videos (both live and pre-recorded), voice recordings and documentation such as leaflets, papers and posters.

Online terrorist activity means any action taken by a person online that forms part of an offence under UK terrorism laws. Generally, it is the means and techniques by which terrorists and their supporters build community, disseminate content and communicate online for terrorist purposes, including through the exploitation of differing services and accounts.

Drawing on these frameworks, ISD's working definition focused on accounts expressing support for groups or organisations proscribed under the Terrorism Act, as well as the broader online behaviours by which terrorists and their supporters build community, disseminate content and communicate online for terrorist purposes, outlined in the Home Office Interim Codes of Practice. While most of the accounts identified expressed support for proscribed groups, a small number of accounts promoted groups or behaviours judged to fall under the UK Terrorism Act's definition of terrorism.

Annex B: Hate Speech Models in Ensemble Classifier

Hatebert. This is a model trained using a transformer-based machine learning technique called Bidirectional Encoder Representations from Transformers or BERT. It is trained on a large dataset from Reddit (called RAL-E) of comments banned for being offensive, abusive or hateful.[64] It determines whether a post is hateful or not. Subset models include Hateabuse based on the Hatebert approach above, but instead is trained to identify abusive posts.

Hateoffence is also based on the Hatebert approach above, but instead is trained to identify offensive posts. **Hateval** is based on the Hatebert approach above, but instead is trained to identify hateful posts.

Dehatebert. This was an attempt to detect hateful speech in 9 languages across 16 different sources. It was a comparison of different approaches in different languages.[65] **Mono** is a version of Dehatebert to identify hateful posts.

HateXplain was an attempt at automated hate speech detection, also to identify the target community and identify what study calls the 'rationales'; the portion of the post on which the labelling decision most depended. This is intended to increase the interpretability of the model.[66] **Rational2** determines if a post is abusive or not, whilst **hate-explain-bert-base-uncased** determines if a post is hateful, offensive or neither.

Detoxify. These are a set of models that provide a score on how likely a post is to contain certain 'toxic' traits.[67] **The Original**[68], **Unbiased**[69] and **Multilingual**[70] models each give each post a score on the following attributes:

- Toxicity
- Severe toxicity
- Obscene language
- Threatening language
- Insults
- Identity attack
- Sexually explicit language (in the case of the latter two).

Hate alert-counter. These models focus on counter-speech, language that is calling out or undermining, opposing or mocking hateful speech in some way. The models usually classify these as hateful speech, so these models are useful to increase the precision of the hybrid ensemble but removing counter-speech as examples of false positives. **Binary** identifies if a post is counter-speech or not. Multi-label identifies what kind of counter-speech is being used, including:

- Presenting facts
- Hypocrisy or contradiction
- Warning of consequences
- Showing affiliation with the group
- Denouncing the hate speech
- Humour
- Posts that have a positive tone
- Posts that are hostile to the hate speech poster

A series of additional models also identify counter-speech specific to posts targeting Black, Jewish and LGBTQ+ communities.[71]

HateALERT-EVALITA. These are a series of models trained for 'Automatic Misogyny Identification' (AMI), which won a prize at EVALITA2018, a period campaign to assess the performance of NLP tools.[72] This includes an overall decision about whether a post is misogynistic, whether the post targets an individual or a more general group, and the type of misogyny being expressed, covering:

- Discrediting
- Derailing
- Dominance
- Sexual harassment
- Stereotype

Hatesonar. An approach that used crowdsourcing to train models to distinguish between hateful and other instances of offensive language.[73]

Rewire. This is a commercial model to determine if a post is hateful or not hateful. We have been granted access to this model for the purposes of this research.[74]
Tisane. This is a commercial model which evaluates each post and can give annotations in the following areas:

- Personal Attacks
- Bigotry
- Profanity
- Sexual advances
- Criminal activity
- External contact

- Adult only
- Mental health issues
- Spam
- Generic

Within these definitions there are also further rationales that could be used to identify which combination of annotations are useful to identify posts that fall under our definition of hate speech.

Lexicons

In addition to the models described, messages can also be analysed more simply by whether or not they contain a given word. First, a number of externally compiled corpora have been identified. Rather than being used as a basis for detecting hate speech, these broad lexicons were used as an initial filter, with the combination of other annotators instead being the basis for decision making. As such, no single annotator determining a post as hateful results in the final decision of a message being classified as hate speech.

T-davidson. 178 words that are commonly used in hate speech were manually curated as a list. Each has a score of how likely the post is to be hate speech when the phrase is included.[75]

Hatebegets-hate. A list of 187 offensive terms that are used against different groups of people commonly in hate speech posts.[76]

Spread_Hate_Speech_WebSci19. A list of 81 offensive terms commonly present in hate speech.[77]

Across a number of different projects, ISD teams have maintained a series of lists, or 'lexicons', of specific offensive terms and identifiers for particular groups. These are splits into slurs and group-specific identifiers to aid the identification of target groups.

Endnotes

[1] Online Nation 2022. Ofcom.
Accessed at: https://www.ofcom.org.uk/__data/assets/pdf_file/0023/238361/online-nation-2022-report.pdf

[2] "Internet and radicalisation pathways: technological advances, relevance of mental health and role of attackers", UK Ministry of Justice (2023). https://www.gov.uk/government/publications/internet-and-radicalisation-pathways-technological-advances-relevance-of-mental-health-and-role-of-attackers.

[3] Davey, Jacob and Milo Comerford. "Between Conspiracy and Extremism: A Long COVID Threat? An Introductory Paper."
Institute for Strategic Dialogue (2021).

[4] Berger, J.M. "THE ALT-RIGHT TWITTER CENSUS." Vox Pol (2018). https://www.voxpol.eu/download/vox-pol_publication/AltRightTwitterCensus.pdf; Lewis, Rebecca. "Alternative influence: Broadcasting the reactionary right on YouTube." (2018); Guhl, Jakob, and Jacob Davey. "A safe space to hate: White supremacist mobilisation on telegram." Institute for Strategic Dialogue (2020); Crawford, Blyth, Florence Keen, and Guillermo Suarez-Tangil. "Memetic irony and the promotion of violence within chan cultures." (2020); Gaudette, Tiana, et al. "Upvoting extremism: Collective identity formation and the extreme right on Reddit." New Media & Society 23.12 (2021): 3491-3508; Walther, Samantha, and Andrew McCoy. "US extremism on Telegram." Perspectives on Terrorism 15.2 (2021): 100-124.

[5] "The most popular social networks in the UK", YouGov. https://yougov.co.uk/ratings/technology/popularity/social-networks/all

[6] Crawford, Blyth, Florence Keen, and Guillermo Suarez-Tangil.
"Memetic irony and the promotion of violence within chan cultures." (2020)

[7] Rogers, Richard. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." European Journal of Communication 35.3 (2020): 213-229; Amarasingam, Amarnath, Shiraz Maher, and Charlie Winter. "How Telegram disruption impacts Jihadist platform migration." CREST report (2021).

[8] https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice

[9] Mudde, Cas, The Ideology of the Extreme Right, (Oxford University Press, 2000).

[10] Comments from users 4chan and YouTube were excluded. The rationale for this was that while these comments come from individual users that are taking part in discussions on a board known for extremism (4chan's /pol/ board) or reacting to extremist content (on YouTube), the individual users have not been classified as extremist. These comments would thus be more comparable to replies on Twitter, which we likewise did not include for this mapping exercise. It was therefore decided these comments were not appropriate to include for this specific network map of extremists, unlike the posts analysed from Facebook, Instagram, Reddit, Twitter and Telegram, which come directly from accounts classified as hateful, extremist or terrorist by subject matter experts.

[11] The all-mpnet-base-v1 can accept up to 512 word-pieces (approximately 300-400 English words). Longer sequences are truncated. All messages are treated individually, threads of Tweets are not grouped. https://www.sbert.net/examples/applications/computing-embeddings/README.html#input-sequence-length

[12] Accessed at: https://developer.twitter.com. This situation may change with planned shifts to researcher data access provisions.

[13] Accessed at: https://crowdtangle.com

[14] Accessed at: https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking

[15] Accessed at: https://core.telegram.org/

[16] Accessed at: https://www.reddit.com/dev/api/

[17] Accessed at: https://praw.readthedocs.io/en/latest/code_overview/other/listinggenerator.html

[18] Accessed at: https://developers.google.com/youtube/v3/

[19] Fiesler, Casey, et al. "What (or who) is public? Privacy settings and social media content sharing." Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 2017.

[20] Guhl, Jakob, and Jacob Davey. "A safe space to hate: White supremacist mobilisation on telegram." Institute for Strategic Dialogue (2020); Crawford, Blyth, Florence Keen, and Guillermo Suarez-Tangil. "Memetic irony and the promotion of violence within chan cultures." (2020); Walther, Samantha, and Andrew McCoy. "US extremism on Telegram." Perspectives on Terrorism 15.2 (2021): 100-124.

[21] Op. cit. YouGov.

[22] List of organisations proscribed by the UK government can be accessed at:
https://www.gov.uk/government/publications/proscribed-terror-groups-or-organisations--2

[23] Ibid.; United Nations Security Council Consolidated List, accessed at:
https://www.un.org/securitycouncil/content/un-sc-consolidated-list

[24] "Memetic Irony and the Promotion of Violence Within Chan Cultures", ICSR (2021).
https://icsr.info/2021/01/05/memetic-irony-and-the-promotion-of-violence-within-chan-cultures/

[25] Nagle, Angela. Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right. John Hunt Publishing, 2017; Crawford, Blyth, Florence Keen, and Guillermo Suarez-Tangil. "Memetic irony and the promotion of violence within chan cultures." (2020).

[26] American Jewish Committee, Translate Hate: 'Globalist', https://www.ajc.org/translatehate/globalist

[27] Paige Williams, "Kyle Rittenhouse, American Vigilante", The New Yorker, 28 June 2021. https://www.newyorker.com/magazine/2021/07/05/kyle-rittenhouse-american-vigilante

[28] Hoffman, Bruce, Jacob Ware, and Ezra Shapiro. "Assessing the threat of incel violence." Studies in Conflict & Terrorism 43.7 (2020): 565-587; Mamié, Robin, Manoel Horta Ribeiro, and Robert West. "Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube." 13th ACM Web Science Conference 2021. 2021.

[29] Gallagher, Aoife and Ciaran O'Connor, "The Great Reset," ISD, https://www.isdglobal.org/explainers/the-great-reset/

[30] Accessed at: https://telegram.org/tos/terms-of-service-for-telegram

[31] Guhl, Jakob, and Jacob Davey. "A safe space to hate: White supremacist mobilisation on Telegram." Institute for Strategic Dialogue 26 (2020); Hope Not Hate. "Terrorgram Network: A Spiral Towards Bloodshed—State of Hate 2020." HOPE Not Hate (2020).

[32] Guhl, Jakob, and Jacob Davey. "A safe space to hate: White supremacist mobilisation on Telegram." Institute for Strategic Dialogue (2020); Telegram, 'Terms of Service', https://telegram.org/tos.

[33] Accessed at: https://www.un.org/french/sc/committees/consolidated.htm

[34] Frenett, Ross, "Who Supports Dissident Irish Republicanism? A Snapshot Of Sympathisers On Facebook In 2015", VOX-Pol (2019). https://voxpol.eu/who-supports-dissident-irish-republicanism-a-snapshot-of-sympathisers-on-facebook-in-2015/

[35] Op. cit. "Between Conspiracy and Extremism", ISD.

[36] Comerford, Milo and Sasha Havlicek, "Mainstreamed Extremism and the Future of Prevention", ISD (2021). https://www.isdglobal.org/wp-content/uploads/2021/10/ISD-Mainstreamed-extremism-and-the-future-of-prevention-3.pdf

[37] "BBC documentary 'provokes furious response from Sikhs", Times of India, 18 January 2010, https://timesofindia.indiatimes.com/world/uk/BBC-documentary-provokes-furious-response-from-Sikhs/articleshow/5465239.cms

[38] Jhutti-Johal, Jagbir and Sunny Hundal, "The changing nature of activism amongst Sikhs in the UK today", UK Commission for Countering Extremism. https://www.gov.uk/government/publications/the-changing-nature-of-activism-amongst-sikhs-in-the-uk-today

[39] Leidig, Eviane. "Hindutva as a variant of right-wing extremism." Patterns of Prejudice 54.3 (2020): 215-237.

[40] Griswold, Eliza, "The Violent Toll of Hindu Nationalism in India", The New Yorker, 5 March 2019. https://www.newyorker.com/news/on-religion/the-violent-toll-of-hindu-nationalism-in-india

[41] "Hindutva/Hindu Nationalism Explainer", ISD, https://www.isdglobal.org/explainers/hindutva-hindu-nationalism/

[42] "Exposed: The Scale of Transphobia Online", Brandwatch. https://www.brandwatch.com/reports/transphobia/

[43] Op. cit. "Between Conspiracy and Extremism", ISD.

[44] Hart, Mackenzie et al "An Online Environmental Scan of Right-wing Extremism in Canada", ISD (2021). https://www.isdglobal.org/wp-content/uploads/2021/07/ISDs-An-Online-Environmental-Scan-of-Right-wing-Extremism-in-Canada.pdf

[45] Please note that we have slightly re-written messages here to convey the original meaning of the message, but to avoid the identification of individuals sending these posts.

[46] Clark, Roland, "Kalergi Plan: The Undying "White Genocide" Conspiracy Theory", Rantt Media, 2 May 2020. https://rantt.com/the-kalergi-plan-explained

[47] Guhl, Jakob and Jacob Davey, "A Safe Space to Hate: White Supremacist Mobilisation on Telegram", ISD (2020). https://www.isdglobal.org/isd-publications/a-safe-space-to-hate-white-supremacist-mobilisation-on-telegram/

[48] Buntain, Cody, et al. "Cross-platform reactions to the post-january 6 deplatforming." Journal of Quantitative Description: Digital Media 3 (2023).

[49] Bernstein, Michael, et al. "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community." Proceedings of the international AAAI conference on web and social media. Vol. 5. No. 1. 2011.

[50] Nagle, Angela. Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right. John Hunt Publishing, 2017; Wendling, Mike. Alt-right: From 4chan to the White House. Fernwood Publishing, 2018; Tuters, Marc, and Sal Hagen. "(((They))) rule: Memetic antagonism and nebulous othering on 4chan." New media & society 22.12 (2020): 2218-2237.

51 "Hate Fuel: the hidden online world fueling far right terror", Community Security Trust (2020). https://cst.org.uk/news/blog/2020/06/11/hate-fuel-the-hidden-online-world-fuelling-far-right-terror

52 Leidig, Eviane, "Odysee: The New YouTube for the Far-Right", GNET, 17 February 2021, https://gnet-research.org/2021/02/17/odysee-the-new-youtube-for-the-far-right/

53 Scott, Mark and Tina Nguyen, "Jihadists flood pro-Trump social network with propaganda", Politico, 2 August 2021. https://www.politico.com/news/2021/08/02/trump-gettr-social-media-isis-502078

54 Thomas, Elise, "The Extreme Right on D Live", ISD (2021) https://www.isdglobal.org/isd-publications/gaming-and-extremism-the-extreme-right-on-dlive/

55 Kaplan, Alex, "The growing links between Gab CEO Andrew Torba and Holocaust denier Nick Fuentes", Media Matters, 18 February 2022. https://www.mediamatters.org/gab/growing-links-between-gab-ceo-andrew-torba-and-holocaust-denier-nick-fuentes

56 Zakrzewski, Cat "YouTube alternative Rumble highlights conservatives' move to more hands-off social networks" Washington Post, 16 November 2020. https://www.washingtonpost.com/politics/2020/11/16/technology-202-youtube-alternative-rumble-highlights-conservatives-move-more-hands-off-social-networks/

57 A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if— (a) he intends thereby to stir up racial hatred, or (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby.

58 A person is guilty of an offence if, with intent to cause a person harassment, alarm or distress, he— (a) uses threatening, abusive or insulting words or behaviour, or disorderly behaviour, or (b) displays any writing, sign or other visible representation which is threatening, abusive or insulting, thereby causing that or another person harassment, alarm or distress.

59 A person who uses threatening words or behaviour, or displays any written material which is threatening, is guilty of an offence if he intends thereby to stir up religious hatred.

60 The Crime and Court Acts removed the word "insulting" from the Public Order Act.

61 J. M. Berger, Extremism, MIT Press, 2018

62 Mudde, Cas, The Ideology of the Extreme Right, (Oxford University Press, 2000).

63 https://www.icct.nl/publication/understanding-conspiracist-radicalisation-qanons-mobilisation-violence

64 Accessed at: https://arxiv.org/abs/2010.12472

65 Accessed at: https://arxiv.org/pdf/2004.06465.pdf

66 Accessed at: https://arxiv.org/abs/2012.10289

67 Accessed at: https://github.com/unitaryai/detoxify

68 Accessed at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

69 Accessed at: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

70 Accessed at: https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

71 These models are produced by the same team as hate-alert-counter. An explanatory paper can be found online here https://arxiv.org/pdf/1808.04409.pdf, as well as their Github https://github.com/hate-alert/Countering_Hate_Speech_ICWSM2019

72 Accessed at: https://arxiv.org/pdf/1812.06700.pdf

73 Accessed at: https://arxiv.org/pdf/1703.04009.pdf

74 Accessed at: https://rewire.online/

75 Accessed at: https://github.com/t-davidson/hate-speech-and-offensive-language

76 Accessed at: https://arxiv.org/abs/1909.10966

77 Accessed at: https://arxiv.org/abs/1812.01693

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

Copyright © Institute for Strategic Dialogue (2023). Institute for Strategic Dialogue (ISD) is a company limited by guarantee, registered office address PO Box 75769, London, SW1P 9ER. ISD is registered in England with company registration number 06581421 and registered charity number 1141069. All Rights Reserved.

www.isdglobal.org