ISD | Powering solutions to extremism and polarisation

Hate of the Nation

A Landscape Mapping of Observable, Plausibly Hateful Speech on Social Media

Jacob Davey, Carl Miller, Jakob Guhl

WARNING: This report contains descriptions of posts that readers may find upsetting. These include samples of hate speech containing hateful and dehumanizing language targeting women and minority communities.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Overview

This report provides an overview of public English-language messages collected from Facebook, Instagram, Twitter, Reddit and 4chan across the month of August 2022 which we class as 'plausibly hateful'. This is where at least one of the reasonable interpretations of the message is that it 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.

Executive Summary

Researching hate on social media is one of the most important but difficult kinds of online research to do. On the one hand, it is a phenomenon that is vital to understand in order to fully describe the nature of online spaces and the experiences of different communities that live within them. On the other hand - as this report discusses - accurately identifying hate speech across a range of platforms in a sensible, honest and robust way is a formidable research challenge, both definitionally and technologically.

This report is the result of a research project aimed at identifying hate speech on Facebook, Instagram, Reddit, Twitter and 4chan's /pol/ board across a month, and also identifying the various data, methodological and epistemic considerations associated with this research practice. There is far too much activity across social media to ever be amenable for comprehensive human analysis, and at the core of this research effort was the training, deployment and evaluation of a natural language processing (NLP) apparatus to detect hate speech algorithmically. The automated classification of hate speech has been the object of both academic and commercial interest for a number of years now, and to build on the progress made by other groups, we combined many hate classification models together into a model of models, or an 'ensemble'. The strengths and limitations of this approach are also discussed below, and it is essential that the findings presented in this report are read with these caveats in mind.

Whilst the research covers a number of platforms, they are very different from each other, and inter-platform comparisons should not be drawn from this study. Twitter, 4Chan, Instagram, Facebook and Reddit each differ in their size, who uses them and how they fit into peoples' lives. They are different too in whether they have policies on hate speech, what those policies are, and how they are enforced. Perhaps most importantly, the volumes of hate identified in our research for each platform is greatly influenced by the scales of data each platform makes discoverable and available for collection, the differing interaction of our data collection criteria with each platform, the differing recall performance of our system on each platform, and the take-down activity of each platform which are guided by platform-specific community guidelines.

Hate speech is highly contextual. Often, it is not possible even for human analysts to determine whether a particular post is in fact hateful when removed from its context. During the analysis for this report, analysts trying to assess if a post was hateful or not often lacked crucial information about the identity of the sender and recipient of a post, or the broader context in which it was made. As hateful terms are often reappropriated and reclaimed by their target groups, it is therefore difficult to confidently determine the intent behind the use of such slurs. At the same time, hateful sentiments can also be communicated in a more ambiguous, implicit and subtle manner.

Because of these caveats, both human coders and machines struggle with edge cases where there is uncertainty around whether a post is hateful or not. To address this challenge, we introduce the category 'plausibly hateful' to describe posts for which multiple different interpretations existed and one reasonable interpretation was that it was indeed hateful. These posts were coded as plausibly hateful and will be referred to as such throughout this report.

The research provides a window into plausible hate speech on social media is therefore not a representative one, nor one free from the limitations attendant and inherent to the methodologies that the research uses. We believe however that it remains vital, important and relevant as society continues to debate how to build digital environments which are also tolerant and diverse.

Key findings

Over August 2022, the month of our study:

- We collected 3,140,324 public messages between 01 August 2022 and 31 August 2022 sent on 4chan, Facebook, Instagram, Reddit and Twitter that contained at least one of 334 keywords or key phrases associated with hate speech that we identified.

- Of these, 422,681 messages were classified as 'plausibly hateful', where at least one reasonable interpretation of its meaning was that that it sought to dehumanise, demonise, express contempt or disgust for, exclude, harass, threaten, or incite violence against an individual or community based

on a protected characteristic. For the purposes of this research, we defined protected characteristics as race, national origin, disability, religious affiliation, sexual orientation, sex, or gender identity.

Across platforms, we identified:

- **394,753** plausibly hateful messages on Twitter.
- **26,085** plausibly hateful messages on 4Chan.
- **1,540** plausibly hateful messages on Facebook.
- **162** plausibly hateful messages on Instagram.
- **141** plausibly hateful messages on Reddit.

- **These numbers do not constitute either the full amount of hate on each platform, nor a representative sample from each platform.** Our findings should be interpreted as at least this many plausible hateful messages were present on these platforms over August 2022, rather than indicative of absolute counts of hateful content.

- **It is essential that these findings are viewed against the realities of platform size and data access.** The volume of hateful messages on Twitter is a product of that platform providing far greater data access to analysts over the course of this research. Due to discrepancies in data access and a number of other reasons, these findings should not be used to make comparisons around the volume of hate speech on each platform.

- **The nature of hate speech differs depending on the cultural norms of platforms.** A qualitative analysis of randomly sampled plausibly hateful messages suggested that slur terms are used in regular discussion on 4chan, suggesting that some users on the platform have normalised the use of hateful language. The same analysis observed that hate on Facebook, Instagram and Twitter seemed to primarily manifest in the use of slurs, and in more ambiguous language, such as the presentation of conspiracy theories which demonise minority communities.

- **Hate speech does not trigger greater levels of interaction than non-hateful messages.** On Reddit and Twitter, collected non-hateful messages actually achieved more likes per post. The picture was more ambivalent on the Meta-owned platforms, as hateful messages received fewer likes and reactions on average than non-hateful messages on Facebook but were shared and commented on more.

- **Some of the plausibly hateful messages identified were observed to persist on the platforms a month later, whilst others were inaccessible.**
  - On Reddit, 26% of plausibly hateful messages identified were no longer available a month after being collected.
  - On Twitter, 18.2% of plausibly hateful messages identified were no longer available a month after being collected.
  - On Instagram, 14.8% of plausibly hateful messages identified were no longer available a month after being collected.
  - On Facebook, 11.5% of plausibly hateful messages identified were no longer available a month after being collected.

In reading these findings, a number of caveats are important to consider. They are detailed in greater length below, but include:

- **The machine classifier introduces false positives into our results**. This is identified through the evaluation of the model for 'precision'. The precision of the model has been measured as: 91% precision on 4chan; 71% precision on Twitter; and 63% on Facebook and Instagram. The collected Reddit data was manually classified. This means that roughly three in ten of the Twitter messages, four in ten of the Facebook and Instagram messages, and one in ten of the 4chan messages predicted by the model to be 'plausibly hateful' were found to be not, in fact, plausibly hateful on human evaluation. This itself demonstrates the challenges of identifying hate speech – even with a complex methodology and multiple models – and this fact, combined with the disparities in data access discussed above are notable challenges which future studies similarly seeking to analyse hate speech at scale and across platforms may have to contend with.

- **Keyword-based collections do not create representative datasets, and the results here cannot be generalised or extrapolated to provide an estimate of the total amount of plausibly hateful activity on each platform**. Accordingly, it is not possible for us to measure the overall recall of the workflow. The differing levels of data access provided by platforms accordingly introduces a fundamental challenge to any research project attempting to compare the levels of hate speech across social media.

- **Different platforms have different moderation policies, and the definitions of hate speech used in this study do not necessarily represent hate speech as defined by any given platform's Terms of Service or community guidelines**. Therefore, the paper does not claim that the persistence of a plausibly hateful message necessarily constitutes enforcement failure from the platform in question.

Methodology and Approach

Technical glossary

- **Ensemble classifier**: The name of the combination of classifiers forming overall system used in this report for classifying messages as hateful / not hateful
- **Features/signals**: Outputs from the individual models in the ensemble
- **XGBoost**: The classification algorithm used to transform all the models' features into a hateful / not hateful classification
- **Lexicon**: A list of keywords/phrases that can be used to 'look-up' content in a message and identify any matches.
- **Filters**: Methods used for automatically removing noisy messages from the dataset.
- **Zero-shot Classifier**: A classifier that is built without using any training data, only a query and a large language model.
- **Precision**: The proportion of messages that are actually hateful from the messages that were predicted hateful
- **Recall**: The proportion of messages that were predicted hateful from all the actually hateful messages.
- **Transformer model**: A deep learning model that uses self-attention to give a weighting to each part of the input text, which can then be utilized for a specific task.
- **False positives**: Irrelevant messages that we are trying to remove.

Defining hate speech

In nearly all cases internationally, hate speech is differentiated from offensive speech. To maintain strong democracies even speech that is seen as offensive is understood to be necessarily permissible and protected by the right to freedom of expression. However, speech that threatens an individual's rights (such as their right to live free from discrimination) or calls for violence against certain groups is not simply speech that offends, but is liable to cause harm. It is for this reason that hate speech has been characterised as a distinct category of speech.

In the UK there are a number of different laws which define legal thresholds for hate speech, including a range of criminal provisions within the Public Order Act 1986, that relate to the promulgation of racial and religious hatred, and hatred on the basis of sexual orientation.[1] There is also legislation around hate crime, whereby any crime can be prosecuted with hate as an aggravating factor if the offender has either a) demonstrated hostility based on race, religion, disability, sexual orientation or transgender identity, or b) been motivated by hostility based on race, religion, disability, sexual orientation or transgender identity.

Beyond legislation, there are a range of other conceptions of hate speech. Some are proposed by advocacy groups and charities, and still others are established by private companies such as social media platforms. Based on these differing conceptions of hate speech it was important to arrive at a clear definition to be the core of this work. A workshop was held with Ofcom colleagues and a number of these existing definitional frameworks of hate speech were reviewed and considered. As a result, we produced the following definition for hate speech:

> *"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."*[2]

It should be noted that while this definition of hate speech was created based on a review of UK legislation and platform terms of service, the objective of this project was not to identify illegal hate speech. Accordingly, we make no claim as to the illegality of the content identified in this study, nor that it defines violative content on any specific social media platform.

Interpreting hate speech

The meaning of any post on social media often depends on context, and can be opaque and difficult to interpret. In a majority of cases identified in this study, hate was expressed in terms of derogatory slurs targeting people on the basis of their protected characteristics, and the hateful nature of posts was evident.

However, across most categories of hate speech, and on most platforms except for 4chan, we also

encountered posts of a more ambiguous character, and hate sometimes took more subtle forms. 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 (e.g. "we need to stop these migrant rats from flooding our country"), 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.

Other examples of ambiguity exist where implicitly derogatory claims are made about a group which reference conspiracy theories. For example, one interpretation of a post reading "the filthy Rothschilds are a part of this globalist cabal who run the world in the interests of their people" is that the post references established antisemitic tropes which are used to demonise and harass Jewish people.

A final example of ambiguity is in shorter messages using slurs which could also be examples of members of a minority community reclaiming speech. For example, in the identification of hateful speech we found messages such as "this ni**** here" or "what's this p*ki sayin?". In these instances there is a possibility that hateful terms could be used by the original target groups of this hate colloquially, as an example of reclamation of a slur. However, without viewing the content in its original context coders are unable to determine the precise intent behind the use of a slur.

Given the way data for this report was collected, analysts often lack context when determining if a message was hateful or not. Analysts are generally not aware, for instance, of the identity of the sender of the message, whom it is directed towards, or the broader context of the conversation that the message occurs within. The automated classification method we describe below similarly only makes decisions on the linguistic content of the message itself rather than the entire conversation thread from which the message was possibly drawn.

This raises important issues of interpretation. Groups targeted by hatred re-appropriate terms that were originally hateful slurs; counter-speech and appropriated speech can look, linguistically speaking, extremely similar to hate speech; and hateful messages do not always use explicit slurs. For all these reasons, defining a message as either hate speech or not is challenging for human coders, let alone machines, and there will always be edge cases where there is genuine uncertainty around whether or not messages should be classified as hateful or not.

To address this challenge, when coding for hate speech we used a category of speech called 'plausibly hateful'. For documents where multiple different interpretations of its meaning existed, it was coded as hateful where at least one reasonable interpretation fell within the definition of hate defined above. Due to the use of this coding category, we refer throughout this report to "plausibly hateful" speech.

Data collection

Identifying keywords associated with hate

There are a number of different ways that social media data related to hate can be collected. In some cases, work on online hate speech has collected data based on it being created by members of hateful communities that the researchers have identified,[3] that are directed at particular individuals,[4] or which contain a thematically-relevant keyword.[5] Previous work has collected hateful data based on lists of hateful slurs that have been identified by the researchers in conversation with experts, community organisations and victims' charities.[6]

To guide this research we gathered messages against keywords which had been identified as relevant to hate speech. 334 keywords were discovered via the following process:

1. The project began with a seed set of 633,408 messages sent by 768 actors manually identified as hateful in an accompanying research report, entitled Tangled Web: the interconnected online landscape of hate speech, extremism, terrorism, and harmful conspiracy movements in the UK.

2. These 633,408 messages were classified into two classes: plausibly hateful and non-hate, according to the automated workflow described later in this

section of the report, but trained on actor-specific messages.

3. The language contained within messages classed as plausibly hateful were compared to a background corpus of non-hateful messages. Two keyword extraction techniques were used (called YAKE and Surprising Phrase analysis[7]). Words that significantly correlated with the hateful rather than non-hateful messages were considered to be used 'candidate' collection keywords for this report.

4. For each, a trial collection was initiated, and a sample of documents were passed through the hate detection workflow. This allowed two estimates to be made: first, the overall volumes of messages each keyword was likely to return, and second, the ratio of hateful to non-hateful messages it was likely to return.

5. An analyst appraised these metrics, and ultimately made a judgement as to whether each keyword should be used in the data collection for this project.

6. This process was iterated. Additional keyword collection terms were added, more data was collected, further linguistic comparisons were made and candidates identified (stage 3), assessed (stage 4) and either added or removed (stage 4). In this sense, we hoped to create a snowballing discovery process, where key linguistic attributes associated with plausible hate speech could be identified and used.

The total list of words that were used are included in the annex at the end of this report.

Collecting data containing keywords associated with hate

Messages containing any of the 334 keywords identified above were collected in-line with the affordances granted by each platform. These were:

- For Twitter, all public Tweets which contained any hate-relevant keyword identified in step 4, excluding Retweets and Tweets that Twitter identified to be non-English language.
- For 4Chan, all posts which contained any hate-relevant keyword identified in step 4 from a user flagged from the UK from the /pol/ board.
- For Facebook, any top-level post from a public group or page indexed by CrowdTangle which contained a hate-relevant keyword identified in step 4, and which Facebook considered to be in the English language.
- For Instagram, any top-level post from a public account indexed by CrowdTangle which contained a hate-relevant keyword identified in step 4 and which Instagram believed to be in the English language.
- For Reddit, any top-level post which contained a hate-relevant keyword that was identified in step 4. (No language filter offered by the search API).

Automated hate speech detection

A great deal of work has been done over the past decade to try to automate the detection of hate, creating a field of published models and training data. Our approach was therefore to build a model of models - called an ensemble - which leverages the strengths of each model to produce an overall decision of whether any given message constitutes plausible hate speech or not. The ensemble contains 22 pre-trained machine learning models. They have been developed with various aims in mind, including those that detect hateful speech towards a single target group, those that cover multiple target groups, as well as those that aim to detect toxicity, threats, and counter-speech. The full ensemble apparatus involved four stages of processing: initial filtering, target annotation, model labelling and the final hate/not-hate decision. Each is described below.

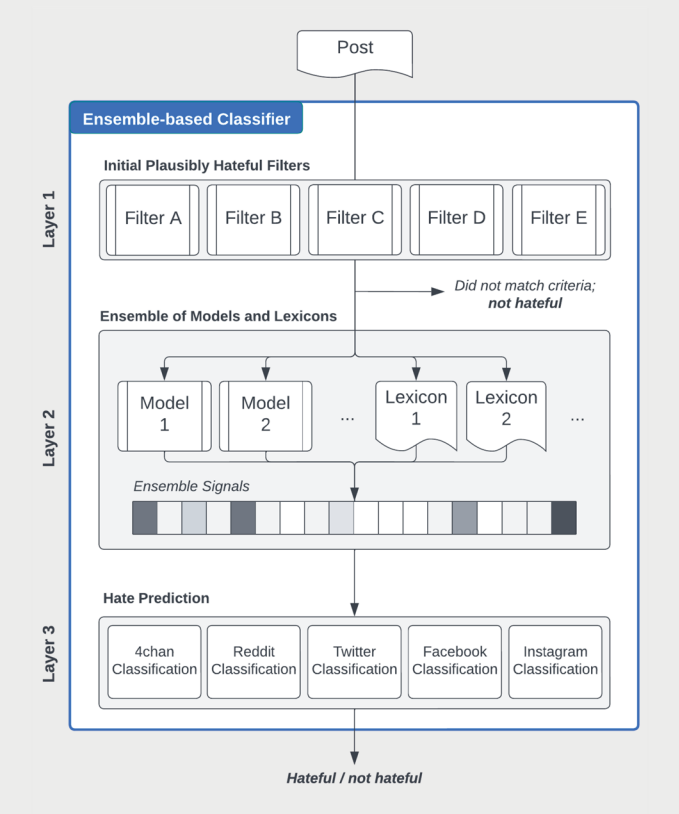

Figure 1. Overview of the layers of annotation in employed in this project

Layer 1. Initial filtering

The data returned contained a substantial proportion of non-hateful material as well as hateful messages. To confront this challenge, our strategy was to create a series of 'high-recall' filters that removed as many of the non-hateful messages as possible, without removing a significant number of hateful messages at the same time. Five filters were used, described below as Filters A, B, C, D and E.

Filter A was a general filter that all documents were passed through. After this, a series of platform specific filters were built, each trying to remove different sorts of platform-specific 'noise' (that is, non-hateful data that was collected) that predominantly appeared on some platforms and not others. Filter B was used on messages from all platforms apart from 4Chan. Filters C, D and E were all used to filter just Facebook and Instagram messages.

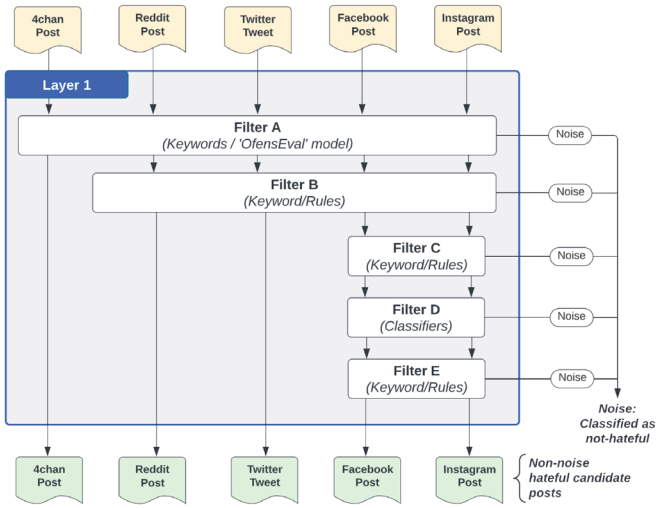

Figure 2. Overview of false positives filtering process across platforms

Filter A

For our first high-recall, cross-platform filter, we removed any document that did not satisfy two criteria. Documents were (a) removed if they did not contain any one of a substantial list of a substantial list of slurs and aggressive terms compiled from a number of sources: 489 identified by ISD researchers, 187 from Hatebegetshate, 178 from Tdavidson, and

81 from websci-19 (see Appendix). Documents were also removed if they were (b) was classified as non-hateful by one of the models identified to be used in the ensemble, called 'Hatebert', when trained on the offenseval dataset (see Appendix). This proved to be highly effective.

Based on manual inspection of 10,019 posts from the keyword-based data that was collected, we established that this filter removed 21% of the data overall, whilst only losing 3% of the hate speech that was in the original dataset.

Filter B
Filter B is a keyword filter consisting of 71 keywords to remove false positives related to pornographic content.

As shown in Figure 2, Filter B is applied to all platforms except 4Chan. There were two prevalent types of non-hateful material identified in the dataset which this filter was designed to remove: pornographic content, and online dating content, both involving the use of sexually aggressive language. The following three examples illustrate this:

- *Your balls belong to me, so let's crush them. findom femdom pay pig goddess humanatm cashcow paypig cuck beta Dom sub cbt*
- *ur disgusting, ur moms disgusting, ur girlfriend/ wife's disgusting findom finD cuck paypig paypiggy humanATM cashcow beta femdom finbrat finsub femsub asiandomme*
- *Hey faggots! It's time to tribute your fuckin master! #cashfag #findom #cashrape #finacialdomination* #cashapp #paypig

Once pornographic and dating material has been removed by filter B, no further filtering was needed for the Twitter and Reddit datasets.

Filter C
Filter C was used on the Instagram and Facebook messages. This is a keyword based filter that removed documents if they contained either 386 high precision keywords or phrases or 2 or more of a list of 30 keywords that alone do not necessarily indicate spam but multiple instances do. There was also a list of 41 keywords that caused documents to be removed when the text also contained the word 'negros'. It was designed to deal with the presence of collection keywords that had different meanings in non-English languages. For example, the word 'paki' means 'please' in Filipino, and the word 'negros' is the name of an island in the Philippines.

There were two reasons why these non-English terms were appearing in our data. First, the API (CrowdTangle) used to collect Facebook and Instagram is not able to reliably limit the data returned to English language data. A second reason was that a certain number of posts include a mix of different languages.

By examining examples of non-hateful material involving these words, we were able to identify terms that when present were indicative of the post being non-hateful, and it is these terms that formed Filter C.

Three examples are as follows:

- "See you!!! U*LSTREET ApparelCa STEALS! 17218 City Airport Baclaran, Parañaque PIPS MY PAT 18+ TWERK BAR presents DOORS 5PM ENTRANCE FEE (P120 W/ FREE BEER) BATTLE ALL STYLES MIC W/ OPEN 8.28.22 MAMI AZZI R MIGGY Baranagu BELLA Hosted MC RHBYN Music DJJULIUS KÌM L. GAB MUSIC TALKWHAT BADBOY PHOBLEY G DRE KALBARYO LOW MRKB JTHAN LA CHAKE CHING CHONG 18K MUSIC DRAFTSMUSICGRASYA MICHHIKO"*
  Removed by the term *"entrance fee"*
- *"#MuruFilm is the powerful and explosive new film from acclaimed filmmaker Tearepa Kahi. Experience it in mind-blowing V-Max NOW at EVENT Cinemas Manukau! Book your tickets here bit.ly/MuruNS CLIFFCURTIS CURTIS JAY RYAN MANU BENNETT TAMEITI SIMONE KESSELL RIA TE UIRA PAKI MURU SERVE OR PROTECT? VIOLENCE EEC OFFENSIVE LANGUAGE +Hg E HE nOD NOW SHOWING EVENT V-MAX BOOK NOW"*
  Removed by the term *"book now"*

Filter D
Filter C was used on the Instagram and Facebook messages. This filter involves the following three Machine Learning classifiers.

1. A zero-shot classifier that removed automotive related material, which appeared in the dataset as a

result of the inclusion of "tranny" in our collection terms. In this context, "tranny" is used to refer to transmission, e.g. "Going in for some tranny work today."

2. A second zero-shot classifier that removed material related to pets, which appeared in the dataset as a result of the inclusion of "coon", "blackie", and "mutt", each of which is used in a non-hateful way in the context of discussion to do with pets. An example of this is "Love my coon Luna very soft and loves her cuddles".

3. A generic spam removing, transformer-based classifier that was trained on a sample of data that passed all of the earlier filters. This classifier was found to remove 80% of the spam. The following illustrate non-hateful material that this filter is intended to remove.
    - *"Search your feelings....you know it to be true: Kimmy's Pick of the Week comes from the Dark Side....of charcuterie boards! These round and paddle-shaped boards are made by a local woodworker and feature wood from the beautiful Purpleheart Amaranth tree. Our round boards also include a handle made from bison hide in order to hang when not in use. Stop in and get one, and may the Force be with you! #CharcuterieBoards #CuttingBoards #GourmetChef #Minot Goy Gou Minot"*
    - *"HAHAHA PAKI SUPPORT GUY'S Fishball pranks umaapoy daw"*
    - *"Download our new Mississippi Green Book App to discover and explore over 100 Black business districts and sites listed in The Negro Motorist Green Book across the state! You can download the app for free on your mobile device or visit onelink.to/msgb The Negro Motorist Green Book special exhibit is open at the Two Mississippi Museums through September 25, presented by the Smithsonian Institution Traveling Exhibition Service and Candacy Taylor."*

Filter E

Filter E was used on the Instagram and Facebook messages. It entailed the following.

1. A regex (pattern matching) filter to remove anything in the form of "[number]:[number]". This filter was to remove the large number of religious quotes that were present in the dataset. Although such material may be depicting acts that may be considered hateful, quoting religious texts itself is not considered hate speech.

2. A filter to remove posts referencing money (currency symbols), which was created to remove posts about ticketed events and marketplace sales. While there is a risk that this could remove some hateful posts in which symbols were used to obfuscate letters in order to avoid moderation, our review of posts containing currency symbols exclusively identified a large volume of posts unrelated to our potential target groups for hate speech. Given the 'needle in a haystack' challenge of identifying hate speech amongst a much larger set of non-hateful posts, we accordingly recognised that whilst this approach may limit the identification of a small number of hateful posts, it was necessary to apply this when balanced against the greater challenge of filtering out non-hateful posts.

3. An additional high precision generic keyword filter that had no overarching theme, it was just composed of terms we established could be used to remove non-hateful content, without removing hate speech. This included terms such as "quick fag", "call or whatsapp" and "doors open at". This consisted of 20 keywords.

Layer 2. Model labelling

All messages that had not been removed by the high-recall filters were then passed through Layer 2. Each message was classified by 22 pre-trained machine learning models and 28 lexicons, which are described in greater detail in the Annex to this report. This resulted in each message being annotated with 90 features from machine learning models. These annotations are used by the platform specific models in layer 3.

Layer 3. Hate/not-hate decision

The final layer of this workflow made an automated decision as to whether a message was plausibly hateful or not. To do so, an XGBoost classifier was trained to use

the annotations given in Layers 2 to make a prediction as to whether or not the post was plausibly hateful. XGBoost is a supervised learning algorithm, therefore requiring a labelled dataset. The training for this classifier occurred in two steps.

Initial Training (Actor-based dataset)
Initially the XGBoost classifier was trained on 6,496 messages collected across Twitter, Facebook, Instagram, Reddit, and 4chan sent by actors manually identified as hateful. 1,091 of these messages were identified as hateful On this dataset the classification accuracy of the model was 0.79 precision and 0.76 recall giving an F1 score of 0.78.[8]

Further training (keywords-based dataset)
Next, the XGBoost classifier was trained on data drawn from the main, keywords-based collection, and specifically from Twitter, Facebook and Instagram. Reddit and 4Chan data were not included at this stage. The Reddit dataset was sufficiently small after filtering (987 posts) that a decision was made to manually determine for each post whether or not it was hateful. Out of the 987 posts that were collected only 141 of them were hateful. The 4Chan data was very similar in nature to the material produced by extremist actors analysed in detail in the accompanying report in this series that was trained to classify hateful material. We therefore used that apparatus to classify the 4Chan data in this dataset. The training data for this classifier contained 488 4Chan posts, 201 of these were hateful.

For the purposes of training and evaluation, datasets for Twitter, Facebook and Instagram were generated by randomly sampling messages from the collected messages for each target group, and then the manual labelling of this sampled data. For Twitter, this resulted in 669 training posts (302 were hateful) and 287 for evaluation (143 were hateful). For Facebook and Instagram, we had 1,526 posts used for training (322 of these were hateful) and 382 for evaluation (79 were hateful). In order to evaluate these classifiers, cross-validation was used to determine the settings of all hyper-parameters (e.g. learning rate). The evaluation data was held back then used to establish model performance.

Evaluating the workflow

Each layer of methodological complexity introduces the potential for bias or inaccuracy, and the use of machine learning in this project was exceedingly complex. Accordingly, accurately understanding the performance of the ensemble model is particularly important.

There are two ways that automated classification of speech can make mistakes: false positives and false negatives. **False positives** are messages which the ensemble would classify as plausibly hateful that would not be considered such by human analysts. **False negatives** are plausibly hateful messages that the system either did not collect in the first place or did not classify as hateful. This error can either be caused not just by how the data was classified, but also how it was collected, and is challenging to measure comprehensively. Reducing false positives boosts what is called the 'precision' of the process, and reducing the number of false negatives boosts what is called the 'recall'.

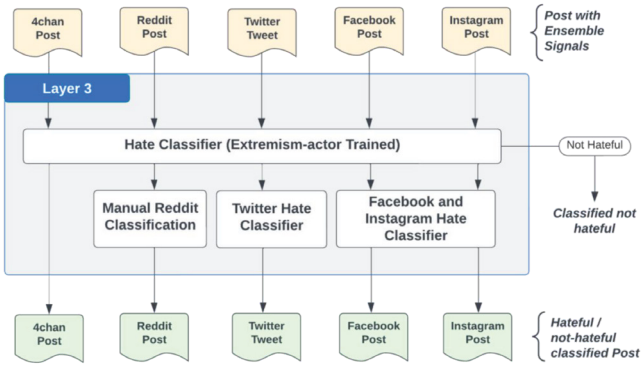

Figure 3: Overview of producing the final hate/not-hate decision for each platform.

Evaluating precision

Separate evaluation datasets were created for this project to assess how well the overall workflow was worked, drawn from each platform. They are not used to train the models, only to evaluate them. They were created through the random sampling of messages from the overall dataset (in the scales mentioned below).

Platform	Potentially hateful labels	Not hateful labels	Total
4chan	280	133	413
Facebook and Instagram	82	800	882
Reddit	NA	NA	NA
Twitter	171	366	537

Table 1. Overview of evaluation data set

Each of these messages were then blind coded[9] by two or more human analysts for whether each was plausibly hateful or not. Then, the output of the workflow was assessed against this human mark-up. Where the algorithmic decision agreed with the human analysts it was deemed to be correct, and where it disagreed it was deemed to be incorrect.

The precision of the overall workflow for each platform was therefore measured as:

- 4chan 91%
- Twitter 71%
- Facebook/Instagram 63%

Due to the low number of posts, the sample from Reddit was coded manually.

Evaluating recall

It must be noted at the outset that the ultimate recall of the workflow is unmeasurable. This is because it is not possible to identify the plausibly hateful messages that did not contain any of the 334 collection keywords this project used, and so were not collected for the project. This important limitation is discussed in greater detail in the caveats section below.

It was however possible to evaluate the recall of the workflow from the data that the project did collect. To do so, data was randomly sampled from the collected data and annotated by human analysts using the same process discussed above. For any message considered plausibly hateful by the human analysts, it was then observed whether the automated workflow had also identified the same message as plausibly hateful.

- For Twitter, 30 out of 150 posts were manually annotated as plausibly hateful. 21 of these were classified as hateful by the ensemble, giving a recall of 0.70.
- For Facebook and Instagram, 200 posts were manually annotated, only 1 of them found to be plausibly hateful, and the ensemble classifier correctly labelled this example, giving a recall of 1.0.
- For Reddit, 200 posts were manually annotated and 7 were found to be plausibly hateful. The ensemble classified 5 of them as hateful, giving a recall of 0.71. The 2 plausibly hateful posts that were missed were filtered out in Level 1.
- For 4Chan 100 posts were annotated, 74 were found to be plausibly hateful, and the hate ensemble correctly identified 70 of these as hateful giving a recall of 0.94.

The above evaluation highlights the difficulty in establishing a reliable measure of recall when the proportion of hateful data is very small. Especially for Facebook, only a single message was considered hateful from the randomised sample of 200.

Measuring persistence of observed hate speech

To measure the degree to which plausibly hateful messages remained present on each platform we performed an additional "live" data collection over a two-week period (from 5th to the 18th of September 2022). On the 27th, we then attempted to collect the same posts again, to see if they were still acquirable from the platform.

Due to differences afforded by each platform's API, three different methods were taken to identify whether a post remained accessible on the platform. For Facebook and Instagram, the CrowdTangle API was used to collect all messages matching the search criteria over the same two-week timeframe. For Twitter, the Tweet IDs were specifically recollected using Twitter's API. For Reddit, the URL to each post was automatically visited and the availability of the post was recorded.

Limitations and Caveats

There are a number of extremely important caveats and limitations associated with this research methodology that should be borne in mind during the reading of our findings.

Applying definitions of hate speech to social media data is challenging

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 in order to develop our shared understandings of hate speech through practical examples. To calculate inter-rater agreement a random sample of 300 posts and comments were independently assessed by two researchers and rated as either 'plausibly hateful' or 'not hateful'. Cohen's kappa was used to measure the proportion of inter-rater agreement over and above chance agreement to determine if the researchers were consistent in their grading.

The analysis demonstrated that the level of agreement between the two researchers was almost perfect,[10] and thus that the rating procedure was highly reliable ($\kappa = .860$; $p < .001$). This would indicate that if other researchers were to duplicate the analysis, they would likely reach the same substantive conclusions based on the rating procedure and available evidence.

Our other response was to use the class of 'plausibly hateful' as the key analytical framework that analysts were asked to decide about. This was intended to help analysts navigate situations where - whether through ambiguous language, lack of context, or unclear meaning - a number of different, legitimate interpretations might be drawn from a given message. This is problematic where some interpretations can lead to the message being understood as hateful, and others not. 'Plausibly hateful' was a useful concept to navigate this and, as discussed above, the inter-annotator agreement suggests it has been applied consistently to the data. However, the approach risks classifying some ambiguous texts as hateful when they are not.

The datasets are not representative of the entire platforms they are drawn from

The data collections for this project could only be carried out on the basis of keywords. This was the only way that data could be collected from the different in-scope platforms in a way that was consistent. It is not possible to measure the recall of these keywords in relation to the total number of plausibly hateful messages on the platform, because the total number is not known. The keyword-based collections created datasets therefore do not represent the total volume of plausibly hateful messages on each platform, nor can they be extrapolated to make this estimate. Accordingly our findings should be interpreted as *at least* this many plausible hateful messages were present on these platforms over August 2022 - other caveats withstanding - rather than indicative of absolute counts of hateful content.

Making meaningful comparisons between platforms is not possible

The results of this study do not allow the prevalence of hate to be compared between platforms. This is for a number of reasons, outlined below.

Data access

First, different platforms made different scales and sorts of data available to researchers:

- **On Twitter**, during the research phase of this project, 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 on other platforms.

- **The Meta-owned tool, CrowdTangle**, provides access to parts of both Facebook and Instagram. 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.

- **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. Searching for top-level posts and comments is possible through the API, however there are limits on how many results are returned per query. While the Reddit API is not explicit about this, is often reported in related documentation to be at most 1,000 items for a given query.

Another crucial factor is exactly what kind of online activity 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 4Chan, both the top-level posts and comments can be collected.

Classifier performance

The analytical workflow used in this project makes mistakes, and does so more frequently on some platforms than others. As discussed above, the measure of 'precision' measures the number of false negatives - messages misclassified as plausibly hateful when they are not. This can mean that more false positives are likely to be included in the results for Facebook and Instagram (precision 63%), and Twitter (71%) than for 4Chan (91%).

Different norms and meanings of language between 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, with hate against Jews involving anti-Jewish conspiracy theories more than specific slurs or attacks.

The machine learning process introduces errors to our results

There are a number of important caveats regarding the performance of the machine learning process designed to detect plausibly hateful messages at scale.

The overall recall of our method not possible to measure

Due to the use of keywords in order to collect data from the platforms studied, it is not possible to ascertain the recall of these keywords to the overall amount of plausibly hateful activity occurring on the platform. This is an important limitation attendant upon all social media research that uses keywords as a collection criteria, and a key caveat of this report.

The recall we can measure is of an unrepresentative dataset

It is possible to measure the recall of the classifier related to the total amount of data gathered for this study. However, given that this dataset is a biased non-random sample, this measure is less intrinsically meaningful.

The process also introduces false positives into our results

The results of the ensemble contain a number of false positives, as discussed in the evaluation section. A number of reasons have been identified for this.

- **Reclaimed language**: The use of words like 'faggot' and 'queer' to self-identify as LGBTQ+ rather than being used as a slur.
- **Other languages being present**: Facebook and Instagram in particular, have a lot of posts that are a mix of languages. These include some words like 'paki', which means 'please' in Filipino language. The use of which the models within the ensemble had not been exposed to before as they have only been trained on the English language
- **Other uses of language**: Some language has multiple uses, which wasn't encountered in the original dataset, for example 'coon' being a Maine coon' or a racoon, leading to language like 'coon hunt' which, unless the other context of the word is known could be considered violent and hateful.
- **Nonsensical posts**: The extremist and terrorist actors covered in the accompanying report in this series, on whose posts the ensemble was originally trained mostly used understandable and well-formed English, but when widening the collection to everyone this is not the case. As this type of often nonsensical use of language had not been encountered before, the initial model was poor at identifying this sort of post as false positives. Nonsensical posts include random strings of words and characters which analysts could not determine the meaning of.
- **Non-hateful aggressive or derogatory speech**: Within the original dataset, there was a limited volume of pornographic messaging, in which people are requesting to be called certain names or derogatory terms as a form of 'roleplay'. This extends to people recounting specific events that have happened to them, where the event itself was hateful, but them talking about it is not.
- **Counter speech**: Within our original dataset there were a small number of people trying to hold people to account for using hateful language or addressing certain events that have happened as unacceptable. This is more common on a dataset collected using keywords and is not hateful, even though very similar language is used.

All of these reasons can produce posts that have (1) similar features (signals) to that of hateful content, or (2) completely new feature distributions or combinations that the classifier had not encountered before, forcing the need to reduce the dataset and to retrain the classifier to learn the domain shifted language (signals/feature space). This was achieved through the layer 1 filters.

Measures of the persistence of plausibly hateful messages should not be confused as a judgment of the enforcement success of any platform

The definition of 'plausibly hateful' does not reflect the different Terms of Service and Community guidelines that each platform maintains regarding hate speech. This means that it is likely that whilst some messages classed as plausibly hateful are likely to be classed as violative content by any given platform, many others will not be. The results of the persistence analysis should therefore not be interpreted as a judgment regarding the success or failure of any given platform's enforcement activity.

The scope of this work did not allow for the empirical comparison of different classification approaches

The intent of this work has been to create a responsible and clear mapping of plausibly hateful activity across the period in question, and also to provide information regarding the possibilities and limitations of this form of research. However, the practical constraints of the project did not allow for the systematic comparison

of different Machine Learning approaches. The aim was therefore to deploy a working classifier, not the best classifier possible. It is therefore very possible that different techniques or methodological choices would have performed better (in terms of precision and measurable recall) than those used and presented in this report.

Findings

The Scales and Venues of Observed Plausibly Hateful Content

We collected 3,140,324 messages between 01 August 2022 and 31 August 2022 from 4chan, Facebook, Instagram, Reddit and Twitter that contained at least one of the 334 keywords or key phrases associated with hate speech that we identified. Of these, 422,681 messages were classified by the ensemble as plausibly hateful.

As discussed in the caveats section above, due to differences in data accessibility, the different volumes of plausibly hateful speech identified on each platform should not be read as indicative of the absolute volumes of plausibly hateful content appearing on each platform over August 2023. Instead, these findings should be interpreted that at least this many plausible hateful messages were present on the platforms over the time period studied, whilst also recognising the presence of false positives in the data.

As a result of these differences in data access, and our inability to make claims around absolute volumes of plausibly hate speech, it is also important that these findings are not treated as comparable. The high volume of plausibly hateful content identified on Twitter is very likely a reflection of the relatively open data access afforded to researchers at the time of conducting the research, whilst the low volumes of plausibly hateful messages identified on Meta platforms are similarly likely a reflection of the restricted data access afforded by the platform, which does not, for example, allow analysts to gather user comments on posts. This would help explain why our dataset contains significantly more messages from Twitter than from Reddit, Facebook and Instagram, despite having a lower number of users than the two Meta platforms (both in the UK and globally).

A notable observation of these findings is the differences between the number of messages containing keywords associated with hate speech, and the number of plausibly hateful messages identified by the ensemble of classifiers used here.

These discrepancies can potentially be explained by a number of underlying factors. One possible explanation relates to the norms of conversation on each platform. On 4chan's /pol/ board, which is reported to be a forum associated with far-right extremists,[11] almost 80% of all messages analysed (26,085 out of 32,903) were classified as plausibly hateful. This may reflect the activity of an online community where hateful behaviours are more normalised and therefore endemic.

On other platforms these discrepancies are likely a reflection of the fact that keywords associated with hate were also used in non-hateful ways. This might be due to them being appropriated or used in counter speech. However, the additional layers of filtering outlined above also point towards the presence of spam, advertising and pornographic content as elements which impede the identification of plausibly hateful content. This finding itself is helpful for future research endeavours seeking to better understand online hate speech.

Given the differences in data availability, we cannot know - and it is not possible to know - the true levels of plausibly hateful activity on each platform. Here it is important to reinforce the observation that providing less data to researchers is not the same as having

Platform	Number of messages collected containing hate-associated keywords	Plausibly hateful messages	Accounts sending hateful messages	Avg. messages per account
4Chan	32,902	26,085	19,098	1.37
Facebook	339,857	1,540	1,137	1.35
Instagram	22,392	162	159	1.02
Reddit	6,829	141	119	1.18
Twitter	2,738,344	394,753	273,645	1.44

Table 2. Breakdown of hateful messages by platform

less hate occurring on a platform, and in many ways this is one of the most crucial learnings to be gleaned from this study, as it has significant implications both for analysts and policy makers. Anyone seeking to understand the scale, nature or persistence of hate speech across popular social media platforms will not be able to construct representative bases of evidence unless data availability improves across the social media industry.

The Severity and Nature of Hate Speech

To assess the severity of the posts identified, a random sample of 350 plausibly hateful messages on 4chan, Facebook and Twitter and the entirety of the plausibly hateful content identified on Reddit and Instagram were qualitatively analysed by a team of ISD analysts. Posts were examined for who was being targeted, and whether the message were directed at individuals or larger groups, contained violent terminology, threatening speech, or calls to action. Please note, this analysis is intended to provide indicative, supplementary and qualitative contributions alongside the quantitative results of this report. The qualitative analysis does not claim to be representative of all of the data that was collected or of the platform overall, but rather as indicative of the variety of hate speech observable across the platforms analysed in this study. It is possible that other researchers may reasonably have reached other conclusions.

Several key trends emerged from the qualitative analysis. First, overall, and across all platforms, hateful activity generally was expressed through the use of derogatory slurs that related to the protected categories included in this project's definition of hate speech.

It is clear that parts of the 4chan community have a distinct and characteristic vocabulary, that includes using specific hateful and racist slang to refer to different communities. In many posts anti-LGBTQ, racist and antisemitic slurs feature in the same post, suggesting a normalisation of hate speech amongst the 4chan community. This trend was also noted on Reddit, albeit in smaller quantities. On 4chan, the observed hateful activity primarily consisted of the use of slurs and expressions of enmity towards protected groups, but it was also more likely to contain violent rhetoric and aggressive speech than on most other platforms. This was also true for Reddit. Qualitative analysis revealed a small number of messages posted on 4chan and Reddit, where hateful discourse targeting Black people, Muslims and Jews was accompanied by violent rhetoric and calls to action (for example a frequently posted message on 4chan which reads "Kill ni*****. Behead ni*****. Roundhouse kick a ni**** into the concrete. Slam dunk a ni**** baby into the trashcan. Crucify filthy blacks.").

Messages were observed where Muslims were described as a disease (e.g "The Muslim surely is a cancer"), and violence against Muslims was not only supported but encouraged as the only way to confront the alleged threat (e.g. "it is OK to kill Muslims"). Antisemitic posts were observed to sometimes depict Jews through the use of conspiracy theories which blamed them for broader societal issues, challenges and grievances (e.g. "The jews own the governments and media in the west. The jews use their power to pass anti-white laws, run anti-white media, censor black crimes and intensely promote miscegenation, LGBT and "diversity." The jews are funding massive illegal migration in USA and Europe. Jews genocided 60 million people, fabricated the holocaust as a distraction, framed Hitler as the bad guy, and then took over the education system to prevent people from finding out").

Messages observed to constitute racist hate often included the use of specific racial slurs referring to non-white communities. Multiple posts in the samples from 4chan and Reddit contained glorifications of violence, or other types of aggressive rhetoric targeting non-white people, immigrants, Jews and Muslims. These messages included calls to assault members of the LGBTQ+ community, fantasies about gassing people or robbing Black drug dealers or calls for self-harm and suicide. While there were examples of such messages in our Facebook, Instagram, and Twitter datasets, these particularly severe forms of hate speech were mostly found in the data collected on 4chan and Reddit.

These findings are not necessarily surprising. 4chan in particular has been frequently associated with controversy throughout its history, and is frequently associated with right wing extremists globally.[12] These more severe manifestations of hate speech should, however, be viewed in the context of broader platform dynamics. 4chan only has 22 million monthly users globally,[13] of which analysis conducted on Similarweb

		Likes/Favorites/ Reddit Score		Retweets/ Shares		Reactions		Comments/ Replies	
		Average	Median	Average	Median	Average	Median	Average	Median
Reddit	Hateful	26	6		N/A		N/A	8	4
	Not Hateful	46	3					20	3
Twitter	Hateful	7	0	1	0		N/A		N/A
	Not Hateful	10	1	1	0				
Facebook	Hateful	30	1	13	0	58	2	12	0
	Not Hateful	40	0	11	0	62	0	9	0
Instagram	Hateful	1157	91		N/A		N/A	77	6
	Not Hateful	937	70					30	3
4chan	Hateful		N/A		N/A		N/A		N/A
	Not Hateful								

Table 3. Average interactions with hateful and non-hateful messages

suggests only 5% (1,100,000) are based in the UK. This stands in contrast to the 44.84 Million Facebook users and 19.05 million Twitter users in the country.[14] 4Chan is a fringe platform, and it is probable that many of its users will be actively choosing to engage with its content, as opposed to platforms like Twitter and Facebook, where the greater global user base, and relatively open functionality mean it is more likely for individuals to inadvertently be exposed to hateful material.

Plausibly hateful messages on Facebook, Instagram and Twitter were more likely to contain less explicitly hateful language and instead manifest in more ambiguous ways, through the referencing of harmful stereotypes, conspiracy theories, and dehumanizing language. On Facebook, the qualitative analysis observed that hate against particular groups was more likely to manifest in the shape of lengthy posts, for example in dehumanizing messages referencing illegal immigrants, or expansive content containing plausibly hateful antisemitic conspiracy theories. However, this finding should again be interpreted through the levels of data access afforded by platforms. The CrowdTangle API only allows access to posts on public pages and groups, as opposed to direct messages, comments under posts, or comments on individual users' pages, and accordingly it is possible that forms of messaging targeting individuals occurs on Facebook but was not available for collection.

Interactions with Hateful Content

To assess wider engagement with hateful content, we compared platform metrics such as likes, shares, reactions and comments on plausibly hateful messages and compared these with other messages collected via the research process described above, but which were classified as not plausibly hateful (referred to below as non-hateful messages).

On Reddit and Twitter, the non-hateful messages we collected achieved more likes per post. On Reddit, plausibly hateful messages also received fewer comments than non-hateful ones. On the Meta-platforms, plausibly hateful messages received fewer likes and reactions on average than non-hateful messages on Facebook but were shared more and commented on more. On Instagram, plausibly hateful messages received more likes and comments on average. It is interesting to note that plausibly hateful messages received greater engagement on Instagram and Facebook compared to Twitter, even though we were able to identify much lower volumes of such messages.

Persistence of Plausibly Hateful Messages

As well as assessing the volumes of plausibly hateful speech which are collectible on platforms, we also

sought to understand how persistent these messages are over time. To achieve this, we sought to measure the degree to which identified plausibly hateful messages remained on each platform.

An additional data collection was performed, beginning on the start of 9th September 2022 and ending at the end of 18th September 2022. Over this time-window, all messages containing the same 334 keywords used in the rest of this report were collected. All messages were then classified as either plausibly hateful or not using the automated workflow described in the methodology section, above. On 27th October 2022, we then sought to recollect all the plausibly hateful messages identified in this time-window. We then reported the number of plausibly hateful messages still collectible (and therefore still present) on each platform.

It should be noted that there are numerous reasons for why a post may no longer be available for collection on a platform. This includes being removed by the platform, by moderators, or the original author of the message. These findings should not be interpreted as an assessment of the efficacy of platform enforcement. It is probable that whilst some of the plausibly hateful content identified in this study will breach platform terms of service, some of it will be permitted by platforms. Indeed, the lack of transparency around how platform terms of service around hate speech are interpreted by moderators in practice is one reason why it would not be possible to accurately engage in an independent research project which sought to solely measure hateful content which transgresses platform terms of service (as interpreted by the platforms) at scale.

The results from this analysis are as follows:

	Overall		
	Total	Inaccessible	%
Facebook	1,083	124	11.45%
Twitter	146,921	26,678	18.16%
Reddit	192	50	26.04%
Instagram	115	17	14.78%

Table 4. Proportion of posts persisting on each platform

Concluding Remarks

This report aims to facilitate a better understanding of plausibly hateful activity on a number of platforms, notwithstanding a range of important limitations and caveats regarding the nature of the data that was collected, and how it was analysed.

Whilst our window into plausibly hateful activity on social media is imperfect, it is also important, providing insight into the differing ways plausibly hateful content manifests online. These insights are important as they help furnish an understanding of what digital spaces are like for the minority communities that use them, and ultimately for the minds and lives that are shaped, in part, by digital life.

This report also foregrounds the importance of data availability for research into online harms to continue. Since this report has been conducted, a number of platforms have restricted data accessibility to mean that similar work in the future becomes very difficult, if not impossible. For everyone who is impacted by, or who works to confront online harms, clear empirical work is vital to drive good policy-making, good decision-making and thoughtful, well-evidenced responses.

As new regulatory paradigms take shape in the UK and globally, these discrepancies in data access afforded to independent researchers will pose significant barriers to accurately assessing the scale and nature of hate speech and other forms of harmful content. If independent, public-interest research into hate speech and other online harms is to continue, then it is essential that solutions are found to these challenges.

Technical Annex

Platform Definitions of hate speech

Platform	Terms of service
Facebook and Instagram[15]	We define hate speech as a direct attack against people — rather than concepts or institutions — on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanising comparisons that have historically been used to attack, intimidate or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics such as occupation, when they're referenced along with a protected characteristic.
Twitter[16]	You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.
Reddit[17]	Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and users that incite violence or that promote hate based on identity or vulnerability will be banned. Marginalized or vulnerable groups include, but are not limited to, groups based on their actual and perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families
4chan[18] (note this research will focus on the /pol/ board)	Global rules: You will not post any of the following outside of /b/: 1. Troll posts 2. Racism 3. Anthropomorphic ("furry") pornography 4. Grotesque ("guro") images 5. Loli/shota pornography 6. Dubs or GET posts, including 'Roll for X' images /pol/ board rules: You are free to speak your mind, but do not attack other users. You may challenge one another, but keep it civil!

Defining hate speech: relevant behaviours and target groups from platform terms of service

Behaviours	Target groups
Dehumanising Speech	Race
Harmful Stereotypes	Ethnicity
Statements of inferiority	National Origin
Expressions of Contempt	Disability
Expressions of Disgust	Religious Affiliation
Expressions of Dismissal	Caste
Calls for Exclusion	Sexual Orientation
Threatening activity	Sex
Abuse	Gender identity
Inciting violence	Serious disease
Inciting hatred	Migration status
Bullying	Pregnancy
Harassment	Victim of serious event
	Veteran status
	Age

Collection keywords

An iterative process was adopted to (1) discover new candidate collection keywords, and (2) sort through candidate keywords to select some to be used for the study as collection keywords. The first step was completed by identifying words that more frequently appeared in plausibly hateful messages than non-hateful messages. The second was performed through a combination of both data-driven and manual appraisal exercises, assessing terms based on the volume of messages they would likely return, and the proportion of returned posts likely to be hateful. Our principle was to continue to add words and phrases as new collection criteria as far as the technical realities allowed.

1. We started with a manually identified initial list of keywords from our dataset of hateful posts written by UK extremist and terrorist accounts identified and analysed in an accompanying research report entitled Tangled Web: the interconnected online landscape of hate speech, extremism, terrorism, and harmful conspiracy movements in the UK.

2. We selected a subset of terms and performed a 1-week historic collection across all platforms. Collected messages were classified as hateful/not hateful by a version of the hate classification ensemble (trained on data from hateful actors only).

3. The language in the plausibly hateful messages and not hateful messages were then compared. For this we applied two complimentary keyword extraction techniques: **YAKE** is an automatic keyword extraction method that is applied to each individual post separately. YAKE utilises various statistical features extracted from each post's text and does not rely on dictionaries or training on any corpora. **Surprising Phrase** analysis is an automatic keyword extraction method that is applied to a collection of messages. Surprising Phrase analysis compares a foreground corpus of text (hateful messages) with a background corpus of text (non-hateful messages) to identify words and phrases that are 'surprisingly common' in the foreground text[19]. A total of 10,000 terms were extracted using YAKE, and 1,000 using surprising phrase analysis. These were the candidate keyterms.

4. All candidate keyterms were then assessed. For each, a trial collection was initiated, and a sample of documents were passed through the hate detection workflow. This allowed two estimates to be made: first, the overall volumes of messages each keyword was likely to return, and second, the ratio of hateful to non-hateful messages it was likely to return.

5. Two analysts appraised this information, and ultimately made a judgement as to whether each keyword should be used in the data collection for this project. In general, analysts attempted to identify keywords that had ratios of hateful messages to not-hateful messages.

6. This process was iterated, such that additional keyword collection terms were added, more data was collected (stage 2), further linguistic comparisons were made and candidates identified (stage 3), assessed (stage 4) and added (stage 5). In this sense, we hoped to create a snowballing discovery process, where key linguistic attributes associated with plausible hate speech could be identified and used.

7. This iteration continued, in order to identify the highest ratio of hate to non-hateful messages possible, and to not exceed the data collection and practical data collection and analysis constraints that the report operated within. In total 334 keywords were identified to be used as data collection keywords.

agender
altjew
aspie
bames
basedjew
basedjudaism
batiman
batty
blaccent
blackie
blackies
blood libel
boseulachi
ching chong
chink
chinks
chuckie
cock carousel
cocksucker
cocksuckers
cohencidence
cohenincidence
commingled
coon
coverage for sex changes
cretin
crossdresser
crossdressers
cuck
cuntboy
cunter
cunts
curry muncher
darkies
deadname
degeneracy
derp
detransition
dhimmi
dhimmitude
dim-witted
dindu nuffin
dindunuffin
downey
dyke
dykes
eeslam
elders of zion
englandistan
eurabia
expose the nose
fag
faggot
faggots
faggy
fags
femboi
femboys
feminazi
feminazis
femoid
femoids
fenians
fleg
flid
foid
gay agenda
gay mafia
gaytard
gender dysphoria
genderism
genderqueer
gentiles
gimmegrants
gimp
gimpy
globohomo
goat fucker
goat fuckers
goat herder
goatfucker
goatfuckers
goatherder
goatherders
gollywog
goy
goyim
goypac
goyz
groid
gulliboyim
haredi
hoe
holohoax
homo
hormone blockers
hypergamy
ilegals
illegal immigrant
illegals
imported paedophiles
incelption
islamification
islamisation
islamism
isra-hell
israelhell
israhell
jew d'etat
jew elite
jew supremacy
jew world order
jew-jitsu
jewed down
jewess
jewhadi
jewish new world order
jewish question
jewishprivilege
jewmerica
jewpuppet
jewry
jewslaughter
jewtube
jihadi's
jihadism
jihadists
jihads
jizya
jq
kaffir
kalergi
kike
kikes
kikesarenotcool
koshernostra
kuffar
kyke
kykes
lavender mafia
lgb
lgbt-lobby
men can't give birth
men in dresses
mestizoes
mgtow
minstrel
misogynist
mizrahi
mohammedan
mohammedans
mong
mongol
mongoloid
mongs
moorish
mordecai
moslem
moslems
moslim
mtf
mulatto
mulattoes
mussie
mussrat
mutt
muttmerica
mutts
muzlims
muzzie
muzzrat
negress
negro
negroes
negroid
negroids
negros
negus
never be a woman
newfag

newfags
nig nog
nig-nog
nig-nogs
nigcels
nigg3r
nigg4h
niggar
niggars
nigger
niggers
niggy
niglet
niglets
no womb you have no
nog
nogs
non whites
normalfag
normalfags
octoroons
only women have a cervix
outbreeding
pajeet
paki
pakis
papery
papism
perversions
pinknews
pisslam
poof
poofta
popery
porch monkey
pozo
protocols of the elders
quadroons
queers
rabbinic judaism
rabbis
rad fem
radfem
rag-head
rag-heads
raghead
ragheads
rapefugee
rapefugees
rapist imports
remember the 6 trillion
retard
rice nigger
roastie
rothschild
rotschild
sambo
sand nigger
sandnigger
sandniggers
schizo
schizos
send them back
send them the fuck back
sexnotgender
shabbosgoy
sheckelgrabb
shekelgrabb
shemale
shit skin
shit skins
should be deported
shylock
sissy
skank
skanks
slag
slags
slut
sluts
slutty
snout
soros shekels
spakka
spastic
spastics
spaz
spear chucker
spic
subhuman
subjewgat
super straight
supercrip
t-girl
tadhg
taig
taqiya
taqiyya
tar babies
tar baby
tard
teague
terfs
the chimp
thegreatreplacement
torah
towelhead
trannies
tranny
trannys
trans activists
trans exclusionary
trans ideology
trans lobby
trans-identified male
transgendered
transgenderism
transgenders
transphobe
transphobic
transsexuals
transvestite
transwoman
troon
troons
twink
tyrone
welfare queen
wetback
whores
window licker
window-licker
witches
wog
wogboys
wogs
women don’t have penises
women have a cervix
women have penises
wop
wormkike
yaxley lennon
yehudi arabia
yid
zio
zioberg
zionazis
zionist occupation government
zionist occupational government
zionist occupied government
zionist world order
ziowood
zog
πiggas

Data acquistion

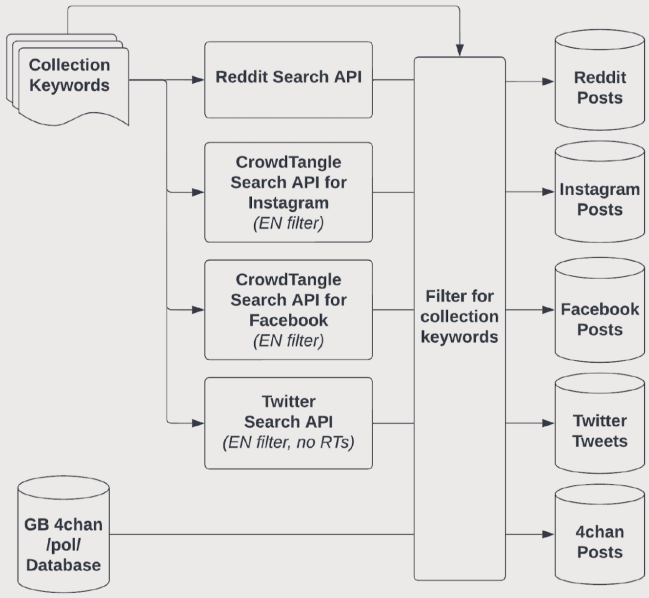

Table 7. Overview of data collection workflow

Due to estimated data volumes, we selected the time window of 1st August 2022 until 1st September 2022 for collection and analysis.

Facebook and Instagram were collected using CrowdTangle's full-platform Search API to obtain top-level posts available from CrowdTangle indexed public pages and groups. The search query was configured to collect any post during 2022-08-01 to 2022-08-31 matching at least one hateful keyword. To reduce the volume of requests, a search filter was applied to only request English language messages, as detected by the platform.

Twitter was collected using Twitter's official full- archive Search API. The search query was configured to collect all Tweets from 1st August 2022 to 31st August 2022 matching at least one hateful keyword, as well as exclude non-original Tweets (Retweets), and to only request English language messages as detected by the platform.

Reddit was collected using the Search API to obtain top-level posts from across the platform. The query was configured to collect all top-level posts from 1st August 2022 to 31st August 2022 matching at least one hateful keyword. The historic data that we are able to obtain through for Reddit is less transparent than data collected for the other platforms. The posts returned by the Reddit Search API results in an apparent subset of posts skewed towards more recent posts.

4chan posts were filtered from an existing collection of 4chan data. This source 4chan data constitutes all messages posted to the 4chan /pol/ board and is collected on an ongoing basis. Due to 4chan only archiving/saving the most recent messages, the platform does not afford itself to measuring takedowns. Due to high volumes of traffic the /pol/ board receives, we estimate that each post is only available for a number of minutes to hours.

After collecting data for each platform, all posts were filtered to ensure the post's text contained the original collection terms. There were a number of reasons data was returned by the platform but was not preserved by this filter.

For Twitter, many Tweets were returned by the platform due to user names being used to identify a match; these did not match our filter and were removed. For Reddit, a large number of posts were returned due to the subreddit name itself being matched (e.g. r/agender). Many Reddit posts did not contain textual content, instead only containing links, images, videos, or other media. These messages were found to be included based on the post's title or link. For Facebook and Instagram, many posts were returned due the keywords being part of hashtags. Such posts that only contained keywords as or within hashtags were filtered out.

Models/lexicons used in the ensemble

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.[20] It determines whether a post is hateful or not. Subset models include **Abuseval** based on the Hatebert approach above, but instead is trained to identify abusive posts. **Offenseval** 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.[21] **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.[22] **Rational2** determines if a post is abusive or not, whilst hate-explain-bert-base-uncased determines if a post is hate, 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.[23] The **Original**,[24] **Unbiased**[25] and **Multilingual**[26] 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 LGBT communities.

Perspective. These are a series of models that can be accessed via an API on the Google Cloud Platform. Originally created to help moderators moderate online conversations, they use finely tuned multi-lingual BERT-based models distilled into single-language Convolutional Neural Networks. These models are then used to evaluate the probability of a comment having an attribute of toxicity. Perspective evaluates the following attributes:

- Toxicity
- Severe toxicity
- Identity attack
- Insult
- Profanity
- Threat

With more experimental models also classifying for:

- Sexually explicit
- Flirtation

It is important to note the probability scores from these models do not correlate to the severity of the toxicity, just the likelihood of the comment being toxic.

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.[27] 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.[28]

Lexicons

In addition to the models described above, messages can also be analysed more simply by whether or not they contain a given word. First, several externally compiled corpora have been identified.

T-davidson. 178 words that are commonly used in hate speech- manually curated. Each has a score of how likely the post is to be hate speech when the phrase is included.[29]

Hatebegets-hate. A list of 187 offensive terms that are used against different groups of people commonly in hate speech posts.[30]

Spread_Hate_Speech_WebSci19. A list of 81 offensive terms commonly present in hate speech.[31]

Across a number of different projects, the ISD team have maintained a series of lists, or 'lexicons', of specific offensive terms and identifiers for particular groups. These are split into slurs and group-specific identifiers.

ISD generated word lists containing words and phrases that were more likely to be associated with the following groups:

- Black people
- Disabled people
- East Asians
- Trans People
- Hindus
- Jewish people
- Muslims
- Non-UK citizens
- Different racial groups
- Protestants and Catholics in Northern Ireland
- Lesbian, Gay and Bisexual people
- Sikhs
- South Asians
- Women

ISD also created lists with slurs likely to be used in hate targeting certain groups:

- Anti-Black slurs
- Anti-disability slurs
- Anti-east Asian slurs
- Homophobic slurs
- Misogynist slurs
- Anti-Muslim slurs
- Antisemitic slurs
- Anti-south Asian slurs
- Racial slurs
- Sectarian slurs
- Transphobic slurs

Model annotations

These detail the full annotations applied by the ensemble:

IMSyPP-inappropriate
IMSyPP-offensive
IMSyPP-violent
abuseval
dehatebert
detoxify-multilingual-identity_attack
detoxify-multilingual-insult
detoxify-multilingual-obscene
detoxify-multilingual-severe_toxicity
detoxify-multilingual-sexual_explicit
detoxify-multilingual-threat
detoxify-multilingual-toxicity
detoxify-original-identity_attack
detoxify-original-insult
detoxify-original-obscene
detoxify-original-severe_toxicity
detoxify-original-threat
detoxify-original-toxicity
detoxify-unbiased-identity_attack
detoxify-unbiased-insult
detoxify-unbiased-obscene
detoxify-unbiased-severe_toxicity
detoxify-unbiased-sexual_explicit
detoxify-unbiased-threat
detoxify-unbiased-toxicity
hate-alert-counter-binary
hate-alert-counter-black
hate-alert-counter-jew
hate-alert-counter-lgbt
hate-alert-counter-multi-Affiliation
hate-alert-counter-multi-Denouncing_speech
hate-alert-counter-multi-Hostile

hate-alert-counter-multi-Humor
hate-alert-counter-multi-Positive_tone
hate-alert-counter-multi-Warning_of_consequences
hate-alert-counter-multi-extra1
hate-alert-counter-multi-extra2
hate-alert-counter-multi-facts
hate-alert-counter-multi-hypocrisy_or_contradictions
hatealert-evalita
hatealert-evalita-active
hatealert-evalita-derailing
hatealert-evalita-discredit
hatealert-evalita-dominance
hatealert-evalita-passive
hatealert-evalita-sexual_harassment
hatealert-evalita-stereotype
hatesonar-hate
hatesonar-offense
hateval
hatexplain-hate
hatexplain-offense
hatexplain-rat-2
offenseval
perspective-IDENTITY_ATTACK
perspective-INSULT
perspective-SEVERE_TOXICITY
perspective-SEXUALLY_EXPLICIT
perspective-THREAT
perspective-TOXICITY
keyword.match/Blacks
keyword.match/Disabled-people
keyword.match/East-asian
keyword.match/Gender-Identity
keyword.match/Hindu
keyword.match/Jews
keyword.match/Muslims
keyword.match/Race
keyword.match/Sectarian
keyword.match/Sexuality
keyword.match/Sikh
keyword.match/South-Asian
keyword.match/Women
keyword.match/antiblack-slur
keyword.match/antidisability-slur
keyword.match/antieastasian-slur
keyword.match/antisemitism-slur
keyword.match/antisouthasian-slur
keyword.match/hatebegetshate
keyword.match/homophobic-slur
keyword.match/misogony-slur
keyword.match/muslim-slur
keyword.match/national-origin
keyword.match/racist-slur
keyword.match/secritatian-slur
keyword.match/tdavidson
keyword.match/transphobic-slur
keyword.match/websci19
roberta-hate
roberta-off

Endnotes

1 The offences in the Public Order Act 1986 relevant to hate speech include broadly the following actions which are intended to or are likely to stir up racial hatred, religious hatred, or hatred on the grounds of sexual orientation : the use of threatening, abusive or insulting words or behaviour, or the display of any written material or the publication or distribution of written material or a recording which is threatening, and in the case of racial hatred, abusive or insulting. It also includes offences relating to fear or provocation of violence, and harassment, alarm or distress which may be prosecuted as a hate crime where racially or religiously aggravated (section 31 of the Crime and Disorder Act 1998).

2 Please note that activity targeting these protected characteristics is considered hateful both when targeted at actual and perceived affiliates of these groups. The protected characteristics covered in this definition are more expansive than those outlined in UK hate crime legislation. These characteristics were considered relevant for this research project based on an analysis of relevant legislation and platform terms of service (as outlined in the Annex of this report).

3 https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0278511#sec003

4 https://www.ofcom.org.uk/__data/assets/pdf_file/0019/242218/2021-22-tracking-twitter-abuse-against-premier-league-players.pdf

5 https://www.nature.com/articles/s41598-021-01487-w

6 https://core.ac.uk/download/pdf/51343449.pdf

7 More information on this process is contained in the technical annex to this report.

8 While YouTube and Telegram were not analysed for this report, as they do not allow keyword-based collection of posts, we trained the classifier on a dataset of material from extremist actors that included data from YouTube and Telegram.

9 To assess inter-coder reliability see section 'Applying definitions of hate speech to social media data is challenging' below for details.

10 Landis and Kock classify a kappa value of .810 – 1.00 as denoting 'almost perfect' inter-rater agreement. https://www.jstor.org/stable/2529310

11 https://onlinelibrary.wiley.com/doi/10.1111/nana.12780

12 https://www.jstor.org/stable/26984798

13 https://expandedramblings.com/index.php/4chan-statistics-facts/

14 https://thesocialshepherd.com/blog/facebook-statistics; https://thesocialshepherd.com/blog/twitter-statistics

15 https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

16 Please note, Twitter's Terms of Service have been adjusted since this analysis was completed. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

17 https://www.redditinc.com/policies/content-policy

18 https://www.4channel.org/rules

19 Robertson, Andrew David, 2019. Characterising semantically coherent classes of text through feature discovery (Doctoral thesis, University of Sussex).

20 https://arxiv.org/abs/2010.12472

21 https://arxiv.org/pdf/2004.06465.pdf

22 https://arxiv.org/abs/2012.10289

23 https://github.com/unitaryai/detoxify

24 https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

25 https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

26 https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

27 https://arxiv.org/pdf/1812.06700.pdf

28 https://arxiv.org/pdf/1703.04009.pdf

29 https://github.com/t-davidson/hate-speech-and-offensive-language

30 https://arxiv.org/abs/1909.10966

31 https://arxiv.org/abs/1812.01693

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org