MAKING SPACE FOR HATE

IDENTIFYING RADICALISERS IN FAR-RIGHT CHAT GROUPS ON TELEGRAM

Patrik Hermansson, Dr James Allen-Robertson, Mustafa Ahmad

September 2023

HOPE not hate uses research, education and public engagement to challenge mistrust and racism, and helps to build communities that are inclusive, celebrate shared identities and are resilient to hate.

We monitor far-right extremism and produce in-depth analysis of the threat of the politics of hate in the UK and abroad.

CONTENTS

PATRIK HERMANSSON

Patrik Hermansson is a senior researcher at HOPE not hate focusing on far-right online culture and organising and is one of the authors of The International Alt-Right: Fascism for the 21st Century? (2020, Routledge Studies in Fascism and the Far Right).

DR. JAMES ALLEN-ROBERTSON

Dr. James Allen-Robertson is a Computational Social Scientist in the Department of Sociology, University of Essex. His research focuses on the ideologies and cultures of technology with his most recent project examining the role of far-right ideology in Silicon Valley. His contribution was supported in part by an 'ISRF Flexible Grants for Small Groups (FG6)' award.

MUSTAFA AHMAD

Mustafa Ahmad is a data analyst at HOPE not hate with a masters in Big Data Analytics from Birmingham City.

EXECUTIVE SUMMARY

Multiple arrests, cases of violence and large amounts of far-right activity has put a spotlight on Telegram as an important space for far-right radicalisation. The report develops a method to identify individuals that are key in the spreading of hateful ideology in far-right Telegram chats. We use a combination of deep learning and analysis of user activity to assess the potential for interactions between users to spread far-right narratives.

Our metric takes into account if a user has expressed any of 13 different far-right topics, interactions with content that express these topics and whether they go on to express it after interacting with the content. The method enables us to surface users who are disproportionately likely to influence other users in terms of taking up far-right ideas.

We analyse over 3 million Telegram messages from 12 public far-right chats and identify two types of impactful users which we call *influencers* and *reinforcers*. Our scoring system orders users on a sliding scale from most to least effective but we identify these groups around the top 1% of users.

Our metric shows that *influencers* have an unusually high probability in posting messages that lead to the take up of new far-right ideas among other users. Other users engage with their message at a much higher rate than the average user. They tend to incorporate more narratives into their messages, tying multiple forms of hate into single frames. We argue that this is an effective way to entrench and expand someone's far-right views.

The impact of *reinforcers* on the other hand comes primarily from the tenacity of their engagement. They tend to send disproportionately many messages containing simpler narratives but the amount of messages, combined with stature in the community also lead to meaningful engagement with their content. Their messages contain fewer topics and they therefore tend to not spread new narratives to other users but repeat narratives already held by many.

This method has direct practical implications. It can enable monitoring organisations to more effectively counter radicalisation by identifying the most effective spreaders of hate, platforms like Telegram to intervene and it helps build knowledge of how influential users communicate to others. This report applies the approach to Telegram but it can generalise to other platforms as well.

KEY FINDINGS

- We identify a method to identify users who are highly influential in spreading far-right ideology in Telegram chats.
- We identify two types of users who have a high likelihood of pushing other users into more extreme ideas, we call them influencers and reinforcers.
- Influencers tend to draw together topics into single narratives, framing for others how different topics can be joined together to form a wider narrative.
- Reinforcers post high volumes frequently, largely reinforcing existing narratives. As prominent chat members, they likely hold significant influence.
- Exposure to hateful content in Telegram chats is correlated with increased odds of a user subsequently expressing that hate themselves.
- We build a deep learning text classifier based on approximately 12,000 manually labelled messages that can classify Telegram messages into 13 different categories with high accuracy (F_1 score: 0.85).

BACKGROUND

Radicalisation online and offline is a growing challenge. Because of Telegram's design and lack of moderation it has become an effective driver of radicalisation in the UK and abroad. This report develops a new method to identify the drivers of hate in far-right Telegram chat groups. We argue that the hard to predict nature of far-right violence, and the inherent damage caused by the entrenchment of far-right ideas, make it important to directly challenge the spread of harmful narratives at their source. The approach uses textual analysis and a scoring method to identify users which have a high likelihood of pushing other users further into the far right.

Telegram has recently been the focus of increasing attention for its use by the far right in the UK as well as internationally. Its rise in popularity can be attributed to a combination of the platform's low moderation and the perception that other mainstream platforms have become increasingly hostile to the far right as well as features on the app itself. Its support for both direct messages and mass broadcasting allows it to work as an outreach and campaigning platform while its closed private chats are well suited to organising.

Telegram also supports large public chat groups which are relatively easy to find and participate in. These chats are often created around a specific far-right organisation or personality or around a specific topic, for example immigration or Covid-19 conspiracy theories.

Telegram has been connected to several cases of planned violence and terrorism convictions in the UK. Recently Ashley Podsiad-Sharp, the leader of White Stag Athletics Club and an associate of nazi terror group National Action was convicted to an eight year prison sentence for having disseminated terrorist material on Telegram.[1] The judge who sentenced Podsiad-Sharp said he had used a Telegram group as "camouflage" to recruit "ignorant and disillusioned men" and incite them to violence.[2] Earlier this year, Luca Benincasa was convicted for his engagement in the Feuerkrieg Division (FKD) which is a proscribed group.[3] Benincasa activism was almost solely on Telegram. The use of Telegram by the far right and its connection to cases of violence justifies more research into the platform and its possible role in stoking hatred and violence.

IDENTIFYING RADICALISERS

Radicalisation into far-right ideas is harmful in multiple ways. While a lot of reporting and research focus on Telegram is related to support for violence and terrorism, radicalisation is not necessarily, or even usually, physically violent. Radicalisation can covertly erode individual lives and wider societal cohesion. Users with radically extreme views may also work to inspire others to their cause, doing the discursive work of radicalising others, which recursively leads to further radicalisation and increased chances of violent action overall. Stochastic terrorism (demonisation of a group in a way that makes violence against them likely) is reliant upon this indirect relationship of communication in which individuals may have particular beliefs reinforced, new beliefs introduced to them or different beliefs integrated into their worldview by a wide array of different actors.[4]

This report is the first step in a new research project by HOPE not hate that aims to develop our understanding of how social dynamics on far-right Telegram chats could help drive radicalisation and the spread of far-right narratives, and improve our understanding of the radicalisation process. We argue that one of the primary dangers of far-right chat groups on Telegram is the way in which they tie multiple forms of hate and conspiracy theories together into a larger structure (coherent or not) which ultimately supports the feeling of threat towards oneself and the in-group, whilst reinforcing these beliefs by providing social connections with like-minded individuals.[5]

In these spaces users engage in what Michael Freeden refers to as ‘decontestation’, a constant reinforcement and/or renegotiation of what an ideology is, and why it is a “correct” explanation for the way the world works.[6] Decontestation allows ideologies to integrate new information and ideas by framing them in relation to the existing ideology and its values. For example, David Lawrence in a report on misogyny and antisemitism shows how the idea of undue jewish influence in world politics is used to explain the perceived loss of status among men in misogynist spaces which enables both men who hold antisemitic views to take on more misogynist views and those who hold misogynist views to also take on antisemitism.[7]

We hypothesise that certain members are more effective than others in the spread of ideas due to factors like their position within the community, level of activity, the way in which they articulate their ideas and what sort of ideas they express in what contexts. We reason that a user that expresses themselves frequently in a chat but in a way that is within the norms, for example expressing non-violent anti-migrant opinions in an anti-migrant chat, might strengthen and help entrench existing views but is unlikely to expand and deepen them significantly among the recipients. Whilst someone who expresses other forms of racism or conspiracy theories which are not held by the majority in that chat, has the potential to more measurably affect other users’ views, inviting the integration of additional narratives into the community world view.

We recognise that radicalisation is also a complex interaction of human relations, personal experiences, social circumstances and psychological factors that spans both online and offline lives.[8] For that reason it is hard to both accurately assess someone's level of radicalisation as well as how they came to take on their views.[9]

Therefore, we focus on the users that make space for hate, leading conversations into a direction where existing ideas can be reinforced, and introducing new forms of hate. Using aggregate data we can surface users who help create an environment which is more likely to facilitate further radicalisation.

RADICALISATION ON TELEGRAM

We focus on Telegram as the design and configuration of the platform supports users to reinforce and extend circulating narratives, what Freeden calls decontestation. Because of their large number of members, in some cases in excess of 10,000, they are spaces where far-right narratives can be introduced, spread and reinforced easily. Even if chats often largely target a specific topic or question, new far-right discourses, conspiracy theories and hate can be introduced and framed as relevant by members.

By focusing on existing far-right and conspiracy theorist chat groups we seek to examine processes that lead to strengthening of existing far-right and adjacent views. Whilst members of these chats likely already hold some far-right views, and could in some cases already be described as "radical", they will vary in terms of the strength of their beliefs, views on specific minority groups and conspiracy theories they believe in. As well as levels of support for violent action.

Already holding some radical views might also make it easier to take on others. Research on conspiracy theory groups have similarly shown that those who believe one conspiracy theory are more likely to integrate others as well, opening up for the possibility of cross pollination between conspiracy theories of different strains and severity.[10] As conspiracy narratives tend to share an overarching theme of threat towards oneself and their in-group, and offer a compelling explanation for events and experiences, even conflicting theories can be tied together into a loosely connected picture.[11]

As such whilst members may hold radical views under one narrative, Telegram chats allow users to be introduced to new ideas, deepening their conviction in their existing narrative, expanding out the range of potential 'enemies' and, from their perspective, increasing the severity of the threat that their in-group faces.

INFLUENCE ON TELEGRAM

This process can take place in multiple ways on Telegram. A user can forward a message from another chat or share a link or a video. Other member users might take part of the message passively or actively engage with it through responding, reacting to it through a set of predefined emojis. Whilst individual messages in broadcasting channels on Telegram often have a view count, this is not the case for group chats whether large or small. We therefore only have the possibility to gauge what messages receive attention or which messages a user consumes if they actively engage within the chat.

Our focus is therefore on engagement through direct replies. This might not be a direct endorsement of the content, it could be an argument against it, but it is reasonable to believe that replying users have more actively engaged with the message than a passive consumer. Replies also act as a proxy for impressions without engagement: whilst there is no way to track the number of views a message has in group chats, the number replies to a message indicate the bare minimum number of people that consumed the content, and we can reasonably presume that the reality was far larger.

The author of the original message is our primary interest. Original posts in chats are the primary way in which discussions are started and we reason that they thereby are the most important way new information is being introduced into a chat on Telegram. While replies naturally also constitute information travelling in the other direction, we argue that it is the initial post that is most likely to introduce new information and frames the conversation that follows. We view this user as the one that sets the scene and opens up space for reinforcing and introducing various forms of hate and far-right narratives that might not otherwise have taken place. It also communicates to other users that the topic is acceptable in the particular chat, which can aid in decontesting it. This can be conscious or unconscious, a message might not directly relate to a hateful discourse but initiate hateful responses.

Our approach seeks to identify messages that introduce responding users to topics they have not themselves expressed and identify users who frequently or disproportionally effectively author these messages.

CONSTRUCTING THE DATASET

Our dataset consists of messages (n=3,103,292) from 12 public far-right Telegram chats scraped between 1 July 2021 and 14 September 2022. The chats were selected based on the criteria that they are well-known within the British far-right scene on Telegram as identified though HOPE not hate's ongoing monitoring work and that they had been active for the period of the collection. We have aimed to include ideological diversity in the dataset in-order to better understand potential differences among different segments of the far-right.

The chat groups included in the dataset are all originating in the UK. Two are primarily anti-immigrant, islamophobic and anti-lgbt+, three are conspiracy theory chats of which one focuses explicitly on Covid-19 and related issues. Four are broad far-right and fascist chats focusing on a variety of issues popular within the far-right, and three others are more extreme fascist chats. Members in the latter three often directly support fascist regimes and historical figures and occasionally express support for violence against minorities.

A core component of our analysis is classification of the messages. We code each message as containing any of 13 commonly expressed far-right narratives and themes using a deep learning classifier trained on a dataset of 11,998 manually labelled Telegram messages. The classifier reaches an F_1 score, a measure of accuracy, across the 13 topics of 0.85. The topics are largely forms of prejudice such as antisemitism, islamophobia, homophobia and misogyny. Additionally we use conspiracy theory as a category indicating expressions of a range of keywords associated with currently circulating conspiracy theories within the far right.

LIST OF TOPICS [IN CONSTRUCTING THE DATASET BOX ALT. SEPARATE]

- Anti-black racism
- Anti-immigrant
- Anti-leftwing and anti-progressive
- Antisemitism
- Conspiracy theory
- Covid-19 Conspiracy Theory
- Far-right slogans and support for far right
- Homophobia
- Transphobia
- Islamophobia
- Misogyny
- Other racism
- Support for violence and violent threats

PROCEDURE

Each message in our dataset is first labelled using our deep learning based classifier. We then create a record for each user indicating which topics they have expressed in their messages alongside timestamps of expressions of each topic.

We then construct a dataset of replies between users. Each interaction is scored based on its potential for introducing a new far-right topic to the responding user. The scoring system takes into account the topics in the original message and the topics that the responding user has themselves expressed prior to the interaction, and will go on to express after the interaction.

Each interaction between a message and a replying user is given a score. If a message exposes a user to a topic they have not previously expressed, the interaction is scored depending on that user's subsequent behaviour, with the highest score being issued when a user re-expresses the topic themselves. If they do not re-express the topic a slightly lower score is given to account for the author's introduction of new information even without subsequent re-expression. Interactions where the replying user has already expressed the topic contained in the original message are scored lowest, recognising that the message may still play a reinforcing role.

Leaving out messages not coded as containing any far-right themes, this leaves us with the following matrix of possible interactions and their associated scores:

	Has not expressed prior	Has expressed prior
Will not express after	2	1
Will express after	4	1

For each interaction a score is given per topic and these are then summed to give the total score for an interaction. This means that interactions will be given a larger score when a message invokes multiple topics.

For example: A user who has previously expressed only misogynist themes interacts with a message that contains misogyny as well as islamophobia. The message has reinforced existing misogyny and so the score for that topic is 1. For the islamophobia topic, as the user had not expressed islamophobia beforehand it is considered a new topic and is given a score of 2. However if that user goes on to express islamophobia after this interaction the topic will be scored as 4 due to the potential that the interaction led to further dissemination. As such the interaction will either be scored as three (1 for reinforcing misogyny + 2 for introducing islamophobia) or as five (1 for reinforcing misogyny +2 for introducing islamophobia with re-expression).

	Score for exposure	Additional score for re-expression	Total per topic for exposure	Total per topic if re-expressed
Misogyny (expressed prior)	1	+0	1	1
Islamophobia (new)	2	+2	2	4
INTERACTION TOTAL			3	5

We then aggregate all inwards interactions per author of the original messages to give each user a score indicating their potential to radicalise others. We explore two ways of aggregating the score per user in the next section.

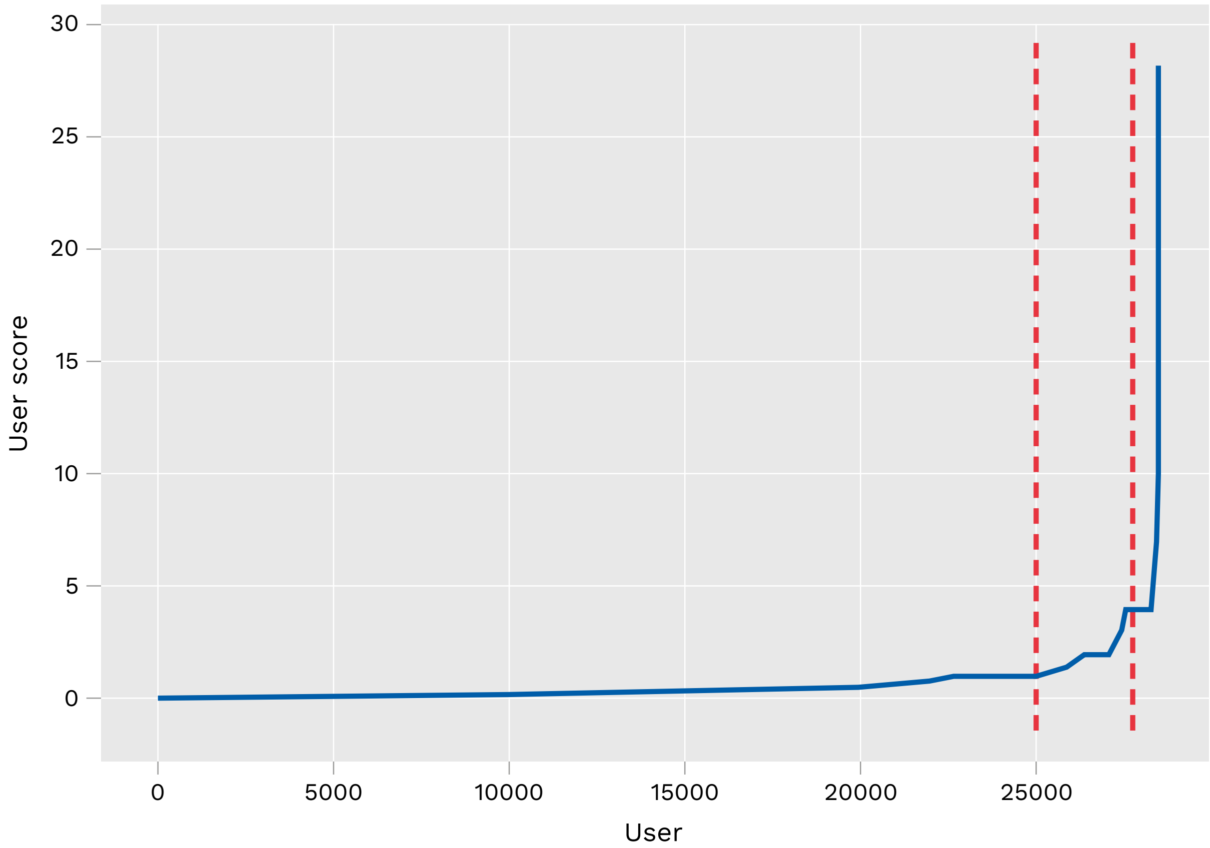
A limitation of our methodology is that we do not separate what could be described as the primary topics of a chat from others. For example, a user engaging with antisemitic content in an antisemitic chat for the first time is scored the same way as if the antisemitic message had been sent in a chat where that kind of content was less common. It is reasonable to believe that chats often are associated with one or several topics and by being inside of them, a user already has some agreement with it. We also do not take time into account, in future revisions we aim to adapt the methodology to weigh the scoring based on the time differential between exposure and expression.

FINDINGS

The resulting dataset can be used to output different types of influential users depending on the way in which the score for the interactions are aggregated, either by summing the score of a user's interactions or by averaging them. Ordering by users' total score emphasises those with a high output, even if their individual messages do not individually have a disproportionately high likelihood of introducing other users to new topics. We call the top users in this category "reinforcers". Alternatively ordering by average score emphasises users whose messages are disproportionately likely to introduce responding users to new far-right topics. We call these users "influencers" because their messages are more likely to introduce other users to new ideas and connect ideas.[12]

To identify the top users in each category we used the Kneedle algorithm to identify the point at which user score significantly increased. By ordering the score in a continuous curve we can observe the point at which users score dramatically changes. This can be interpreted as a significant difference in the interaction behaviour and visibility of users as well as an indicator that the scores of these users is due to a difference in their activity or content rather than pure chance.We view users above this point as the active community, active and engaged with other users in the chats. These are roughly the top 10-12% (~2,500) of users depending on aggregation method. Below this point users have little inward engagement either because of low activity and low engagement with their posts. We then run the same algorithm again on the dataset above the first "knee" and get a third group of the top roughly 1% of the highest scored users, with exponentially higher scores. These are the users of primary interest to us and the basis for what we call the "influencer" and "reinforcer" groups detailed over.

Figure 1. Kneepoints on user scores.



Exposure to hate from another user was correlated with increased odds of expressing that same hate later on. A logistic regression model[13] indicated that the odds of expressing a type of hate were 5 to 6 times higher if that user had been exposed to that same type of hate beforehand. It also indicated that exposure to hate did not fully explain why a user would go on to express it themselves, which is in line with our understanding of radicalisation as a complex process in which these users could play a significant role.

INFLUENCERS

The users with the highest average score per interaction are disproportionately effective in writing messages that spread new ideas to their followers. We find that the top users in this category engage with other users to a larger degree than average. They tend to reply to other users extensively but are also in turn replied to to a similarly high degree. Their response/reply ratio is 0.6, compared to 0.1 for the lower scoring users. Meaning, they are receiving more replies per message posted compared to other users.

The content of their messages singles them out from lower scoring users. Examining the output of these users we find that they have a high amount of original posts and argumentative posts. This is compared to lower scoring users who tend to respond with short, single sentence messages.

The influencer users are also covering a wide range of topics. On average, influencers mention 0.61 topics per message they send in conversations and use 11.8 hate categories in total, spanning much of our list. The rest of the users express just 0.37 of the topics per message on average, half of that of the influencers. This is expected because our scoring system premiers users who draw together multiple narratives but it shows a notable difference in how different users communicate on Telegram. Influencers tend to draw together topics into single narratives, framing for others how different topics can be joined together to form a wider narrative.

We find influencers in all chat groups in our dataset but they are concentrated in the largest groups and the groups which cover a wider set of topics. One might expect users who insert new topics in for example conspiracy theory chats to be a common sort of influencer but this does not seem to be the case. Only a handful of these users are found among the influencers. It is possible that users expressing a wide set of far-right topics find more acceptance in multi issue chat groups, or a part of a larger group's members that are willing to engage with them. This may indicate that particular chat spaces themselves have greater potential for radicalisation due to their less narrow focus.

Counter to our expectations, most of these users' messages are responses. In other words, they tend not to start new conversations but respond to them, often adding new information. Rather than initiating and framing conversations for others, they are possibly instead redirecting and reframing them.

CASE STUDY

One of the highest scoring users in our influencer category engages frequently in three of the chats in our dataset, one focused on anti-lockdown and Covid-19 conspiracy theory and two far-right chats primarily focusing on migration. The user has sent approximately 500 messages which puts them above average in terms of activity but still far away from the top. They almost exclusively reply to existing messages rather than starting new threads.

The user makes strong emotive responses to messages, often expressing strong anger, and weaves in personal stories. Messages often start with "You have said exactly how I feel" or variations therefore before going on to explain a personal story, often vaguely related to the original post. Notably, their posts often stray from the original topic. Messages relating to Covid-19 often lead into discussion about corrupt mainstream media or immigration.

REINFORCERS

The users with the highest total score show expected results. Compared to the *Influencers*, these users post shorter messages and much more frequently. Each individual message contains fewer hateful topics. On average, the reinforcers expressed 0.37 topics per message they sent in conversations which is the same as the average for the whole dataset. Whilst reinforcers totaled many more messages than the influencers with an average of 1,910 messages per user, each interaction with a reinforcer message was on average scored lower than the *influencers*. Whilst they are persistent in their activity, they introduce fewer new topics to the community.

Notably, in this group we find far-right activists well known to HOPE not hate. Several of the highest scorers are engaged far-right organisers. Organisers, and in many cases administrators of the chat themselves, are likely to engage with the community to a large degree while also sending messages that are not widely divergent from presiding norms.

Reinforcers have a lower response/reply ratio than influencers. While they have a large amount of engagement in total, they on average receive less responses per message they send.

Whilst there is lower engagement with their messages overall, the sheer output and engagement with the community likely afford this group of users considerable influence. Perhaps greater than that of the influencer group due to their consistent presence in the chat. A single effective message will not receive as many impressions as many hundreds of messages over a long period of time.

CONCLUSION AND FUTURE RESEARCH

This report has introduced a methodology to assess users' potential for spreading hate and far-right ideology to other users in public chat groups. The method relies on accurate classification of messages into topic categories; however recent advances in the field have made this simpler and more accurate. The result is a practical approach that can be run on large datasets.

Identifying users and chat groups who are disproportionately effective in spreading hateful ideas have theoretical and directly practical implications. It can enable social media and monitoring organisations to more effectively counter radicalisation.

We find that the method can be adapted to surface users with different kinds of influence with the distinction between influencers and reinforcers and that the results are interesting and useful in HOPE not hate's ongoing monitoring work.

The method can also provide further insight into radicalisation on Telegram. The clear clustering of influential users into specific chats is useful as an indicator of not just what users are influential but what chat groups have a higher potential for pushing members towards far-right ideas.

Our methodology has multiple limitations. We do not examine information flowing in the other direction, from reply to the original message poster and we do not examine the role of violent language to any significant degree. We also do not examine in detail what makes a user more effective than another. The influencer group is especially interesting for future research. We believe that more analysis on the content of these users can reveal deeper insight into what kinds of messages and posting behaviour leads to engagement.

In future reports we hope to explore these topics as well as make improvements to our classification method as the analysis should benefit from more nuanced thematic classification, expand to larger datasets and look closer at what kinds of content lead to hateful responses without themselves expressing them.

ENDNOTES

1 https://hopenothate.org.uk/chapter/through-struggle-we-forge-ourselves-into-greater-men/

2 https://www.bbc.co.uk/news/uk-england-south-yorkshire-66671559

3 https://www.bbc.co.uk/news/uk-wales-64404704

4 Amman, M. and Meloy, J.R. (2021) 'Stochastic Terrorism: A Linguistic and Psychological Analysis', Perspectives on Terrorism, 15(5), pp. 2–13.

5 McCauley, C. and Moskalenko, S. (2008) 'Mechanisms of Political Radicalization: Pathways Toward Terrorism', Terrorism and Political Violence, 20(3), pp. 415–433. Available at: https://doi.org/10.1080/09546550802073367; Borum, R. (2011) 'Radicalization into Violent Extremism I: A Review of Social Science Theories', Journal of Strategic Security, 4(4), pp. 7–36.

6 Freeden, M. (1998) Ideologies and political theory: a conceptual approach. 1. issued in paperback. Oxford: Clarendon Press.

7 https://hopenothate.org.uk/2021/09/26/antisemitism-misogyny-overlap-and-interplay/

8 McCauley, C. and Moskalenko, S. (2008) 'Mechanisms of Political Radicalization: Pathways Toward Terrorism', Terrorism and Political Violence, 20(3), pp. 415–433. Available at: https://doi.org/10.1080/09546550802073367; Borum, R. (2011) 'Radicalization into Violent Extremism I: A Review of Social Science Theories', Journal of Strategic Security, 4(4), pp. 7–36.

9 Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist*, 72(3), 278–288; Van der Vegt, I., Kleinberg, B., & Gill, P. (2023). Proceed with caution: On the use of computational linguistics in threat assessment. *Journal of Policing, Intelligence and Counter Terrorism*, 18(2), 231-239.

10 Goertzel, T. (1994) 'Belief in Conspiracy Theories', Political Psychology, 15(4), pp. 731–742. Available at: https://doi.org/10.2307/3791630; Oliver, J.E. and Wood, T.J. (2014) 'Conspiracy Theories and the Paranoid Style(s) of Mass Opinion', American Journal of Political Science, 58(4), pp. 952–966.

11 Douglas, K.M. et al. (2019) 'Understanding Conspiracy Theories', Political Psychology, 40(S1), pp. 3–35. Available at: https://doi.org/10.1111/pops.12568.

12 For the averaging strategy we ignore users with less than 50 replies across the entire dataset as it otherwise risks returning users which have had a very small number of impactful messages.

13 Logistic regression model on 349,518 observations, pseudo r-squared=0.14. P<0.000

HOPE not hate Ltd
Registered office:
Suite 1, 7th Floor, 50 Broadway,
London SW1H 0BL, UK