TACKLING HATE

THE 2024 UK RIOTS:
Tracing the Path from Online Activity to Offline Mobilisation

Matteo Vergani, Andrea Giovannetti, Kewen Liao, Xinzhe Li, Stephanie Ng, Dan Goodhardt

Suggested citation: Vergani, M., Giovannetti, A., Liao, K., Li, X., Ng, S., & Goodhardt, D. (2026). The 2024 UK riots: Tracing the path from online activity to offline mobilisation. Tackling Hate Lab.

This report contains discussions of hateful and distressing language, including explicit examples, as well as themes of violence and discrimination. Some content may be confronting or upsetting for readers.

Please engage with this material at your own pace and seek support if needed.

Contents

3 Executive Summary
5 Introduction
8 From online activity to riots
12 Mobilisation Networks
16 The Hate Contagion
20 Conclusion

EXECUTIVE SUMMARY

This report investigates the role of online activity in shaping and amplifying the 2024 UK riots that occurred between 30 July and 5 August 2024 following a tragic incident in Southport where three young girls were fatally stabbed and ten others injured at a dance class on 29 July. Using geolocated data from X and advanced AI-based language analysis, we assess whether online anti-immigrant and anti-politics narratives preceded unrest, identify key influencers, and examine transnational spread. Our findings offer critical insights into how social media can act as both a mirror and a motor of collective violence.

Key Findings

1. Reposting rates signal the risk of real-world violence

We find that reposting rates (that is, how often each original hate-filled post is shared) spiked prior to major unrest. Unlike raw tweet volume, reposting rates reveal how widely harmful narratives are being amplified. This is a key signal of narrative traction. Rising reposting rates may serve as an early warning for law enforcement and policymakers, indicating when fringe views are gaining mass appeal and offline mobilisation becomes more likely.

2. A small group of users drive most engagement

Our analysis shows that just 100 accounts generated over half of all reactions in the network, with 10 users accounting for 15% of reactions. These users, some linked to Reform UK or other far-right parties, played a central role in shaping online mobilisation. Effective intervention strategies could focus on this narrow band of high-impact influencers rather than attempting broad suppression, making moderation and counter-messaging efforts more targeted and efficient.

3. Coordinated core groups sustain pro-violence messaging

We identified a tightly connected cluster of "hardcore" pro-riot users who not only supported the riots but were deeply embedded within the broader mobilisation network. In contrast, anti-riot users were less connected and more isolated. This structural asymmetry suggests that violent actors coordinate more effectively online. Disrupting these dense core groups—rather than just removing individual posts—could be a more effective strategy for countering digital incitement.

4. Anti-foreigner hate triggered anti-foreigner and anti-politics hate in Australia

Anti-foreigner tweets originating in the UK triggered more online activity, both within the UK and in Australia, than anti-politics tweets. This content crossed borders and influenced discussions abroad, while Australian content had no effect on UK discourse. This finding underscores the global impact of domestic hate narratives and suggests the need for international coordination in tracking cross-border digital threats.

5. Mobilisation tweets are associated with riot locations

Our geographic analysis shows that cities where online mobilisation messages (i.e., explicit calls to action) were posted were also the cities most likely to experience riots. General anti-immigration sentiment alone did not predict unrest. This spatial link highlights that not all online hate is equal: targeted calls to mobilise are a stronger signal of the risk of offline violence. Monitoring and disrupting this content stream may help prevent escalation before it spills into the streets.

6. Implications for Australia

Although the 2024 riots took place in the UK, our findings show that anti-immigrant narratives originating there triggered significant online reactions in Australia, particularly in the form of anti-politics and anti-foreigner discourse. While these digital spillovers did not lead to coordinated offline violence, the transnational transmission of hate narratives underscores a key risk: that domestic unrest in one country can shape the online climate elsewhere, priming audiences for mobilisation. For Australian agencies (at federal, state, and local levels) this highlights the need to strengthen early-warning capabilities for digitally driven threats. Specifically, it suggests value in (1) monitoring reposting rates of hate narratives as a signal of growing traction, (2) identifying high-impact local accounts that amplify imported narratives, and (3) coordinating across jurisdictions to respond rapidly when foreign-origin content begins to trend locally. These findings can inform the development of proportionate, evidence-based responses to emerging digital threats before they escalate into real-world harm. While the UK case offers valuable insights, further research is needed to understand how these dynamics might play out in the Australian context, where political, social, and media environments differ in important ways.

INTRODUCTION

Aims

This report investigates the possible influence of online activity on the 2024 riots in the UK. Drawing on geolocated data from X (formerly known as Twitter), the study has three main objectives:

1. To investigate whether and how online discussions served as early indicators or precursors to the riots.
2. To identify the online networks and influential accounts that contributed to spreading support for the riots and offline mobilisation.
3. To examine how anti-immigration content spread geographically within the UK and across national borders, particularly between the UK and Australia.

Context

Between 30 July and 5 August 2024, the United Kingdom experienced significant social unrest following a tragic incident in Southport, where three young girls were fatally stabbed and ten others injured at a dance class on 29 July.

Although the offender was arrested at the scene, reporting restrictions prevented the public disclosure of his identity as he was a minor at the time of arrest. In this information void, false rumours circulated widely on social media, particularly on X, claiming the attacker was a Muslim asylum seeker. These unverified claims rapidly amplified existing anti-immigrant sentiment and were actively exploited by far-right groups to organise and incite anti-immigration protests across the country.

Riots first broke out in Southport on 30 July near the site of the attack, where a protest outside a mosque escalated into violence. Demonstrators attacked police, set a van on fire, and damaged local businesses. In the days that followed, unrest spread across England and Northern Ireland, with violent incidents reported in major cities including London, Manchester, Liverpool, Sunderland, and Belfast. The disorder included arson, looting, assaults on police and counter-protesters, and attacks on buildings associated with immigrant communities and asylum seekers.

On 1 August, Liverpool Crown Court revealed the offender's identity because of concerns over the lack of confirmed identity fuelling misinformation and riots. The offender was Alex Rudakubana: a Cardiff-born British citizen of Rwandan descent, raised in an Evangelical Christian family. This official communication didn't stop the riots and the anti-immigrant hate online. Most users who supported the riots shifted the narrative from blaming Muslim asylum seekers to blaming immigrants more generally.

By 8 August, 177 people had been sentenced to prison, and by July 2025, over 1,840 arrests and 1,100 charges had been recorded. The riots were largely uncoordinated but driven by far-right networks, including former members of banned extremist groups.

Significance

The 2024 UK riots expose deep fractures in public trust, social cohesion, and the government's capacity at the time to manage mass violence. Sparked by misinformation following the Southport stabbing, the unrest revealed how quickly false claims about the attacker's identity were weaponised by far-right actors to incite anti-Muslim and anti-migrant sentiment. The subsequent violence, including mosque attacks and clashes with police, illustrates how rapidly misinformation can translate online narratives into offline unrest.

Rudakubana was known to authorities and was referred to the anti-extremism Prevent scheme three times. Police and social workers had interacted with him on multiple occasions. He was known to have an obsession with violence. However, he was not enrolled in Prevent due to a lack of evidence of a terrorist ideology. Politically, the case highlighted the limitations of the UK's violence-prevention systems, which remain heavily focused on terrorism with clear ideological drivers, leaving gaps in addressing non-ideological or hybrid forms of mass violence, as evidenced by the Prevent learning review: Axel Muganwa Rudakubana (Home Office, 2025) recommendations.[1]

Understanding how these digital mobilisation dynamics unfold is not only crucial for analysing the UK case but also has direct relevance for Australian agencies tasked with promoting social cohesion, online safety, and preventing hate-driven violence. In this context, community engagement (including the role of trusted leaders, advocates, and support services) is essential both for preventing the escalation of hate and for supporting affected communities in its aftermath.

From a research perspective, the riots provide a unique opportunity to examine the relationship between online language and offline action. X played a central role in amplifying false claims through its algorithms and trending features, making it a critical case study of how platform dynamics can accelerate harmful narratives and facilitate real-world mobilisation.

The review concluded: "The Review identifies several areas for learning to strengthen risk assessments, particularly around understanding indicators of radicalisation where a coherent ideology is not present and recognising the potential risk from repeat referrals. A number of recommendations have been identified through this review. These include strengthening training and guidance, changes to terminology used within Prevent, and improving assurance processes."

UK riots and online activity: What we know and what we don't know

Despite the scale and significance of the 2024 unrest, there has been limited research on the events. Two qualitative reports draw on a small number of illustrative examples of online content (Institute for Strategic Dialogue, 2024; Venkataramakrishnan, 2025). These reports suggest (though do not empirically test) that online activity played a role in shaping the rationale for the riots. Montserrat

One report from the Institute for Strategic Dialogue and CASM Technology (2024) finds that far-right Telegram channels were used to spread anti-Muslim and anti-migrant hate, organise protests, and coordinate offline violence following the Southport attack, with a 327% spike in activity and evidence of posts naming riot locations before unrest occurred. However, the findings are based solely on descriptive data, without statistical testing or causal analysis, which means we cannot determine the strength of the realtionship between online activity and offline violence, or whether one causes the other (Institute for Strategic Dialogue, 2024; Institute for Strategic Dialogue and CASM Technology, 2024; Venkataramakrishnan, 2025). This report, however, provides other important findings about the link between online activity and offline violence.

Data

To perform the analyses, we utilised three datasets:

- **UK dataset:** 67,934 geolocated tweets (19 July–10 August) collected using keyword-based queries on immigration, deportation, anti-refugee sentiment, conspiracy theories (e.g. "great replacement," "white genocide"), anti-immigrant hashtags, and posts linked to events such as the Southport attack and discourse involving UK politicians (e.g. Keir Starmer, Nigel Farage).
- **Australia dataset:** 5,388 geolocated tweets (30 June–31 August) collected with the same keywords as the UK dataset, capturing discourse on immigration, deportation, conspiracy theories, anti-immigrant hashtags, and political debates about UK politics and the riots.
- **Offline incidents dataset:** 85 incidents (29 July–10 August) across 37 towns and cities in England and Northern Ireland, compiled from open-source media, including violent riots, hate crimes, arson, assaults, and large-scale clashes with police.

In our analyses, we focused on detecting two main forms of online discourse: anti-foreigner and anti-politics language.

Anti-foreigner language includes posts that express hate towards people perceived as foreigners. This may target specific ethnic, national, religious, or immigration-related groups (such as asylum seekers, Muslims, or African communities) or include hostile references to foreign countries. We define hate speech as language that attacks, diminishes, or incites hostility or violence against such groups. Importantly, this language may not always be explicit; it can appear in coded forms or be expressed through sarcasm or humour. An example of anti-foreigner tweet is:

"#Southport Why are we importing these low grade savages"

Anti-politics language refers to posts expressing hostility or deep distrust toward political leaders, democratic institutions (such as parliament or elections), or international organisations like the EU or UN. This discourse may include personal attacks, conspiracy theories, or incitement to political violence or disengagement. An example of anti-politics sentiment is:

"Keir Starmer You really are a useless ct"**

Classifier development

To detect online expressions of anti-foreigner and anti-politics language, we developed a set of AI-based classifiers using state-of-the-art large language models (LLMs). We trained and evaluated these models using a rigorous, multi-stage process to ensure accuracy, transparency, and robustness.

We began by asking a sample of 5 annotators from the UK each to manually annotate independently 500 tweets randomly selected from our UK dataset. Annotators were sourced via Prolific, an online platform commonly used in academic research to recruit diverse, pre-screened participants. To ensure reliability, we implemented a screening exercise: each candidate was provided with definitions of anti-foreigner and anti-politics hate and asked to annotate a small sample of 20 tweets. Only those who correctly annotated at least 18 out of 20 tweets, demonstrating a clear understanding of the definitions, were selected for the full annotation task. This was the sole screening criterion used.

These examples provided a foundation for understanding how such sentiments appear in online discourse. We then benchmarked three leading LLMs: RoBERTa (a fine-tuned auto-encoder model), Qwen3 (an open-source generative model), and GPT-4.1 (a proprietary generative model), using both fine-tuning and prompting techniques.

Model performance was evaluated using F1 scores, which combine precision and recall. GPT-4.1 emerged as the most reliable model, achieving F1 scores of 0.78 for anti-foreigner and 0.72 for anti-politics content.

Importantly, before settling on the final classifier, we tested over 100 different configurations across 31 labels and four performance metrics, including tweet volume, volatility, and reposting activity. This comprehensive testing helped us choose a model that was both accurate and generalisable across various types of content.

Overall, this process ensured that the classifiers used in our study could reliably identify shifts in harmful attitudes online. This is crucial for later analyses linking online discourse to offline unrest.

Rationale for focusing on X data

We acknowledge that X (formerly Twitter) is used by approximately 10% of the UK population and around 7% in Australia. In the UK context, this translates to an estimated 6–7 million users. While these numbers are smaller than Facebook's ~40% penetration rate, they are comparable to Instagram, which also maintains around 10% active usage.

Unlike Facebook or Instagram, however, X offers a rare combination of scale, openness, and timeliness: it is one of the few major social media platforms that permits large-scale, real-time data collection, including original posts, engagement activity (e.g., retweets, likes, replies), and crucially, geolocation metadata. The availability of geolocated posts is essential for the spatial analyses presented in this report, including the mapping of mobilisation content and its relationship to offline riot locations: analyses that are not feasible with data from more restrictive platforms such as Instagram or Facebook.

While qualitative studies have effectively used Facebook and Instagram to assess sentiment through public comments, these platforms impose significant barriers to programmatic data access, limiting their use in quantitative, large-N analyses. In contrast, X provides a uniquely powerful dataset for identifying patterns in online hate, influence networks, and digital mobilisation across time and space.

Since Elon Musk's acquisition in 2022 and the subsequent relaxation of moderation policies, X has also emerged as a key venue for observing both mainstream and fringe political discourse. This makes it particularly valuable for tracking the early signals of hate narratives and collective action as they unfold in real time.

In sum, although X does not represent the general population, its open architecture and role as a hub for politicised and often unmoderated discourse make it an indispensable platform for the type of large-scale, empirically grounded analysis conducted in this report. It enables broader generalisability than small-N qualitative studies in platforms like Facebook or Instagram, and offers insights into the dynamics of geospatial digital mobilisation that are otherwise inaccessible through more closed platforms.

FROM ONLINE ACTIVITY TO RIOTS

We analysed three datasets (the UK dataset, the Australia dataset, and a dataset of offline incidents) to examine whether online activity was connected to the UK riots. We used the Australia dataset as a comparison group, based on the idea that if there is a genuine link between online activity and real-world violence, we should only see that connection in the UK data. In Australia, similar types of online discussions occurred, but there were no related incidents of offline violence. This allowed us to test whether the patterns we observed in the UK were unique to that context or part of a broader global trend.

To guide our analyses, we formulated two hypotheses.

- **Hypothesis 1:** Anti-foreigner language would serve as a signal of riot-related violence, particularly given that during 2024 immigrants were the main targets of the unrest.
- **Hypothesis 2:** Anti-politics language would also signal increased risk of violence, as previous work shows that political distrust correlates with violent collective action (Richler, 2023) and similar patterns were observed in our Tackling Hate Lab report on anti-trans hate (Vergani et al., 2025).

Methodological note

To test the relationships between online language and offline events such as riots, we applied time-based analytical models to examine whether shifts in online discourse were linked to, or could help predict, subsequent unrest. These methods allowed us to explore whether online discourse played a role in shaping real-world events, while recognising that such relationships are complex, dynamic, and not necessarily causal in the everyday sense.

Key findings

Firstly, Figure 1 and Figure 2 show that spikes in anti-politics and anti-foreigner language online occurred in the period when offline incidents were most intense. In the figures, the light line represents a moving average of the hourly counts, while the dark line shows the same series after applying a weekly-level detrending procedure. This pattern was present in the UK dataset but not in the Australia dataset, supporting the idea that these online dynamics were uniquely linked to the UK context, where offline violence actually occurred.

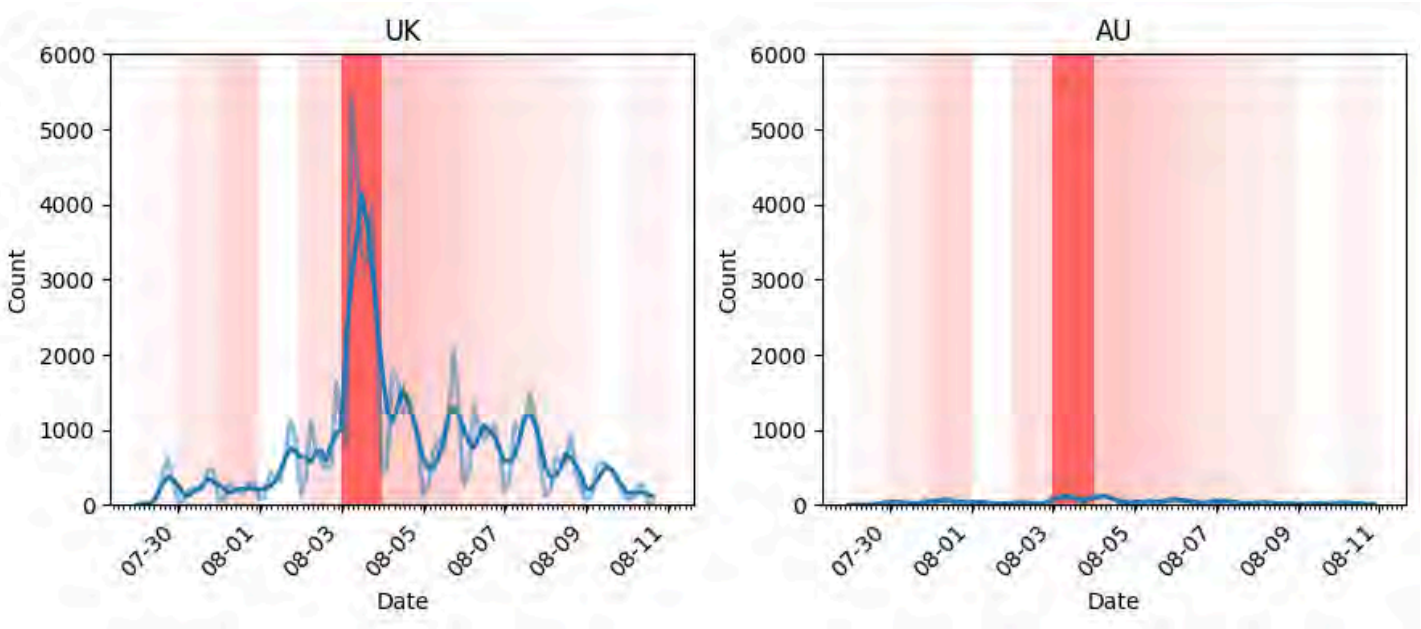


Figure 1. Hourly counts of anti-foreigner tweets in the UK and Australia datasets. Red vertical lines indicate the number of offline incidents recorded each day; darker shades represent days with a higher number of incidents.

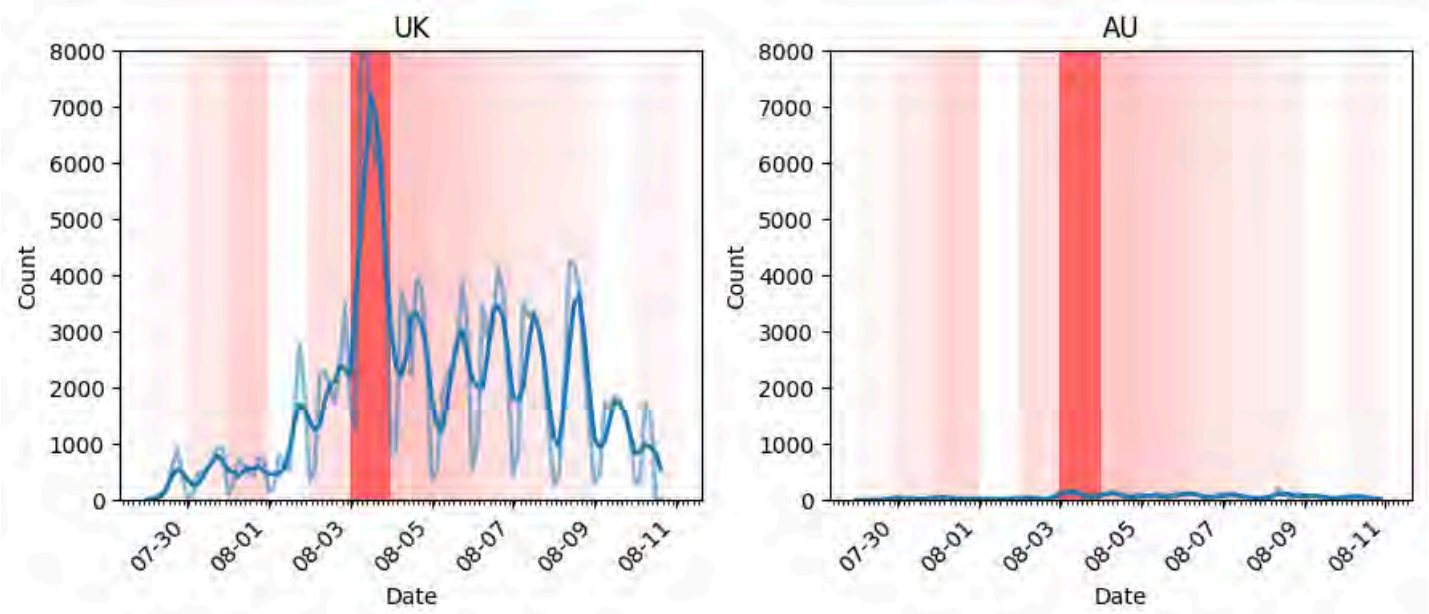


Figure 2. Hourly counts of anti-politics tweets in the UK and Australia datasets. Red vertical lines indicate the number of offline incidents recorded each day; darker shades represent days with a higher number of incidents.

Interestingly, the anti-politics measure shows much greater variation than the anti-foreigner measure during periods of riot activity (Standard deviation: 1733.78 vs 924.33 for the UK signals). This means that the volume of anti-politics tweets fluctuated more dramatically than anti-foreigner tweets when riots were happening. However, this spike in variation coincided with the peak of the riots, rather than preceding them. In other words, the surge in anti-politics content happened at the same time as the violence, not before it, so it does not serve as an early warning. These findings suggest that looking at raw tweet volumes alone, even when large or highly variable, is not sufficient to anticipate unrest in advance.

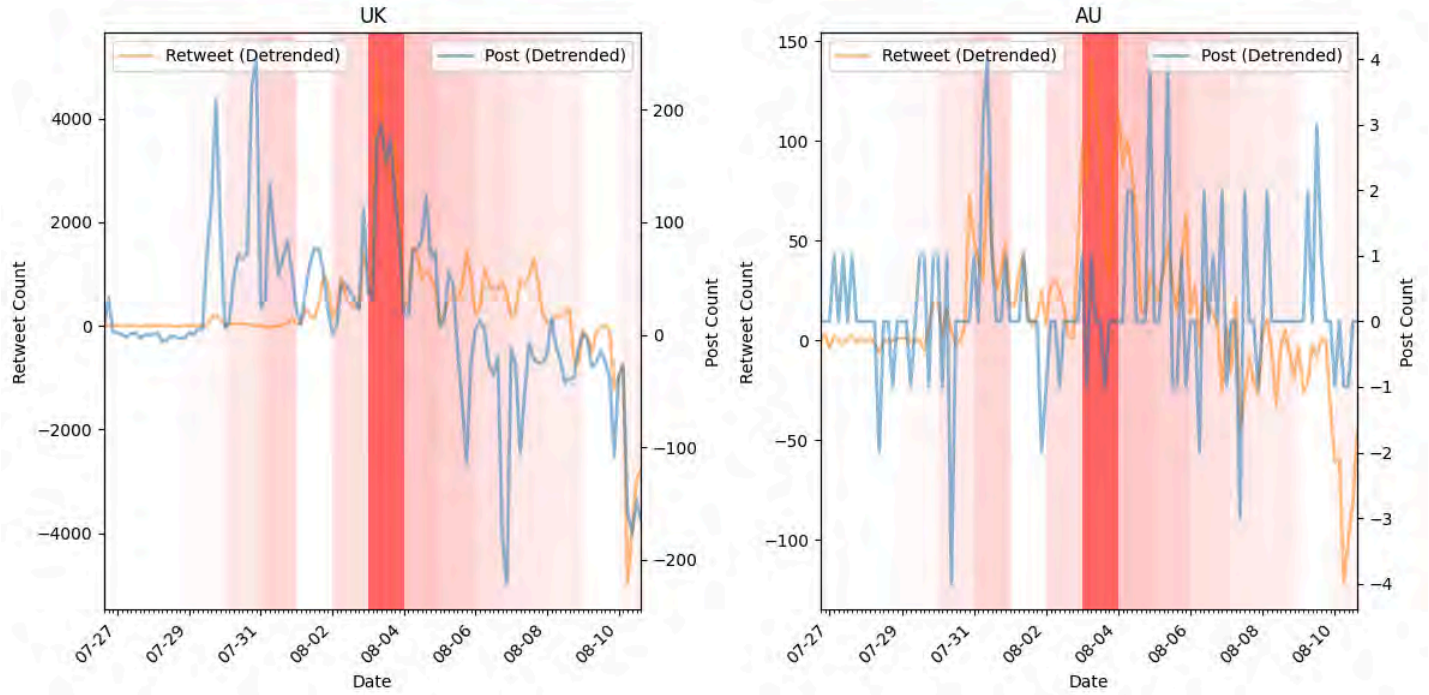


Figure 3. Hourly counts of anti-foreigner tweets in the UK and Australia datasets, separated into original posts and retweets.

Secondly, Figures 3 and 4 show that separating retweets from original posts reveals important differences in online activity patterns. In the UK dataset (but not in the Australian dataset), the volume of original posts containing anti-foreigner and anti-politics language increases significantly before the peak in offline riot incidents. In contrast, retweet volumes tend to rise during the peak period of violence, rather than preceding it. This suggests that original content creation surged in the lead-up to the riots, whereas content amplification occurred more reactively. No comparable surge in such language was observed in the Australian dataset during the same period. Regular fluctuations seen in the figures reflect typical day–night cycles in tweet volumes. Figures 3 and 4 display detrended curves, which help highlight unusual spikes in activity by removing regular weekly patterns from the data. To create these curves, we applied a simple method: for each day and hour in the dataset, we subtracted the value from exactly one week earlier.

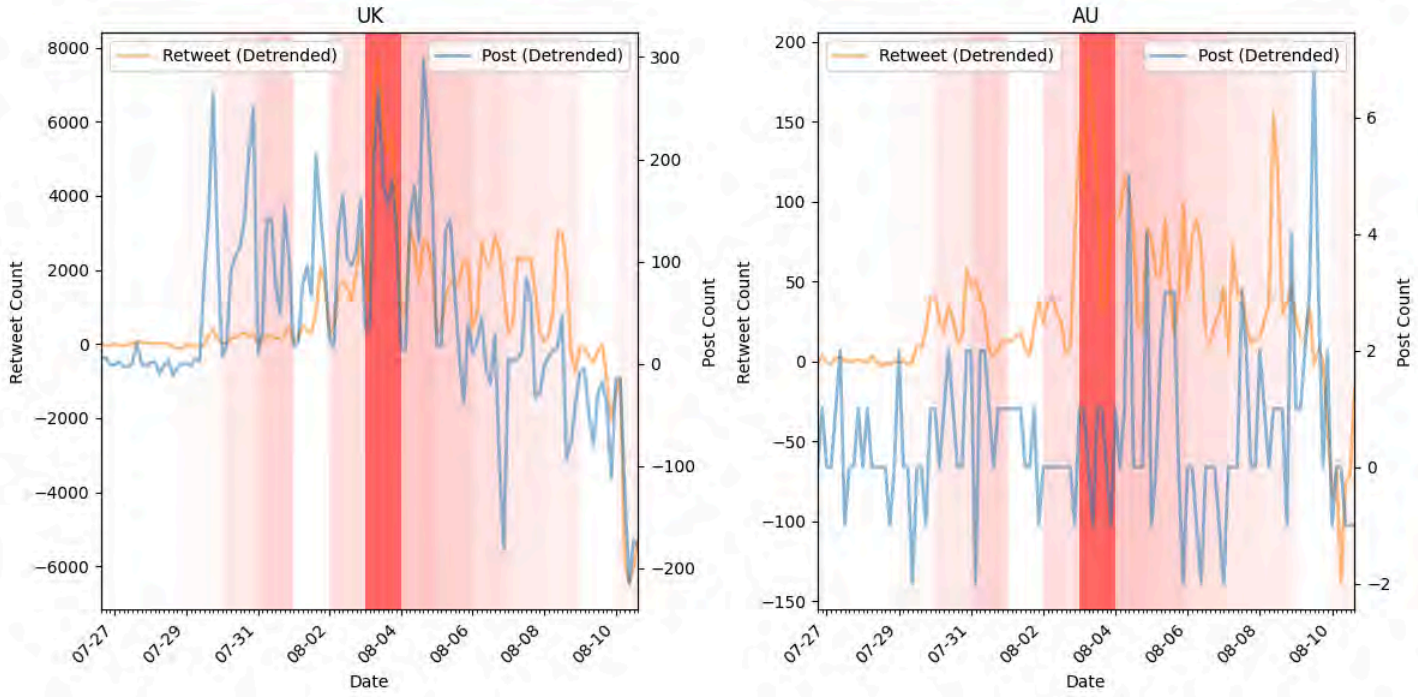


Figure 4. Hourly counts of anti-politics tweets in the UK and Australia datasets, separated into original posts and retweets.

Building on the previous finding, we constructed a measure of reposting rates, defined as the ratio of redistribution actions (including retweets and quotes without added commentary) to the number of original posts. As shown in Figure 5, reposting rates for both anti-foreigner and anti-politics tweets increase noticeably prior to the peak in offline riot incidents. In other words, a rise in reposting rates may serve as an early indicator of potential unrest. While this pattern is suggestive, we emphasise that it reflects correlation rather than causation, and may in part capture heightened online engagement in response to unfolding events. Figure 5 shows that reposting rates—especially for anti-foreigner content—spiked well above baseline levels in the lead-up to the riots.

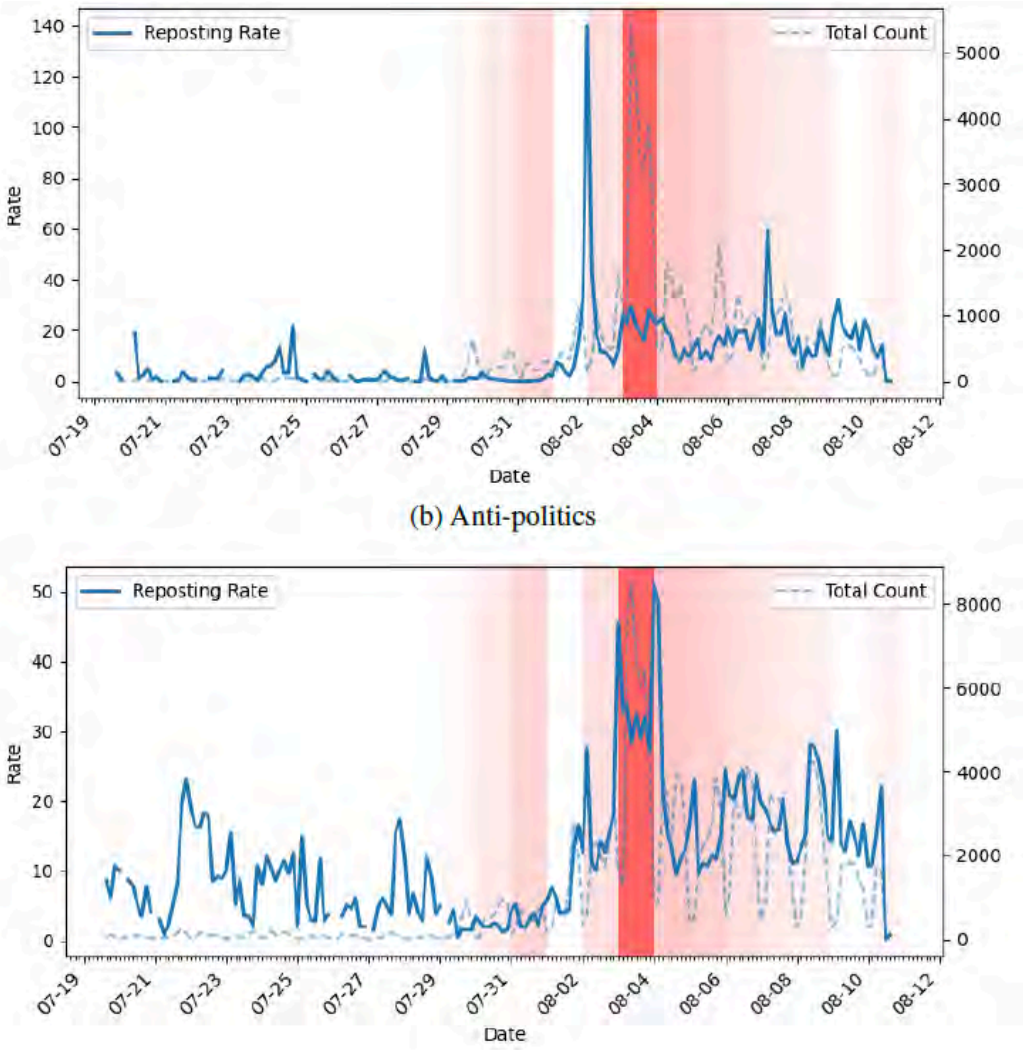


Figure 5. Reposting rates of anti-foreigner and anti-politics tweets.

These patterns can be explained by considering the different behaviours and motivations behind original posting and content amplification. Writing an original post (especially one containing anti-foreigner or anti-politics language) typically requires more cognitive and emotional investment than simply retweeting existing content. It suggests a higher level of personal engagement or mobilisation, which may reflect a readiness to act or an attempt to influence others. The surge in original posts before the peak of offline violence in the UK may therefore signal an intensifying emotional or political climate in which individuals are not only consuming content, but actively contributing to a hostile discourse. In contrast, retweets tend to peak during the riots themselves, when attention and visibility are already high, indicating reactive rather than initiatory behaviour.

Reposting rates, which measure the ratio of retweets and quotes to original posts, offer an especially sensitive signal because they reflect not just volume, but the balance between content creation and amplification. A rising reposting rate means that each original post is being shared more widely, suggesting that certain narratives are gaining traction and spreading rapidly. This form of amplification may help accelerate collective sentiment and synchronise offline mobilisation. Unlike raw retweet counts, reposting rates account for how concentrated the amplification is around particular themes, making them a more informative early warning indicator.

MOBILISATION NETWORKS

We analysed the UK and Australian datasets to identify the online networks and influential accounts that contributed to support the riots in online discussions.

Methodological note

We constructed a cross-sectional influence network to analyse how users interact and contribute to the broader flow of online narratives. This network maps user interactions based on reactions (such as likes, retweets, and replies) to content shared by others. It enables us to assess influence both at the micro level (identifying the influence of individual users) and at the macro level (where we examine the overall structure of the online conversation and the formation of community clusters).

At the micro level, the network approach allows us to determine the position of each user within the wider information ecosystem. This includes measuring how influential a user is in driving engagement and shaping the narrative flow. Influence is operationalised using standard network centrality measures, which help rank users based on how many others interact with their content, and how strategically located they are within the network. At the macro level, the network structure provides insight into how narratives coalesce into distinct communities, and how these communities are connected to one another.

Key findings

Firstly, to understand the positioning of users across narratives within the dataset, we applied a clustering algorithm, agnostic/unbiased on the content of the tweets and the identity of the users, to the full network. The clustering algorithm splits nodes in clusters, that is groups of users primarily interacting with one another. In our network, the clustering algorithm returns hundreds of clusters, with five major clusters collectively grouping more than 85% of nodes. The largest cluster contains 23.1% of the nodes, followed by the second largest at 20%, the third at 18.75%, the fourth at 16.72%, and the fifth at 8.17%. Each cluster has distinctive structural characteristics, for example, in terms of internal cohesion and connectivity to the wider network.

In this section, we present four main descriptive findings from the network analysis.

Finding 1. Influence is concentrated among a small group of users.

More than half (54%) of all reactions in the network (that is, likes, retweets, and replies) come from just the 100 most influential users. Within this subset, the 10 most influential users obtain around 15% of the network reactions, thus accounting for a significant share of influence on the network. The top10 ranking is reported in Table 1. Three of the top 10 accounts were openly associated with the Reform UK party and Brexit, while two others were linked to the Reclaim Party or ADF (Alliance for Democracy and Freedom) Party. The most influential among them is Darren Grimes, a British Reform UK politician, political commentator, and former pro-Brexit campaigner who rose to prominence through the youth-focused group BeLeave and later as a media figure on platforms such as GB News. Although Grimes's relationship with the far right is subject to debate, his position as the most influential account in this study underscores the strong traction his narratives had in the online mobilisation surrounding the Southport attack and the subsequent riots. This pattern holds true across both the UK and Australian datasets, as the analysis was conducted on a combined dataset.

Table 1. Top 10 accounts by content activity (Outdegree): users who shared the most content that reached others in the network.

Accounts	Reform UK / Brexit	Reclaim party / ADF party	Outdegree centrality
@darrengrimes_	1	0	5275
Anonymised	1	0	3752
Anonymised	0	0	3299
@ActionBrexit	1	0	3100
Anonymised	0	1	2814
Anonymised	0	0	2443
Anonymised	0	0	2441
Anonymised	0	1	2207
Anonymised	0	0	2173
Anonymised	0	0	1951

Note: In line with ethical research protocols, we only report identifiable account names for public figures or verified users (e.g., elected officials, political commentators). All other account names have been anonymised.

Finding 2. Influencers have a capability to operate across clusters.

To illustrate this finding, Figure 6 presents Grimes's ego-network (that is, the set of accounts that reacted to his posts, along with any connections that exist between those accounts). The figure highlights the highly unbalanced structure of his influence: no other node in the network approaches Grimes in terms of influence. The accounts in his ego-network divide into two distinct groups. On the right side of the figure, we see a large group of users who do not interact with one another, most of whom belong to a single cluster (represented in blue). On the left side, a smaller but still substantial group includes users from multiple clusters (notably those shown in blue, purple, and orange). This group forms several tightly connected micro-communities, indicating complex patterns of interaction that occur independently of Grimes's direct involvement. While interactions on the right are largely isolated and sporadic, those on the left are denser and more sustained, with higher levels of engagement. This illustrates how a single influential account can affect multiple user communities in different ways, either triggering widespread but shallow reactions or fostering deeper, more interconnected conversations among specific subgroups.

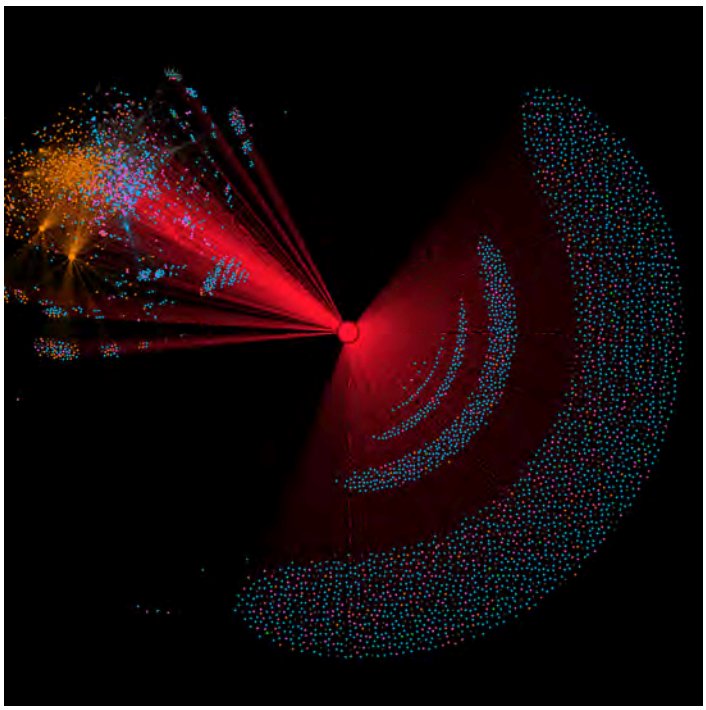


Figure 6. A visual representation of Darren Grimes' influence network. Node size indicates the extent to which a user influences others; node colour represents different clusters.

Finding 3: Identifying a tightly connected group of ‘hardcore’ pro-riot users.

To identify who are the group of users that are more strongly disseminating messages in support of the riots and offline violence, we analysed the structure of the network based on the type of content users shared. Specifically, we tagged posts that either supported or opposed the riots and mapped them in Figure 7 using green links for anti-riot content and red links for pro-riot content. Overall, around 58% of users in the network expressed a clear stance on the riots. Of these, approximately 42% shared pro-riot messages, while 16% posted content opposing the riots.

Our analysis shows a striking difference in how these two groups are distributed within the network. Users expressing anti-riot views are mostly concentrated within a single cluster, suggesting a more limited and contained presence. In contrast, those supporting the riots are spread across several clusters, indicating a broader and more dispersed reach.

One cluster in particular, shown in orange, stands out. The users within this group are much more closely interconnected than those in other parts of the network. This suggests the presence of a tightly-knit subgroup of “hardcore” pro-riot users who are not only active but also strongly embedded within the wider online mobilisation. This pattern points to a highly concentrated and cohesive core of pro-riot sentiment that contrasts sharply with the smaller, more isolated group of users voicing opposition. The structure of the network reveals how support for the riots is not only more widespread but also more deeply rooted in certain parts of the online space.

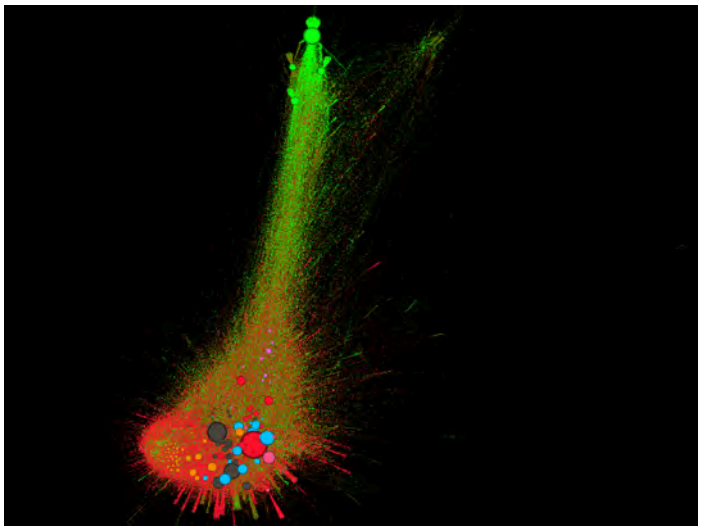


Figure 7. Edge colouring attributes green to “anti-riot” and “red” to “pro-riot” links, respectively. Node colours reflect the associated cluster for all nodes but for far-right, plot in red. Node size indicates influence exerted on the network.

Finding 4. Distinguishing influencers from amplifiers of pro-riots content.

To understand how networks sustained offline mobilisation, we distinguish between two types of users: influencers (those who primarily generate content that others engage with) and amplifiers (those who mostly react to or share that content, helping it spread further through the network). In both the pro- and anti-riot communities, we observe a striking imbalance: a relatively small number of influencers drive the conversation, while a much larger group of amplifiers helps circulate their messages. We visualise this distinction in Figure 8 and Figure 9, where node size reflects the amount of influence a user exerts (Figure 8) or receives (Figure 9).

By measuring both how much influence users exert and how much they receive, we can identify not only who is shaping the discourse, but also how concentrated or distributed that influence is. This distinction is crucial for recognising where power lies within online communities and where potential intervention or regulation may be most effective.

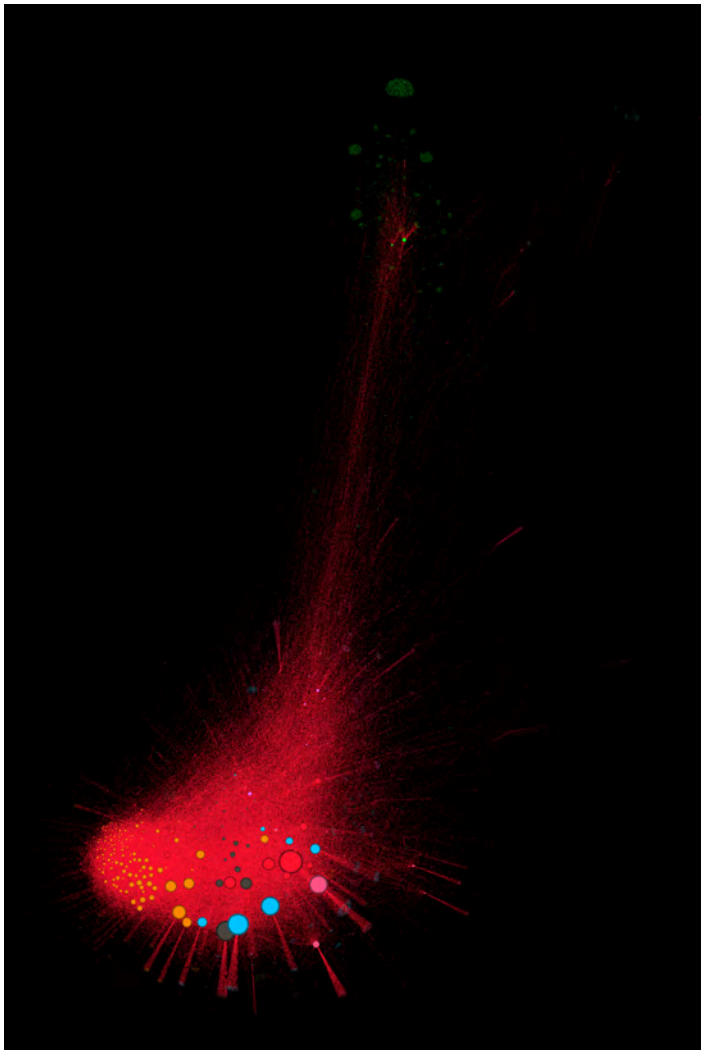


Figure 8. Network visualisation of pro-riots influencer nodes. Node size reflects in-degree, indicating how many reactions each account received from others. Larger nodes represent users who exerted the greatest influence within the network by generating content that attracted widespread engagement.

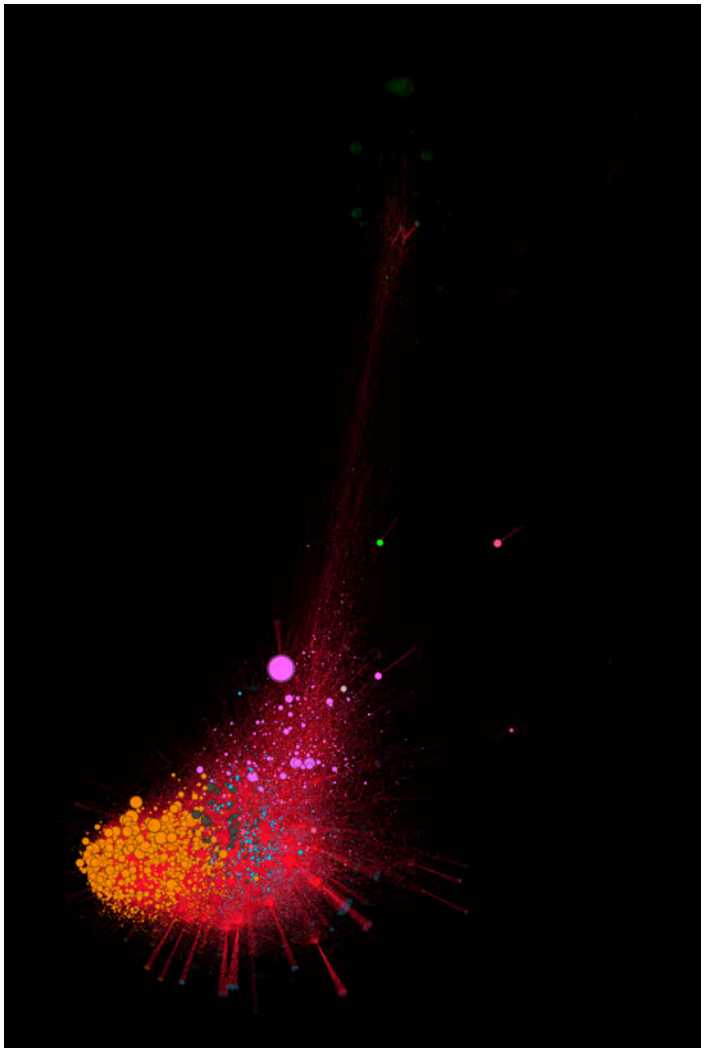


Figure 9. Network visualisation of pro-riots amplifier nodes. Node size reflects out-degree, indicating how many reactions each user gave to others. Larger nodes represent the most active amplifiers, that is, those who most frequently engaged with others' content.

THE HATE CONTAGION

We analysed three datasets (the UK dataset, the Australia dataset, and a dataset of offline incidents) to examine how anti-immigration content spread geographically within the UK and across national borders.

Methodological note

We analysed how anti-foreigner and anti-politics content spread online in the UK and Australia by modelling four separate streams of tweets (UK anti-politics, UK anti-foreigner, AU anti-politics, AU anti-foreigner) as part of a single interconnected system known as a multidimensional Hawkes process. This statistical model is widely used to capture how events can trigger further events either within the same stream (self-excitation) or across different streams (cross-excitation).

Because we did not want to impose strong assumptions on how these triggering effects unfold over time, we used a non-parametric approach (the conditional laws estimator). This method estimates the influence patterns directly from the data, by calculating how often new events occur given past activity.

To apply this method, we converted the four tweet streams into ordered sequences of timestamps. We then assumed that any given tweet could influence further activity for up to 1.5 hours, the maximum lag in our model. This time window was divided into three shorter intervals, which allowed us to estimate how the likelihood of one tweet triggering another changes over time.

This short-interval calibration strategy ensures that we can capture immediate cause–effect relationships between messages (first-order effects), while filtering out broader background trends or more complex, indirect influences and feedback loops.

Key findings

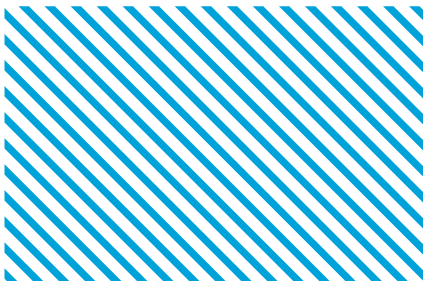
To compare the strength of influence between different narratives, we calculated kernel norms, which measure the overall impact one type of message has on another. Technically, kernel norms represent the area under each estimated curve and can be interpreted as the expected number of new messages in process i triggered by a single message in process j. All norms are grouped in the so called Impact Matrix, where each cell indicates the impact of the associated column on the associated row.

Figure 10 reports the impact matrix. From the matrix, we note that the impact of narratives varies considerably, both within countries and across countries. For instance, a single UK-based anti-politics tweet tends to reinforce only its own stream, generating on average 0.8 additional UK-based anti-politics tweets. By contrast, a single UK-based anti-foreigner tweet has a broader effect: it not only increases the likelihood of further anti-foreigner tweets in the UK but also “spills over” into the Australian context, raising the odds of both anti-politics and anti-foreigner tweets there.

Country-wise, the influence is asymmetric: messages originating in Australia do not affect UK-based streams, which is consistent with expectations. Within Australia, however, we observe stronger cross-series contamination, as both anti-politics and anti-foreigner streams are strongly driven by the domestic anti-foreigner narrative.

Anti-foreigner narratives spread more widely than anti-politics ones because they are emotionally charged, flexible enough to link to other grievances, and embedded in an asymmetric information environment. UK content travels easily into Australia, but not the reverse, while in Australia migration debates strongly spill into broader political discontent. This asymmetry likely reflects the UK’s stronger influence in the global media and political landscape, where developments in British politics and society receive more international attention than comparable events in Australia. As a result, Australian users are more exposed to, and more likely to engage with, UK-origin content than vice versa.

In Figure 10, the x-axis displays the types of messages that trigger a reaction, while the y-axis shows the types of messages that are influenced in response. Each cell quantifies how many additional messages are generated by one message of the type shown on the x-axis. For example, the top-left cell shows how many UK-based anti-politics messages are triggered, on average, by one additional UK-based anti-politics message. Darker colours represent stronger effects.



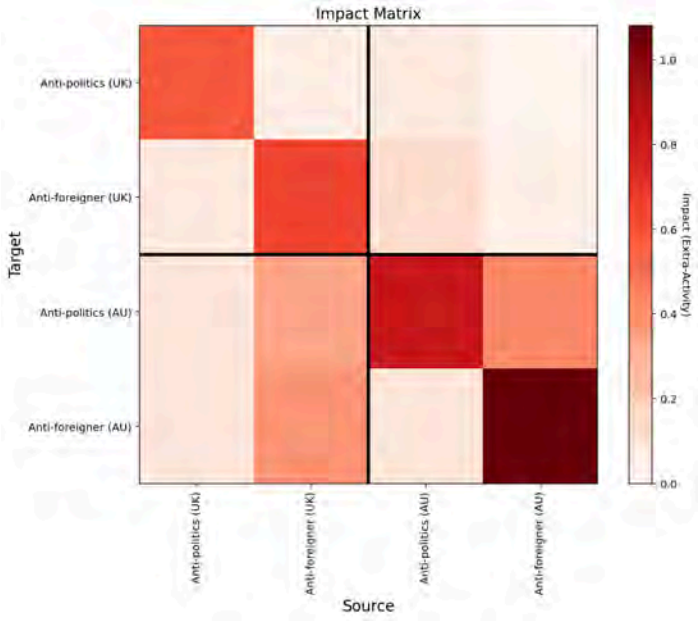


Figure 10. The figure shows narrative interactions, with each column representing the category of a message and each row the type of message it may trigger. Darker colours indicate a higher number of additional messages expected from a single new message.

Figure 11 presents the instantaneous kernel functions, which show how quickly different types of messages trigger further activity. The analysis reveals a clear timing difference: anti-politics content generates almost immediate reactions, while anti-foreigner content spreads more slowly, peaking around an hour later but persisting longer once it gains traction. This likely reflects the nature of the narratives: anti-politics messages tap into existing, generalised frustrations with institutions, making users quick to respond. By contrast, anti-foreigner content may need more framing before it spreads, but once mobilised, it resonates across communities and remains active in discussions for longer.

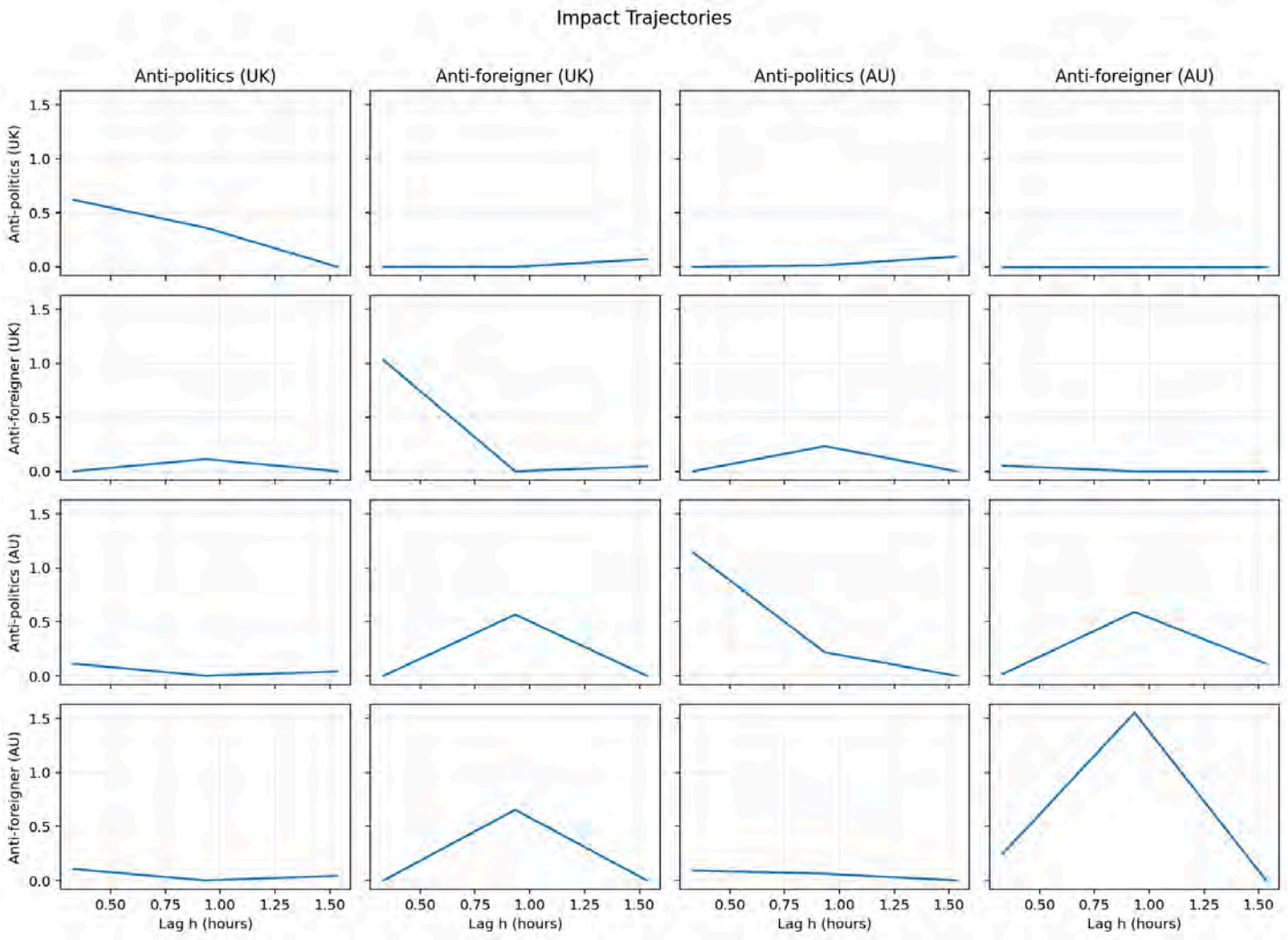


Figure 11. The figure shows instantaneous kernel functions, with each line representing how one type of message affects the timing of others; higher values indicate stronger effects.

Geographical patterns of anti-foreigner language, mobilisation and riot activity

We produced a series of maps to display descriptive spatial patterns of anti-foreigner language, mobilisation tweets, and riot events between 28 July and 10 August. We explained what are anti-foreigner tweets in the first section of this report. Mobilisation tweets are tweets that call for direct mobilisation against foreigners. The following are two examples of tweets classified in the mobilisation category:

> Young kids stabbed to death in a nursery in #Southport - Is this Correct - KILLING OUR CHILDREN ?? START FUCKING ORGANISING NOW AND DOING SOMETHING ABOUT IT - HOW LONG ARE PEOPLE GOING TO EXECPT THIS SHIT??? KILLING OUR KIDS WTF?

> Who is going to be outside City of London Magistrates Court this afternoon to show the Child killer that we stand together and hatred of any kind will never win? #Southport #BurnInHell

Figure 12 below presents daily heatmaps of anti-foreigner tweets, adjusted for local population size. This adjustment enables us to identify areas with a disproportionately high intensity of anti-foreigner sentiment, rather than simply reflecting population density, particularly in urban centres such as London.

On 28 July, the day prior to the Southport attacks, anti-foreigner discourse on Twitter was minimal, with only low-level sentiment observed in the most densely populated areas. By 29 July, coinciding with the Southport attacks, anti-foreigner tweets increased noticeably, particularly around Southport, with some tweets including explicit calls for mobilisation and violent action against immigrants in both Southport and London.

Between 30 July and 1 August, calls for anti-immigrant mobilisation were concentrated in Southport, London, and Sunderland. During this period, anti-immigration content became more widespread nationally. On 2 August, the volume and geographical spread of mobilisation messages intensified. By 3 August, mobilisation messages were evident across numerous cities, closely aligning with the locations of offline riot events.

Areas where online mobilisation tweets appeared were also more likely to experience offline violence. In contrast, regions exhibiting only general anti-immigration sentiment, without mobilisation content, did not typically see such events.

Following 5 August, online mobilisation activity declined. However, elevated levels of anti-immigration language continued, gradually diminishing by 10 August.

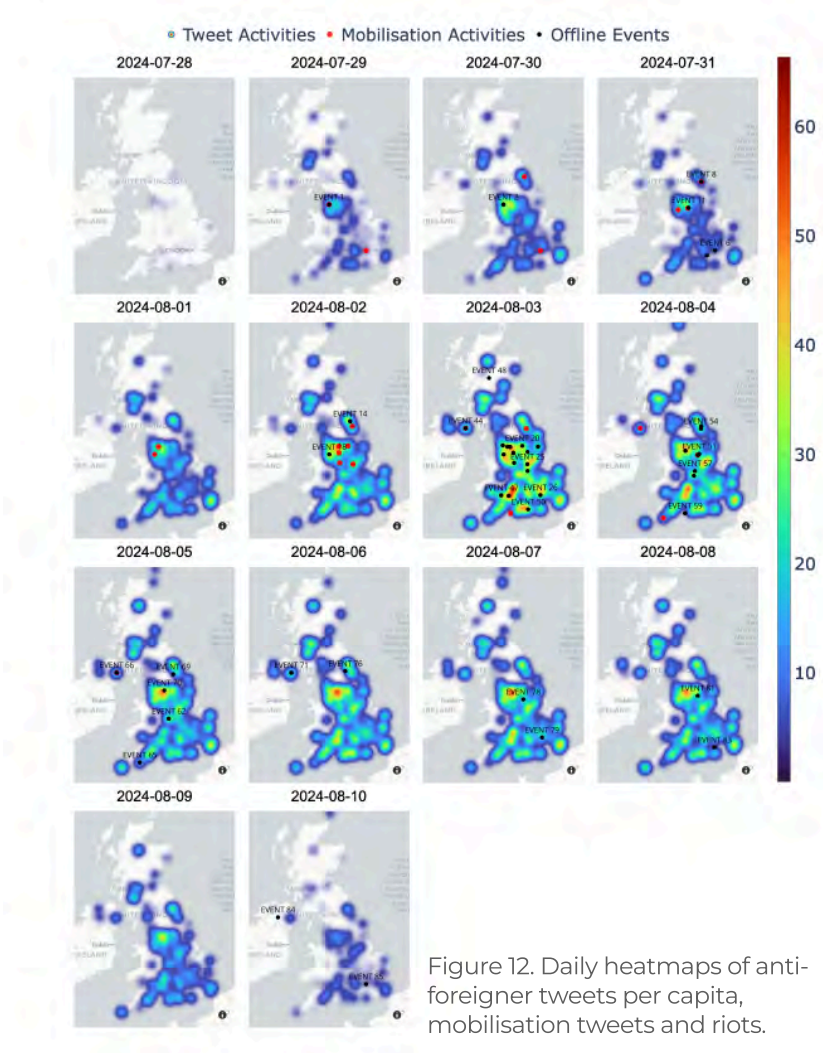


Figure 12. Daily heatmaps of anti-foreigner tweets per capita, mobilisation tweets and riots.

CONCLUSION

This report provides new evidence on the role of online language in signalling and shaping offline unrest in the context of violent political mobilisation.

The anti-immigration movement that emerged in the wake of the Southport incident was not the product of organised political ideology alone, but rather of emotionally driven, decentralised mobilisation that has been efficiently amplified online. In line with social movements scholarship (Castells, 2015), this report highlights that collective violence can emerge not from rational grievances or objective conditions, but from perceived injustice, shaped and amplified through media channels where rumours and misinformation play an important role. Social media platforms like X act as critical spaces where these perceptions are shared, intensified, and spread across highly asymmetric networks of influence.

This report highlights how reposting rates of anti-immigration content offer a sensitive early-warning signal for offline unrest. For law enforcement and crisis response agencies, this provides a valuable tool to anticipate when harmful narratives are gaining momentum, enabling proactive deployments, community engagement, or rapid response planning in high-risk areas. Unlike raw tweet volumes, reposting rates capture when fringe content begins to resonate broadly, signalling the risk of violent mobilisation.

While reposting rates offer a potentially valuable early signal of digitally driven mobilisation, we acknowledge that any operational use of this metric by law enforcement would require clear ethical guidelines, safeguards, and transparency to avoid misuse. Monitoring reposting activity should not be used to criminalise dissent or suppress peaceful protest, but rather to help distinguish between broadly expressed opinion and coordinated, hate-driven incitement. As with other intelligence tools, its value depends not only on technical validity but also on how proportionately and responsibly it is applied in practice.

The findings also show that influence is highly concentrated among a small set of influential figures. This concentration of influence also reflects X's role as a niche but active hub where certain actors, particularly those aligned with fringe or extreme narratives, congregate and coordinate. For government regulators and digital governance bodies, this shifts attention from moderating individual posts to understanding and addressing the network dynamics that enable such influencers to shape discourse across online communities. Supporting platforms to detect and mitigate high-impact accounts especially during periods of escalating sentiment can help disrupt mobilisation without over-policing general speech.

Importantly, this work distinguishes between broad hate sentiment and direct calls to action. Explicit mobilisation content urging protest or confrontation was spreading days before the most intense phase of the riots. Law enforcement, prosecutors, and content moderators alike should focus on detecting and responding to this specific content type. Human rights agencies, meanwhile, can use this insight to advocate for proportional moderation that targets incitement rather than opinion.

Although we found limited cross-border contagion into Australia, the asymmetry of this influence, where UK anti-immigration narratives triggered online activity overseas, suggests that domestic hate campaigns can have international reach, even when offline violence remains localised. This may be partly explained by shared political discourse, media ecosystems, and activist networks across the Anglosphere, which enable narratives to resonate and circulate beyond national borders.

The spatial alignment between online mobilisation content and real-world unrest underscores how even platforms with relatively modest user bases like X can offer meaningful insights into emerging risks, particularly when combined with geolocation and network analysis.

Building on these findings, future research can refine how we detect early digital signals of unrest across platforms and contexts. The insights offered here are crucial for developing proportionate, evidence-based responses to online incitement. For countries like Australia—where cross-border narrative contagion was observed—these insights are especially valuable for preventing the escalation of online anti-immigrant hate into coordinated offline violence.

REFERENCES

Castells, M. (2015). Networks of outrage and hope: Social movements in the Internet age. John Wiley & Sons.

Home Office. (2025, June 18). Prevent learning review: Axel Muganwa Rudakubana (accessible) (Redacted version). https://www.gov.uk/government/publications/prevent-learning-review-southport-attack

Institute for Strategic Dialogue, & CASM Technology. (2024, 13 September). Quantifying extremism: A data-driven analysis of UK riot-related far-right Telegram networks. Digital Dispatches. Retrieved from https://www.isdglobal.org/digital_dispatches/quantifying-extremism-a-data-driven-analysis-of-riot-related-far-right-telegram-networks/

Institute for Strategic Dialogue. (2024, 31 July). From rumours to riots: How online misinformation fuelled violence in the aftermath of the Southport attack. Digital Dispatches. Retrieved from https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/

Richler, J. (2023). From non-violent to violent collective action. Nature Reviews Psychology, 2(2), 70-70.

Venkataramakrishnan, S. (2025). From Disinformation to Violence: The UK Far Right and 2024 Riots. Counter Terrorist Trends & Analysis, 17(2).

Vergani, M., Giovannetti, A., & Goodhardt, D. (2025). Tracking the 2023 wave of anti-trans and anti-drag mobilisation in Australia. Deakin University. Retrieved from: https://tacklinghate.org/our-work/new-report-on-anti-trans-and-anti-drag-mobilisation-in-australia/

AUTHORS BIO

Matteo Vergani

Matteo is director of the Tackling Hate Lab and Associate Professor in Sociology at Deakin University and specialises in radicalisation and hate crime, publishing in leading international journals and securing large research grants. He collaborates with numerous institutions and government agencies in Australia and Canada. His research advances the systematic consolidation of knowledge in hate and extremism studies through large-scale systematic reviews and the development of rigorous measurement tools of online and online hate and radicalisation. His research programme fosters multidisciplinary collaboration across social sciences, data science, econometrics and engineering, leveraging advanced technologies for analysing digital archives and social media big data.

Andrea Giovannetti

Dr Andrea Giovannetti is Co-Director of the Tackling Hate Lab, Assistant Professor of Economics at the Australian Catholic University and a member of the Violence Research Centre at the Institute of Criminology of the University of Cambridge, where he previously held a Marie Curie Postdoctoral Fellowship. His research on organised crime, contemporary extremism and social cohesion combines machine-based econometrics with advanced computational methods in network theory to support policymakers and security agencies on a large spectrum of inter-connected issues. Andrea's collaborations with public stakeholders on complex social threats include London Metropolitan Police, Merseyside Police, Home Office and Home Affairs.

Kewen Liao

Kewen is co-director of the Tackling Hate Lab and Associate Professor in Data Analytics and Machine Learning at Deakin University. Kewen is an accomplished algorithms researcher with a PhD in Computer Science. His expertise spans data science, machine learning, and theoretical computer science, with a focus on algorithm design and analysis, clustering and graph mining, time series and streaming analytics, and image and text data analysis. He is a Chief Investigator on an Australian Research Council Discovery Project and previously led a Defence Next Generation Technologies Fund Project. He also holds a US patent with Canon Inc. for a novel object-matching method in computer vision. Kewen is committed to cross-disciplinary research, leveraging data science and AI to drive significant societal and economic impact. He also holds key leadership and professional roles in leading national and international data science and AI conferences.

Xinzhe Li

Xinzhe is a researcher specializing in large language models (LLMs), with expertise in LLM inference and reasoning—particularly search-based or multi-path exploration methods for improving planning and efficiency—as well as the evaluation of LLMs' internal components and learning dynamics. He completed his PhD in Information Technology at Deakin University. In this report, he led the development and evaluation of LLM-based classifiers for multi-label tweet annotation, targeting socially sensitive categories such as anti-immigration and hate-related content. Xinzhe has benchmarked prompting and fine-tuning strategies to assess classification performance, implemented strategies for scalable data labeling, and conducted correlation and causal analyses to examine the relationship between online language and offline riot events.

Stephanie Zi Xin Ng

Stephanie is a postdoctoral researcher at the Tackling Hate Lab. She obtained her PhD in Engineering from Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI). Her research focuses on advancing computational methods that leverage Natural Language Processing (NLP) and Large Language Models (LLMs) to address pressing social challenges. Her work spans a variety of topics, including the policy framing of contemporary social issues, stance detection in parliamentary debates, and agent-based modelling of inoculation strategies against online extremism. Her recent work explores video and image processing, including multimodal analysis of the privacy paradox in short videos and Computer Vision (CV) for hazard detection.

Dan Goodhardt

Dan is a researcher at Deakin University, where he has collaborated for over a decade on studies of hate crime and violent extremism in Australia and Southeast Asia. He has extensive experience in collecting antisemitism data through the Jewish Community's Community Security Group and currently works as a senior manager with Victoria Police. As co-convenor of the Practitioners Working Group on Tackling Hate in Victoria at the Centre for Resilience and Inclusive Societies, he connects security practitioners with researchers to advance efforts in countering hate and extremism.


TACKLING
HATE