



Anti-First Nations Hate in Australia

**Online Dynamics Across the Voice
Referendum and 2025 National Events**

Matteo Vergani, Stephanie Zi Xin Ng, Michael Lavery,
Andrea Giovannetti, Kewen Liao, Dan Goodhardt

Contents



3	Executive Summary
4	Introduction
7	Hate During the 2023 The Voice Referendum Campaign
12	Event-Driven Spikes in Anti-First Nations Hate in 2025
15	Spatial and platform variation in anti-First Nations hate between the 2023 and 2025 datasets
18	Conclusion
19	References
20	Authors



EXECUTIVE SUMMARY

Aims

This report examines how anti-First Nations hate evolved across two distinct online contexts in Australia: the 2023 Voice to Parliament referendum campaign and two nationally salient events in 2025, Anzac Day and March for Australia. Using large-scale social media and news datasets, combined with community-informed artificial intelligence classification, the report analyses how political mobilisation, public controversy, and commemorative events shape the scale, forms, and distribution of anti-First Nations hate.

Data

This report uses two main datasets of online language:

- 2023 Voice referendum dataset: 2,155,134 posts from 153,184 unique accounts collected between January 2023 and January 2024 across X, news media, Reddit, YouTube, Facebook, forums, blogs, Tumblr, Bluesky, and Instagram.
- 2025 events dataset: 876,366 posts from 219,104 unique accounts collected between March and October 2025 focusing on two events: Anzac Day (25 April) and March for Australia (31 August).

Key findings

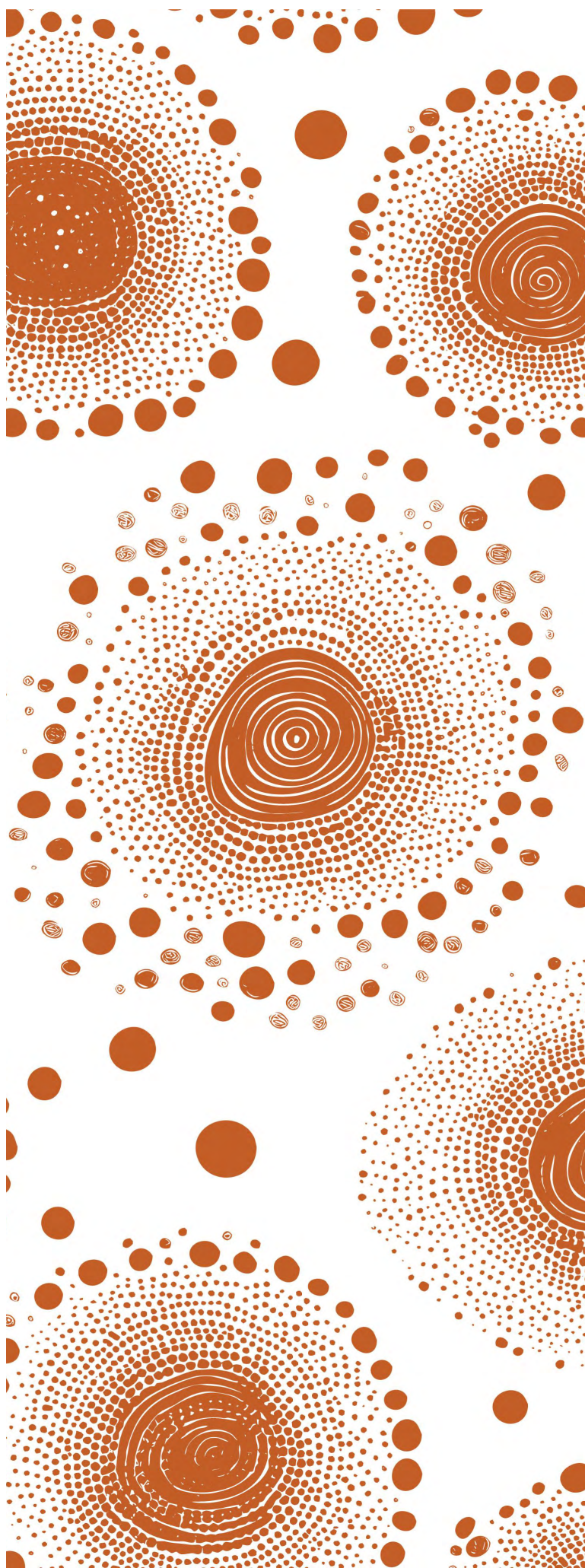
Anti-First Nations hate in Australia is highly context-dependent. During the referendum campaign, 88.8% of detected hate was Voice-based hate focused on political recognition, constitutional change, and claims of “race-based privilege”. By contrast, 96.2% of hate in the 2025 dataset consisted of racial identity hate, particularly White supremacist and dehumanising narratives.

Hate during the referendum escalated in clear stepwise phases rather than random fluctuations. Activity intensified around major campaign milestones, media controversies, and nationally broadcast political statements, peaking immediately before the October 2023 referendum before collapsing rapidly afterwards.

First Nations public figures experienced disproportionately high levels of hate. Figures such as Linda Burney, Lidia Thorpe, Noel Pearson, and Adam Briggs attracted substantially higher proportions of hateful content than non-Indigenous political figures, indicating patterns of selective racialised targeting rather than general political disagreement.

Misinformation played multiple roles in amplifying hostility. Sovereignty-related misinformation increased the visibility and scale of hostile discourse, threat-based misinformation was associated with increased hate levels, and identity-delegitimising misinformation reinforced longer-term racial hostility towards First Nations peoples.

The findings suggest that online anti-First Nations hate operates across a continuum ranging from policy-contingent hostility to explicit racial degradation. The results highlight the importance of context-sensitive interventions addressing misinformation, amplification dynamics, and escalating racialised mobilisation across digital platforms.



INTRODUCTION

1.1. Aims

This report examines how anti-First Nations hate, as perceived through large-scale online discourse and interpreted via community-informed classification, has evolved in Australia across two distinct contexts: the 2023 Voice to Parliament referendum campaign and key nationally salient events in 2025, including Anzac Day and the March for Australia. Drawing on multi-platform social media and news data, the report analyses how different political and commemorative contexts shape the scale, structure, and expression of hate

1. Map the scale, composition, and temporal evolution of anti-First Nations hate in Australian online discourse, comparing patterns across the 2023 The Voice referendum campaign and 2025 event-driven contexts;
2. Distinguish analytically between different forms of hate, particularly racial identity hate and voice-based hate, in order to clarify how hostility targets either Indigenous identity itself or participation in political recognition processes;
3. Examine the distribution of anti-First Nations hate across platforms, networks, and geographic contexts, identifying key sites of amplification, patterns of concentration, and implications for policy and intervention.

1.2. The Australian context

In Australia, anti-First Nations racism is deeply rooted in the ongoing legacies of colonisation, dispossession, and exclusion from political and social institutions. While often framed as a historical issue, research consistently shows that racism remains embedded in contemporary systems and everyday interactions (Moreton-Robinson, 2015; Paradies, 2016). Evidence from recent research indicates that racism experienced by First Nations peoples is both interpersonal and systemic, occurring across workplaces, schools, public spaces, and online environments, and often reflecting broader historical and institutional dynamics (Allison et al., 2025).

In this context, contemporary expressions of racism frequently draw on enduring stereotypes and deficit narratives. First Nations peoples are commonly portrayed as inherently linked to crime or disorder, particularly in public debate concerning the Northern Territory, or as dependent on welfare and undeserving of public resources (Carlson & Frazer, 2021; Allison et al., 2025). These narratives are often accompanied by forms of “Aboriginality policing”, in which individuals’ identity is questioned on the basis of skin colour or perceived ethnicity (Carlson & Frazer, 2021). At the same time, historical revisionism persists in claims that diminish Indigenous cultures or justify colonisation, framing First Nations peoples as lacking civilisation or benefiting from settler intervention (Carlson & Frazer, 2021). Such framings normalise racialised assumptions and provide a foundation for both subtle and overt forms of hostility.

Political events can intensify these dynamics. Debate surrounding the 2023 Indigenous Voice to Parliament referendum, for example, saw the circulation of claims that recognising First Nations perspectives would “divide the nation by race”, reframing structural inequality as preferential treatment and limiting the capacity to address racism at a systemic level. Following the referendum,



community reports collected by the Jumbunna Institute for Indigenous Education and Research through its Call It Out report indicated increased hostility and more overt expressions of racism in both online and offline settings (Allison et al., 2025).

Other nationally salient events have similarly acted as triggers for contestation. Public controversy around Welcome to Country ceremonies reflects ongoing resentment towards Indigenous recognition and misunderstanding of cultural protocols, mischaracterised as exclusionary, unnecessary or derided as fraudulent performative creations rather than being acknowledged and appreciated as deeply important cultural practices.[1] In more extreme cases, this hostility has escalated into violence. The 2025 March for Australia was followed by a targeted attack on Camp Sovereignty in Naarm Melbourne involving members of the Nationalist Socialist Network, suggesting the convergence of racial ideology.[2]

Taken together, these patterns indicate that anti-First Nations hate in Australia operates across a continuum, from everyday stereotyping and exclusion to coordinated and violent acts. As with other forms of hate, online and offline dynamics are closely intertwined, with public discourse, political debate, and media coverage shaping the conditions under which hostility emerges, circulates, and intensifies.

[1] <https://www.abc.net.au/news/2025-04-25/melbourne-anzac-day-welcome-to-country-hecklers/105215124>

[2] <https://www.abc.net.au/news/2026-02-10/thomas-sewell-neo-nazi-camp-sovereignty-melbourne-court/106326988>

1.3. Data

This report analyses two datasets of online content.

1. 2023 dataset (The Voice Referendum). The first dataset comprises 2,155,134 individual entries authored by 153,184 unique accounts. It was collected using a hashtag-based query designed to capture the full spectrum of online discussion surrounding the 2023 Voice referendum between 1 January 2023 and 1 January 2024. The query incorporates official campaign terminology (for example, #referendum, #referendum2023, #thevoice, #voicetoparliament), general descriptors (#voice, #voteearly), and a comprehensive set of campaign-aligned hashtags from both the Yes and No positions. These include affirmative mobilisation tags (such as #voteyes, #yes23, #yestoice, #walkforyes, #writeyes) as well as oppositional and critical tags (including #voteno, #no23, #notmyvoice, #votenotothevoice, #votenotodivision, and related variants). By incorporating parallel and competing campaign framings, slogans, and derivative hashtags, the query captures not only formal campaign messaging but also grassroots advocacy, criticism, parody, and post-referendum reflection. The dataset reflects a broad and contested discursive environment in which political mobilisation, identity claims, misinformation, counter-speech, and everyday political commentary coexist.
2. 2025 dataset (Anzac and March for Australia). The second dataset contains 876,366 individual entries authored by 219,104 unique accounts, and it was collected using a proximity-based Boolean query designed to capture online discussions of First Nations-related events and controversies in Australia between 31 March 2025 and 27 October 2025. The query operates in two principal ways. First, it retrieves posts in which references to Anzac or Anzac Day events (including dawn services and ceremonies) appear in close proximity to terms associated with Welcome to Country or Acknowledgement of Country protocols, Indigenous or Aboriginal Elders, and reactions such as booing, heckling, disrespect, or public backlash. Second, it captures discussions of nationalist or far-right rallies, including “March for Australia”, anti-immigration protests, and related events, when these appear near terms indicating mobilisation, confrontation, violence, or extremist groups (for example, neo-Nazi or white supremacist references), as well as references to sites of First Nations activism such as camps or sovereignty spaces. Geographic filters (including Australian place names, country identifiers, and .au domains) further constrain the dataset to the Australian context. Together, these conditions capture a discursive field in which commemoration, Indigenous recognition, nationalist mobilisation, protest activity, conflict narratives, and public debate intersect.

Most entries come from X, accounting for 1,376,585 posts on the Voice referendum and 670,472 on other First Nations-related events. Content from X includes original posts, replies, and shares, allowing us to observe not only how anti-First Nations hate is expressed, but also how it is amplified through retweeting and reactive exchanges. X is particularly important for this analysis because it plays a central role in real-time political debate in Australia and allows users to rapidly circulate emotionally charged or polarising content.

The dataset includes a substantial volume of news content (491,260 entries on the Voice referendum and 111,564 on other First Nations-related events) drawn from a wide range of Australian media outlets. These include national and metropolitan sources such as The Sydney Morning Herald, Herald Sun, Sky News, and MSN, as well as numerous regional and local newspapers across Australia. This diversity allows us to capture how discussions about First Nations and related issues appear not only in opinionated commentary, but also in news reporting, headlines, and comment sections, which can shape public interpretation of events.

Smaller but analytically important portions of the dataset come from online forums and blogging platforms. Forum content is primarily drawn from Australian discussion boards such as BigFooty and HotCopper, where users often engage in longer, less moderated discussions around politics, national identity, and current affairs. Tumblr contributes a distinct form of content, characterised by highly expressive posts, visual material, and the circulation of memes and slogans, which can play a role in spreading both overt and coded forms of hate. Blogs and other minor sources account for a small number of entries but provide additional context for how narratives circulate across different parts of the online ecosystem.

The remaining content is drawn from platforms including Reddit, YouTube, Bluesky, Facebook, and Instagram. Reddit constitutes the third-largest source of content, with 90,369 and 49,242 entries across the two datasets, followed by YouTube (27,667 and 11,203 entries) and Facebook (15,810 and 3,846 entries). Bluesky and Instagram can be characterised as platforms with comparatively younger and more digitally native user bases. They contribute comparatively smaller volumes, each comprising several hundred entries, reflecting their more limited representation in the sampled data.

Across all platforms, both dataset includes original posts (737,349 and 150,333 entries respectively), replies (431,874 and 211,880 entries respectively), and shares (985,909 and 514,153 entries respectively). This mix is critical, as hate does not spread only through original messages, but also through repetition, endorsement, and contestation, with reposting and replying often extending the reach and lifespan of hostile content.



1.4. Classification

To capture toxicity, we used the toxicity classifier developed by Google's Perspective API. Toxicity is defined as language that is rude, disrespectful, or likely to make users disengage from a conversation. Unlike our community-informed hate classifiers, the toxicity measure is not target-specific and instead captures broader hostile or aggressive communication patterns. Each social media post received a continuous toxicity score ranging from 0 to 1, with higher scores indicating greater levels of toxic language. This measure allowed us to assess broader changes in hostility and conversational aggression at scale across online discussions.

To capture anti-First Nations hate, we developed a two-stage artificial intelligence pipeline combining a community-informed supervised classifier with a generative reasoning model. The first stage consisted of a classifier trained on 2,877 unique social media text samples annotated by trained community annotators. Each sample received multiple annotations, and a soft-labelling approach was used to account for uncertainty by averaging annotator labels and assigning neutral values to missing annotations during weighted training. Final labels were assigned using majority agreement. The final training dataset included 1,731 hateful and 1,146 non-hateful posts. We tested several contemporary language models and training strategies. The best-performing model was RoBERTa-Large using partial fine-tuning, where only a subset of model parameters was updated to adapt the model to anti-First Nations hate while reducing overfitting. The final classifier achieved strong performance, with an AUC of 96.08%, accuracy of 89.64%, precision of 89.04%, recall of 90.91%, and an F1 score of 89.97%, indicating high agreement with community judgements and strong capacity to identify hateful content while minimising false positives and false negatives. Because supervised classifiers may struggle with emerging, coded, implicit, or previously unseen forms of hate, we implemented a second-stage generative model. This model analysed content flagged in the first stage against the same structured hate codebook developed and used by the human annotators, allowing contextual reasoning about targeted groups, dehumanising language, and anti-First Nations hostility. This approach improved interpretability and sensitivity to nuanced forms of harmful content while maintaining conceptual consistency with community-defined standards.

1.5. Racial identity hate vs Voice-based hate

Through the report we distinguish between two analytically related but conceptually distinct forms of hate as experienced by First Nations annotators: Racial identity hate and Voice-based hate. Although both emerge within the broader context of public debate surrounding constitutional recognition, they differ in their primary target and underlying logic.

Racial identity hate refers to hostility directed at Aboriginal and Torres Strait Islander identity itself. Annotators identify this form of hate where the legitimacy, authenticity, or worth of Indigenous identity is questioned, diminished, or degraded. This includes assertions of racial hierarchy, claims that Aboriginal people are “not really” Aboriginal due to mixed ancestry, denial of Traditional Ownership, or framing Aboriginal identity as manipulative, fraudulent, or opportunistic. The defining feature is that the target is who First Nations people are (their skin colour, rituals, ancestry, cultural status, or collective belonging) irrespective of any specific policy proposal. In this category, the referendum may serve merely as a context in which pre-existing but hitherto unspoken racialised narratives are expressed.

Voice-based hate, by contrast, is directed at First Nations peoples in relation to their participation in, or perceived benefit from, the proposed Voice to Parliament. Annotators perceived and identified this form of hate where social media content suggested for example that the Voice would create division, constitute “apartheid”, or grant unfair race-based privilege to First Nations peoples, and content that framed the referendum as pointless denying that racial struggles of First Nations peoples in Australia are a problem for the country. While these expressions may draw upon racialised assumptions, the immediate object of resentment is the proposed political mechanism of representation and recognition rather than Indigenous identity per se. The hate is therefore conditional upon the governance proposal and framed through arguments about equality, fairness, or national unity. Our community annotators perceived this type of language as highly offensive and racist. Importantly, annotators did not classify all opposition to the Voice as hateful. For example, statements were not coded as hate if they reflected First Nations perspectives questioning the legitimacy of the referendum (such as concerns about the absence of a Treaty with all First Nations peoples in Australia), or if they expressed straightforward opposition (e.g. “Vote No”).

In practice, the two forms often intersect. However, community annotators report that distinguishing between identity-based degradation and policy-contingent hostility improves coding reliability and clarifies the different harms experienced by First Nations peoples.

To distinguish between Racial identity hate vs Voice-based hate, we conducted a thematic analysis of the language contained in our two datasets using BERTopic, a machine-learning tool that groups large volumes of text into clusters based on semantic similarity. In simple terms, BERTopic identifies patterns in how people talk about an issue and groups together posts that use similar language or express similar ideas. To ensure that these clusters were interpretable and aligned with the aims of the study, we used a large language model to generate clear, structured labels and descriptions for each topic.



2. HATE DURING THE 2023 THE VOICE REFERENDUM CAMPAIGN

During the referendum campaign, voice-based hate was the most prominent form of hate that our classifier captured (88.8%). Within this content, the most common subtype was Voice as Racism (45.2%), where posts framed the Voice as racist or as granting unfair race-based privilege to First Nations peoples. The second most prominent category was Voice Delegitimisation (43.6%), which characterised the referendum as pointless or useless because it wasn't addressing a real issue, denying the historical and ongoing racial struggles of First Nations peoples in Australia.

About 11.2% of posts were classified as racial identity hate in our dataset capturing conversations about the Voice campaign. Sub-themes included Welfare Grievance Framing, which depicts First Nations peoples as exploiting public resources; Identity-Based Attacks, targeting public figures for their Aboriginality; and Aboriginality Policing, where individuals' Indigenous identity is questioned or invalidated. Explicitly supremacist narratives, criminal stereotyping, and dehumanising content were present but comparatively less frequent. The full breakdown is presented in Table 1.

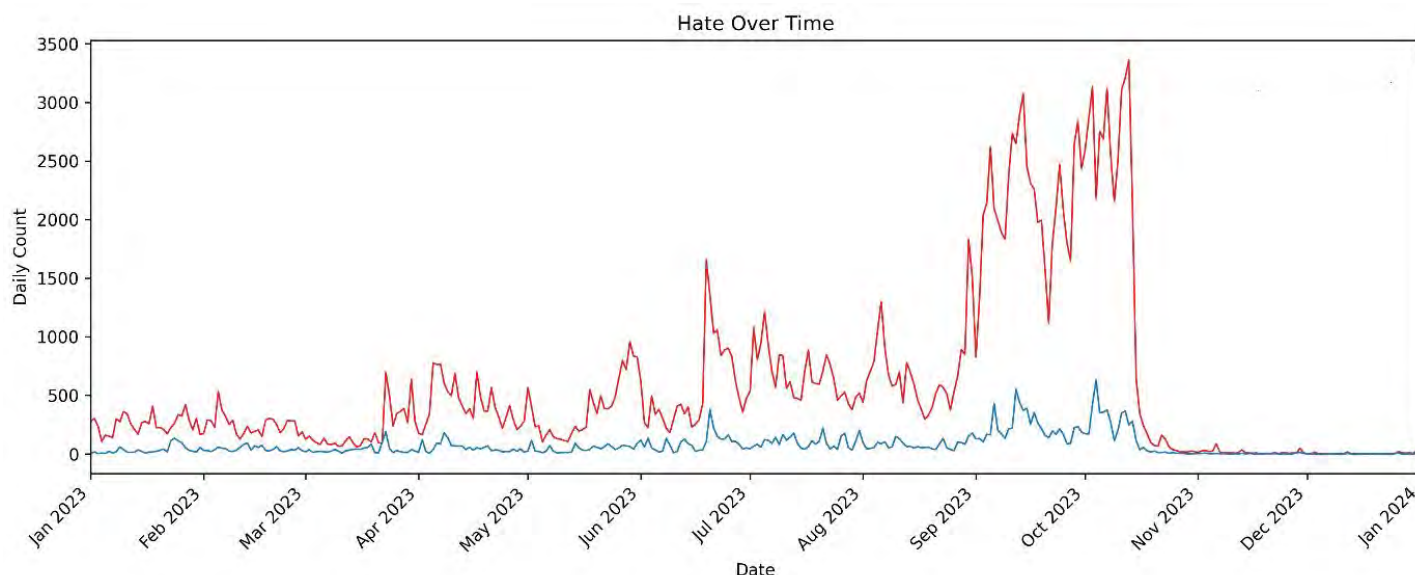
Table 1. Categories of anti-First Nations hate in the Voice campaign dataset.

Type of hate	Category	Description	Frequency	%
Voice-based hate	Voice as Racism	Posts suggesting that the Voice referendum was racist or a form of apartheid.	106,412	45.2
	Voice Delegitimisation	Posts delegitimising the Voice denying the historical and ongoing struggles of First Nations peoples.	102,744	43.6
Racial identity hate	Welfare Grievance Framing	Claims that First Nations peoples are 'free riders', a 'grievance industry', or exploit aid.	13,679	5.8
	Identity-Based Attacks	Posts attacking First Nations public figures for their Aboriginality.	5,344	2.3
	Criminal Stereotyping	Posts associating First Nations peoples with crime, alcohol, or disorder.	2,435	1.0
	Aboriginality Policing	Posts questioning or denying individuals' Aboriginal identity.	1,985	0.8
	Dehumanisation	Posts dehumanising First Nations peoples or their culture.	1,859	0.8
	White Supremacy	Posts implying superiority of White, settler, or European Australians.	973	0.4
	Other (generalised hostility and grievance narratives)	Residual category capturing related hostile content not covered above.	113	0.1

2.1. Hate trends during the referendum campaign

Across 2023, anti-First Nations hate followed a clear pattern of stepwise escalation rather than random fluctuation. As Figure 1 shows, activity remained low until mid-March, followed by escalating phases from late March through May. From mid-June, hate stabilised at a higher baseline, shifting from episodic spikes to sustained intensity. This momentum continued through July and August, peaking sharply before the October referendum, then declining rapidly afterwards.

Figure 1. Daily counts of voice-based hate (red line) and racial identity hate (blue line) during the referendum campaign.



2.1.1. Early campaign period (January–March 2023)

In January and February 2023, hate levels fluctuated within a relatively contained range, typically between 200 and 400 posts per day, with only brief surges that did not establish a new baseline. A more consequential shift emerged in late March, when activity rose sharply and remained elevated across several days. The timing suggests mobilisation and anticipatory media framing building immediately prior to the 31 March Voice to Parliament 'No' campaign tour launch in Tamworth by "Recognise a Better Way" movement 2023, led by Warren Mundine.[3]

A further wave occurred in late May. On 18 May 2023, daily volume reached 587 posts, when Stan Grant publicly stepped away from the ABC on 18–19 May citing racial abuse.[5] This episode demonstrates how elite-level controversy involving a prominent Indigenous public figure coincided with a measurable increase in anti-First Nations hate. Importantly, hate levels did not return to baseline but rather increased in the following week, with a new peak on 29 May, when hate activity reached 1,020 posts, the local maximum of this wave, followed by 876 on 30 May and 915 on 31 May.

2.1.2. Sustained escalation (April–May 2023)

Early April 2023 marks a transition from isolated spikes to sustained high-intensity activity, with daily volumes consistently elevated at around 800 posts across several consecutive days. This indicates a prolonged phase rather than a single anomaly. Activity remained high later in the month, including a notable spike on 17 April, coinciding with the launch of the No advertising campaign and associated media attention.[4] The 5–11 April escalation likely reflects ramping attention in advance of that campaign push.

[3] <https://www.abc.net.au/news/2023-04-01/voice-to-parliament-no-campaign-launched-amid-tamworth-protest/102173596>

[4] <https://www.abc.net.au/news/2023-04-17/no-campaign-launches-advertising-campaign/102232974>

[5] <https://www.bbc.com/news/world-australia-65642975>

2.1.3. Narrative consolidation and step-change mobilisation (June–August 2023)

A more pronounced shift occurred in mid-June, when daily volume surged to nearly 1,800 posts, marking a transition to a substantially higher level of activity. Unlike earlier spikes, this increase was sustained across multiple days, establishing a new operating range. While no single event explains the shift, it aligns with intensified referendum narratives, particularly around “division” and “race-based laws”, and the widespread circulation of the slogan “If you don’t know, vote No”. A further peak emerged in early July, with elevated levels persisting through mid-month. This period coincides with high-profile commentary, including a 17 July opinion piece by John Anderson framing the Voice as perpetual grievance,[6] followed by further national debate involving Lidia Thorpe on 20 July[7] and commentary from John Howard on 26 July.[8] The clustering suggests narrative reinforcement associated with a series of significant campaign-relevant events.

In early August, activity again intensified. On 5 August 2023, daily volume reached 1,146 posts, rising further to 1,385 on 6 August, the August maximum. This surge preceded the 8 August repeal of Western Australia’s heritage laws,[9] indicating that debate was already peaking before the formal policy decision. Subsequent high-profile amplification, including for example Pauline Hanson’s viral video around 10–11 August,[10] likely sustained elevated levels rather than initiating the rise.

A further step-change occurred at the end of August. On 30 August 2023, daily volume rose to 1,983 posts, followed by 1,688 on 31 August. This marks the highest levels observed up to that point in the dataset. The spike occurred when Prime Minister Albanese announced the referendum date (14 October 2023) on 30 August 2023, which also formally kicked off the final campaign period.[11]

2.1.4. The final rush (September–October 2023)

From early September, the time series entered a sustained high-intensity plateau, with daily volumes consistently above 2,000 and peaking at approximately 3,400 posts. This peak aligns with Jacinta Price’s National Press Club statement, marking a key moment in the campaign’s escalation.[12] This is one of the clearest same-day correspondences between a nationally broadcast political statement and peak online hate.

Early October marked the campaign’s peak, with sustained activity above 3,000 posts per day and a maximum of approximately 3,600. This period represents the highest intensity phase of the time series prior to the referendum. The 3 October surge occurred two days before a major No campaign event in Adelaide on 5 October[13] and eleven days before the referendum on 14 October. The entire period from 1 to 13 October reflects campaign culmination rather than a single trigger. Hate levels reached their structural peak immediately prior to the vote.

After 14 October 2023, when daily volume was 2,382 posts, activity collapsed abruptly. On 15 October, volume fell to 753 posts, and on 16 October to 376. By late October, daily totals were below 100. By December, most days recorded single-digit counts or zero. This pattern is best characterised as a regime collapse rather than a gradual decline. Once the referendum concluded, mobilisation and antagonistic framing dissipated almost immediately.

Overall, the data reveal not random volatility but stepwise escalation across distinct waves, each increasing in magnitude from late March through early October. The clearest and most consistent correspondences occur around high-visibility media debates, nationally broadcast statements, and campaign climax moments. Spikes typically occur on the same day as major events or within one to three days, consistent with rapid online diffusion dynamics. The dominant structural pattern is progressive intensification over seven months, culminating in mid-October 2023, followed by immediate post-referendum contraction.



[6] https://www.theaustralian.com.au/subscribe/news/1/?sourceCode=TAWEB_WRE170_a&dest=https%3A%2F%2Fwww.theaustralian.com.au%2FInquirer%2Findigenous-voice-to-parliament-will-be-a-distraction-from-the-real-problems%2Fnews-story%2Ff783bdaeeea2408613af2142dfb2b3f8&memtype=anonymous&mode=premium&v21=GROUPB-Segment-2-NOSCORE&V21spcbehaviour=append

[7] <https://www.sbs.com.au/news/article/window-dressing-lidia-thorpes-blak-sovereign-movement-releases-their-own-voice-pamphlet/w8e9dxg6t>

[8] <https://www.bbc.com/news/world-australia-66309637>

[9] <https://www.abc.net.au/news/2023-08-08/roger-cook-repeals-aboriginal-cultural-heritage-laws/102699678>

[10] <https://www.abc.net.au/news/2023-08-11/one-nation-shares-misleading-voice-parliament-video/102697696>

[11] <https://www.abc.net.au/news/2023-08-30/indigenous-voice-to-parliament-referendum-date-announcement-live/102786994>

[12] <https://www.sbs.com.au/news/article/jacinta-price-likens-indigenous-and-irish-australian-trauma-david-littleproud-applauds/6lalawgxp>

[13] <https://www.abc.net.au/news/2023-10-06/no-campaign-event-with-jacinta-price-kerryne-liddle/102939884>

2.2. Public figures in anti-First Nations hate discourse

We analysed the public figures most frequently mentioned in the dataset and the proportion of posts mentioning them that were classified as hateful. A clear pattern emerges: hate is disproportionately concentrated on First Nations public figures, particularly those associated with the Voice referendum. Among the most prominent examples is Linda Burney, who is mentioned in over 41,000 posts, with more than a quarter of these (27.1%) classified as racial identity hate. Similarly, Lidia Thorpe (13.9%), Noel Pearson (10.0%), and Adam Briggs (10.8%) all attract elevated levels of hate relative to their overall visibility. By contrast, non-Indigenous political figures, even those who are highly visible and attract large volumes of discussion, tend to attract substantially lower levels of hate. For example, Anthony Albanese is mentioned in over 138,000 posts but has a lower hate rate (14.0%) than Linda Burney, while figures such as Peter Dutton (0.6%), Barnaby Joyce (1.1%), and John Howard (1.3%) attract minimal levels of hate relative to total discussion. Instead, these figures are more likely to be associated with general political toxicity.

The data also reveal a distinction between political alignment and hate intensity. While both pro- and anti-Voice figures receive criticism, those advocating for the Voice (particularly First Nations leaders) consistently experience higher levels of hate. Campaign-related entities such as Yes23 attract notable hostility (9.2%), but still less than individual First Nations figures, reinforcing the importance of personalised targeting in the dynamics of hate.

Taken together, these findings indicate that online hostility during the referendum was not simply a function of political disagreement or visibility. Instead, it reflects a pattern of selective, racialised targeting, in which First Nations individuals, particularly those in leadership or advocacy roles, bear a disproportionate share of anti-First Nations hate.

2.3. Misinformation during the referendum campaign

To examine the relationship between misinformation and hate during the 2023 referendum campaign, we draw on a dataset of verified false claims compiled by AAP FactCheck. AAP FactCheck is an independent fact-checking unit of the Australian Associated Press, which provides impartial, evidence-based analysis of public claims across a wide range of topics.^[14] Its work follows verification standards and international best practices, with accreditation from the International Fact-Checking Network. The dataset used here concentrates exclusively on misinformation identified as focusing on the Voice referendum and First Nations peoples in Australia.

Our analysis shows that misinformation can be grouped into three distinct types, each associated with different patterns of amplification and hostility.

The first type consists of high-volume claims that generate substantial increases in attention without proportionally increasing hate intensity. These narratives tend to centre on sovereignty and external control, such as claims that Australia would be “signed over to the United Nations” if Indigenous rights were recognised. Following the

circulation of this claim, posts increased from 1,711 to 3,162, while the mean proportion of hate slightly decreased from 0.179 to 0.166. A similar pattern is observed in other sovereignty-related claims, where large increases in posting volume are not accompanied by corresponding increases in hate rates. This suggests that such misinformation primarily functions to expand the visibility and reach of hostile discourse, increasing the absolute number of hateful posts even when their relative proportion remains stable.

The second type comprises threat-based misinformation, which frames the Voice as a direct risk to individual rights, property, or economic security. These claims are more consistently associated with increases in hate rates. For example, the false claim that the Voice would allow the renaming of Australian cities is associated with an increase in the proportion of hate from 0.082 to 0.121. Similarly, claims that the Voice would enable the seizure of private property show an increase from 0.085 to 0.115, while misinformation about financial obligations (such as “secret” taxes or rent payments) shows a marked rise from 0.050 to 0.144. These patterns indicate that when misinformation is framed in terms of material threat, it is more likely to intensify hate rather than simply increase discussion volume. Such narratives appear to be associated with concerns about loss, fairness, and personal impact, thereby amplifying antagonistic responses.

The third type involves identity-delegitimising claims, which directly target the legitimacy, authenticity, or cultural standing of First Nations peoples. These include claims that the Uluru Statement from the Heart is inauthentic or copied, that Aboriginal cultural practices were invented by non-Indigenous individuals, or that Aboriginal people were not the first inhabitants of Australia. These claims are analytically significant because they bridge voice-based and racial identity hate, shifting the focus from opposition to a policy proposal to broader challenges to Indigenous identity. For instance, the claim that Aboriginal dot painting was invented by a non-Indigenous teacher is associated with sustained high levels of hate (2,862 posts before and 2,619 after, with mean hate levels remaining above 0.14). Unlike threat-based misinformation, these narratives do not necessarily produce sharp increases in hate intensity, but they reinforce and normalise racialised hostility over time.

This typology shows that misinformation and hate interact through multiple mechanisms. Some claims amplify the scale of discourse, others intensify hostility, and others embed and legitimise identity-based harm. This distinction is critical for understanding how misinformation contributes to anti-First Nations hate, and why different types of narratives may require different responses.

To complement this analysis, Figure 2 presents a weekly time series of anti-First Nations hate alongside general toxicity (measured using the Perspective API)^[15], with periods of heightened misinformation activity highlighted. Weeks in which more than ten misinformation claims were debunked are marked with red bars, indicating moments of intensified misinformation circulation. Several of these periods align closely with increases in both hate and toxicity: late March (week of 27 March; 18 claims), mid-June (19 June; 14 claims), and mid-August (14 August; 13 claims). This correspondence is particularly evident during the June inflection point, where both indicators shift to a higher baseline, suggesting a broader escalation in antagonistic discourse.

[14] <https://www.aap.com.au/factcheck/who-we-are/>

[15] https://developers.perspectiveapi.com/s/about-the-api-key-concepts?language=en_US in other

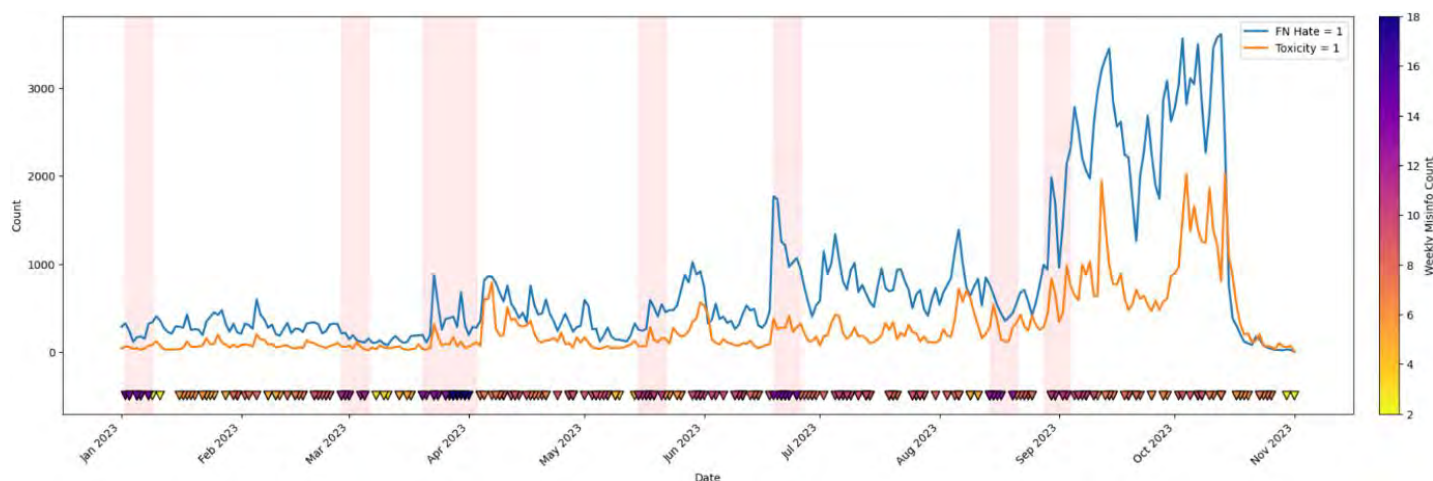


Figure 2. Weekly trends in anti-First Nations hate and toxicity during the 2023 referendum campaign, with periods of heightened misinformation activity highlighted.

However, the relationship is not uniform across all periods. Some spikes in misinformation activity, such as late February (12 claims), are associated with more modest changes in hate intensity, reinforcing the earlier finding that misinformation does not mechanically translate into higher levels of hate. Instead, the figure illustrates that misinformation operates as a contextual amplifier, contributing to periods of heightened hostility when combined with broader campaign dynamics, media attention, and political events.

Importantly, the co-movement of toxicity and hate indicates that misinformation is embedded within a wider deterioration of discourse quality, rather than functioning as an isolated driver. Overall, the temporal alignment between misinformation peaks and increases in hostility supports the conclusion that misinformation plays a facilitating and amplifying role, while its specific effects depend on the type of narrative and the surrounding information environment.





3. EVENT-DRIVEN SPIKES IN ANTI-FIRST NATIONS HATE IN 2025

A Comparative Analysis of ANZAC Day (25 April) and March for Australia (31 August)

This section examines how different types of nationally salient events shape the volume, timing, and character of anti-First Nations hate online. It asks: how do distinct public events, varying in purpose, symbolism, and political context, act as triggers for online hate targeting First Nations peoples? To address this question, we analyse two case studies from 2025: Anzac Day (25 April) and March for Australia (31 August). These events provide analytically useful contrasts between a nationally institutionalised day of commemoration and a political mobilisation. This allows comparison of how different public narratives and forms of participation shape online hate dynamics.

In 2025, public discussion surrounding Anzac Day included controversy over the booing of a Welcome to Country at commemorative events, which drew widespread media attention and public debate.^[1] The incident became a focal point for contestation over the place of First Nations recognition within national rituals. Hateful tweets during the Anzac Day period framed Indigenous recognition as inappropriate or disrespectful within a national ritual, often extending into broader claims about national identity and ownership. This discourse frequently incorporated racialised narratives, including assertions of White national belonging, challenges to Indigenous legitimacy, and derogatory or dehumanising characterisations of First Nations peoples.

March for Australia represents a different type of trigger event, a politically oriented public mobilisation associated with nationalist themes. During the march, tensions were heightened by the attack on Camp Sovereignty in Naarm

Melbourne by members of the neo-Nazi group Nationalist Socialist Network, which became a key point of discussion across online platforms.^[2] This incident foregrounded issues of Indigenous sovereignty, protest, and counter mobilisation, generating a distinct discursive environment compared to Anzac Day. A dominant theme in hateful contents during March for Australia involved framing the Camp Sovereignty attack as either a legitimate expression of national unity or as aligned with white supremacist and neo-Nazi ideology of the Nationalist Socialist Network. Many posts incorporated explicit racialised claims, including anti-immigration sentiment, challenges to Indigenous sovereignty, and dehumanising or exclusionary portrayals.

3.1. The Composition of Anti-First Nations Hate

If we consider the whole 2025 dataset including discussions surrounding Anzac Day and March for Australia, anti-First Nations hate was overwhelmingly concentrated within racial identity categories, accounting for 96.24% of all detected content. Within this, two dominant forms emerged: White Supremacy (46.71%), which asserted or implied the superiority of White, settler, or European identities, and Dehumanisation (42.82%), which characterised First Nations peoples and cultures in degrading terms. Other categories, including Welfare Grievance Framing, Aboriginality Policing, and Identity-Based Attacks, appeared only marginally, while voice-based hate constituted less than 1% of the dataset. The data is reported in Table 2.

[1] <https://www.abc.net.au/news/2025-04-25/melbourne-anzac-day-welcome-to-country-hecklers/105215124>

[2] <https://www.abc.net.au/news/2026-02-10/thomas-sewell-neo-nazi-camp-sovereignty-melbourne-court/106326988>

Table 2. Categories of anti-First Nations hate in the April–October 2025 dataset.

Type of hate	Category	Description	Frequency	%
Racial identity hate	White Supremacy	Posts implying superiority and supremacy of Whites, settlers, Europeans.	4,408	46.71
	Dehumanisation	Posts dehumanising First Nations peoples and their culture.	4,041	42.82
	Welfare Grievance Framing	Posts mentioning that First Nations are free riders, are a “grievance industry”, exploit aid, pull the “race card”, etc.	361	3.83
	Aboriginality Policing	Posts questioning aboriginality (“not being black enough”).	154	1.63
	Identity-Based Attacks	Posts attacking First Nations public figures for their aboriginality.	103	1.09
	Criminal Stereotyping	Posts associating First Nations peoples with crime, alcohol, or disorder.	15	0.16
Voice-based hate	Voice Delegitimation	Posts delegitimising The Voice and framing it as useless, unimportant, pointless, etc.	48	0.51
	Voice as Racism	Posts suggesting that The Voice referendum was racist or a form of apartheid.	38	0.40
	Other	Residual category capturing related hostile content not covered above.	269	2.85

Focusing on the period surrounding Anzac Day (one week before and after), hate discourse was characterised by a relatively balanced distribution between Dehumanisation (47.12%) and White Supremacy (42.27%), which together accounted for the vast majority of detected content (Figure 3). This indicates that online hate associated with the event combined both explicit assertions of racial hierarchy and more generalised degrading portrayals of First Nations peoples. The prominence of Dehumanisation suggests that contestation around Indigenous recognition within commemorative rituals, such as the Welcome to Country,

often manifested through language that diminished the legitimacy, humanity, or cultural standing of First Nations communities. At the same time, the strong presence of White Supremacy points to narratives that positioned national identity as fundamentally non-Indigenous. Less prominent categories, including Welfare Grievance Framing (4.04%) and Aboriginality Policing (1.71%), indicate additional layers of discourse that questioned entitlement to recognition or authenticity, though these remained secondary to the dominant modes of expression.



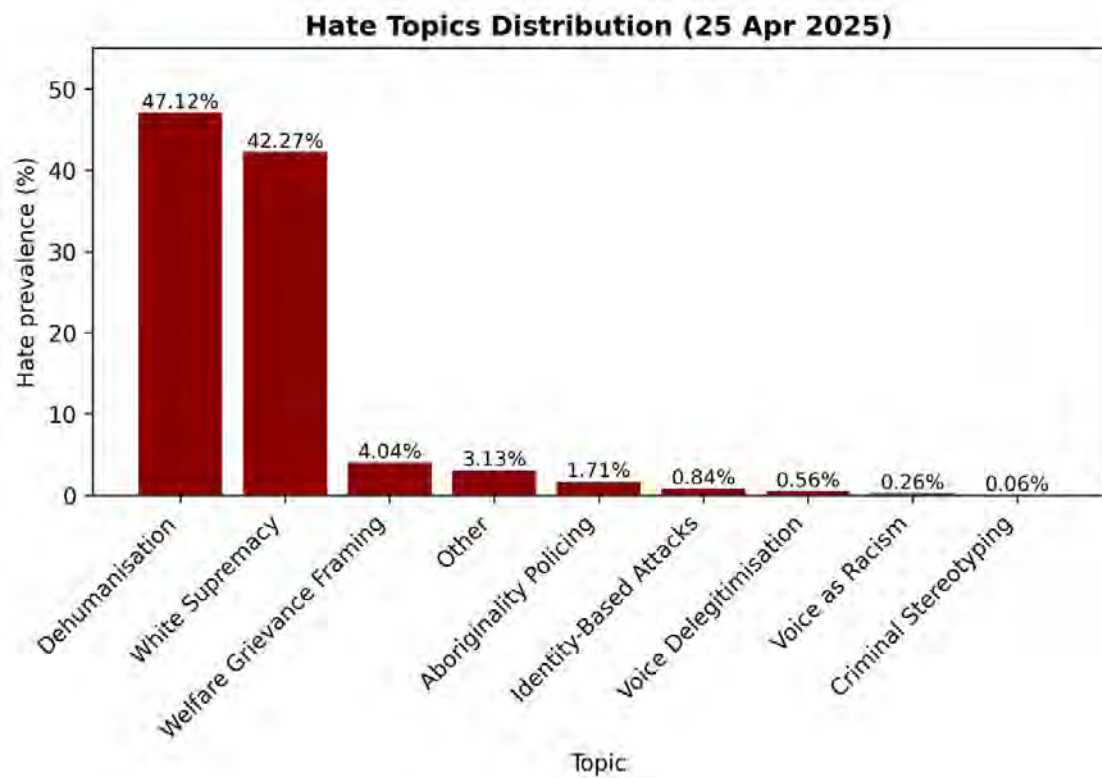


Figure 3. Distribution of hate topics around Anzac Day (25 April 2025).

In contrast, hate associated with March for Australia (one week before and after) was more heavily concentrated within a single dominant frame: White Supremacy accounted for 69.12% of all detected hate, substantially exceeding its share in the Anzac Day discussion (Figure 4). This concentration suggests a more explicit and ideologically coherent articulation of racial hierarchy within discourse surrounding the march. Dehumanisation remained present but at a much lower level (22.11%),

indicating that while degrading representations persisted, they were less central than direct assertions of exclusionary national identity. Smaller categories, including Welfare Grievance Framing, Aboriginality Policing, and Criminal Stereotyping, appeared only marginally, and voice-related themes were negligible. Overall, the distribution reflects a discursive environment shaped more by overt ideological positioning than by diffuse or reactive forms of hostility.

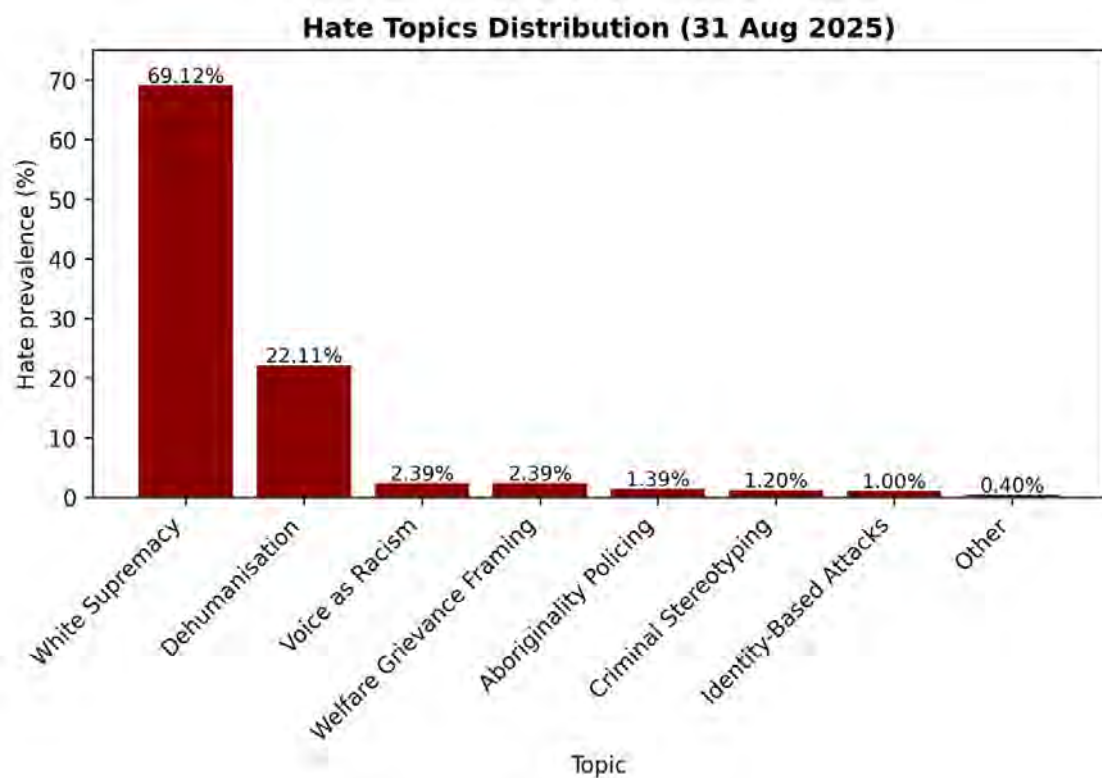


Figure 4. Distribution of hate topics around March for Australia (31 August 2025).

4. SPATIAL AND PLATFORM VARIATION IN ANTI-FIRST NATIONS HATE BETWEEN THE 2023 AND 2025 DATASETS

In the 2023 dataset, voice-based hate is the dominant form across all major platforms and is especially concentrated on high-amplification environments (Figure 5). This content is most prevalent on YouTube (around 30%) and X (around 16%), followed by forums, with more moderate levels on Reddit and blogs. Racial identity hate appears at much lower levels overall but follows a similar pattern, with relatively higher concentrations on forums, YouTube, and X, and minimal presence on other platforms.

In the 2025 dataset, racial identity hate is the dominant form across all platforms and accounts for nearly all observed hate content (Figure 6). This content is most prevalent on YouTube (around 2.4%), followed by X and forums, with lower levels across Facebook public pages, Bluesky, Tumblr, and Reddit. Instagram and blogs show only minimal presence, and news content is negligible. Voice-based hate remains consistently low across all platforms.

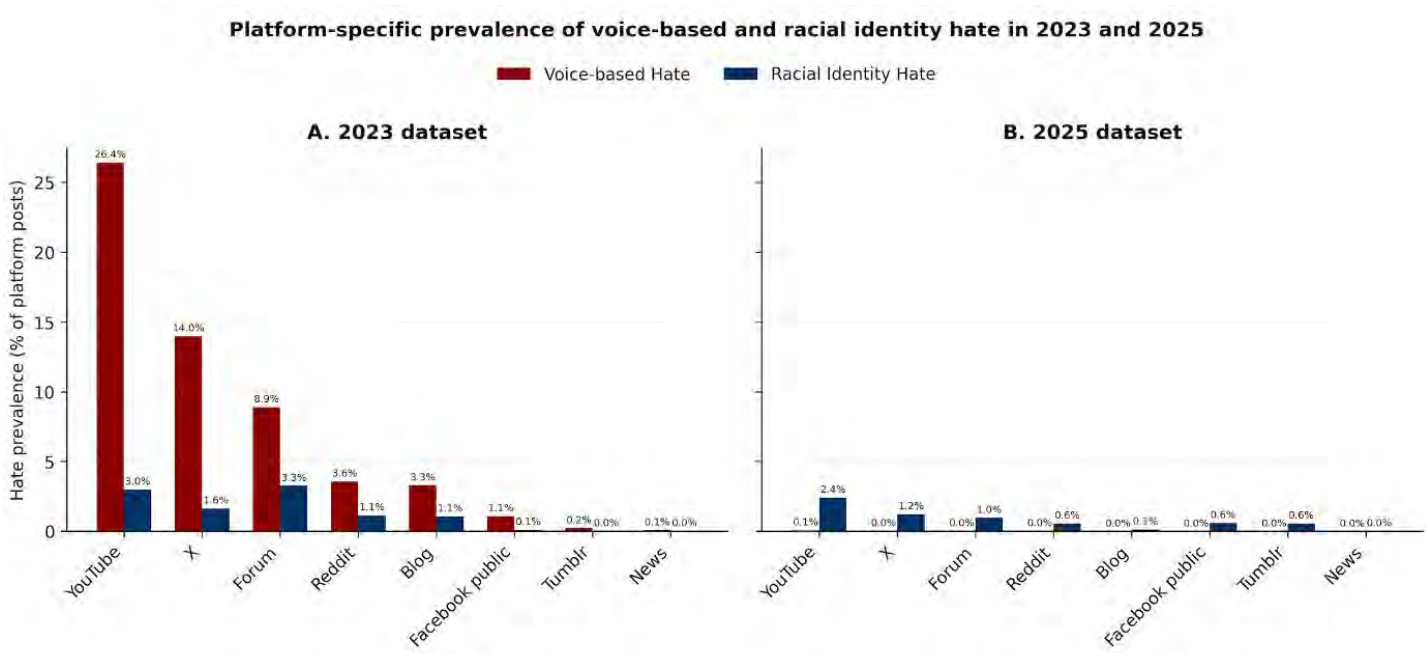


Figure 5. Platform-specific prevalence of voice-based and racial identity hate in the 2023 and 2025 datasets. Prevalence is calculated as the proportion of posts on each platform classified as hate.

As forums exhibit relatively high levels of racial identity hate across both datasets, but include diverse communities, we analysed the top five forums by average daily hate prevalence (Figure 8). In 2023, there is strong variation across platforms. Some forums are dominated by voice-based hate, particularly onlineopinion.com.au (around 25%) and hotcopper.com.au (around 15%), while others show higher levels of racial identity hate, most notably 4chan.org (around 15%). This points to a division between more

structured discussion spaces, where policy-related hostility dominates, and less moderated environments, where identity-based hate is more prominent. By 2025, voice-based hate largely disappears across forums, while racial identity hate persists at lower and more uniform levels. The highest rates are observed on platforms such as hotcopper.com.au and spacebattles.com (around 1.5–2%), with previously prominent sites like 4chan.org declining to below 1%.

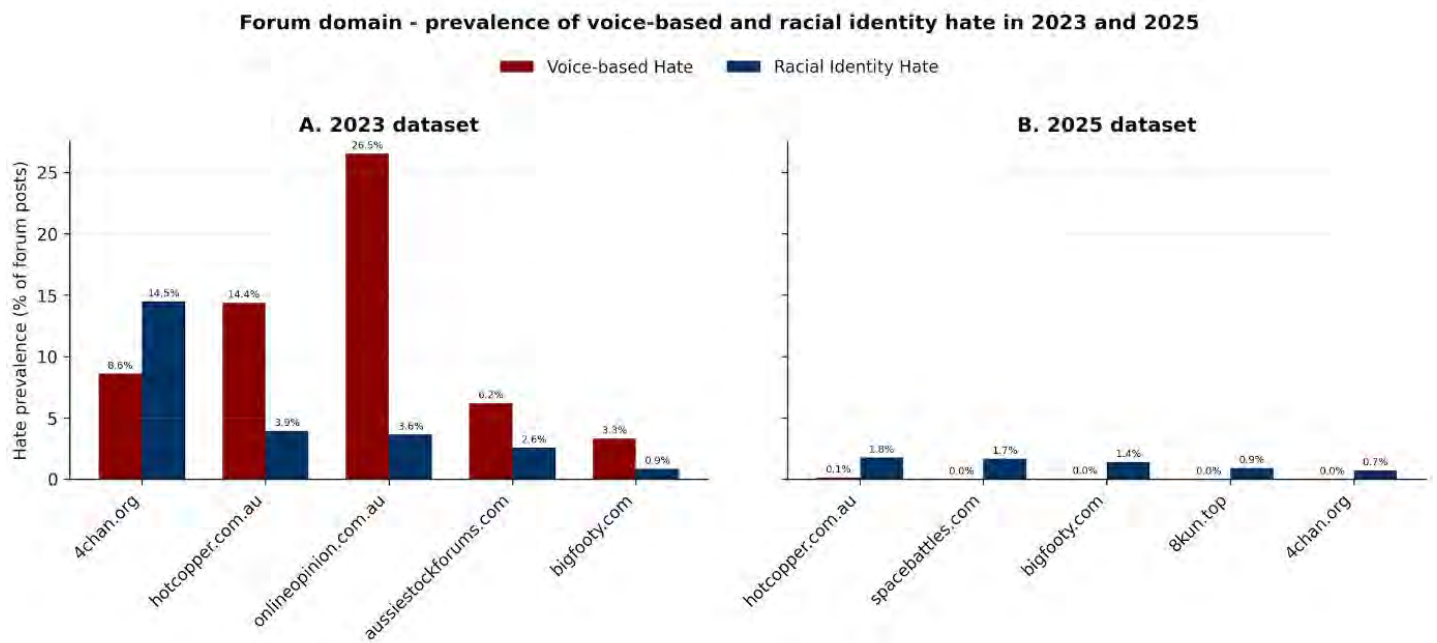


Figure 6. Forum-specific prevalence of voice-based and racial identity hate in the 2023 and 2025 datasets. Prevalence is calculated as the proportion of posts on each forum classified as hate.

In terms of geographic distribution, it is important to note that this analysis is based only on the subset of posts with available geolocation metadata from X. As geolocation data is present in only a small proportion of posts, these results should not be interpreted as estimates of the true

prevalence of hate across states. Instead, they provide a comparative view of how hate activity is distributed geographically within the observed sample. In 2023, hate activity is unevenly distributed across states (Figure 7).

State-level Map

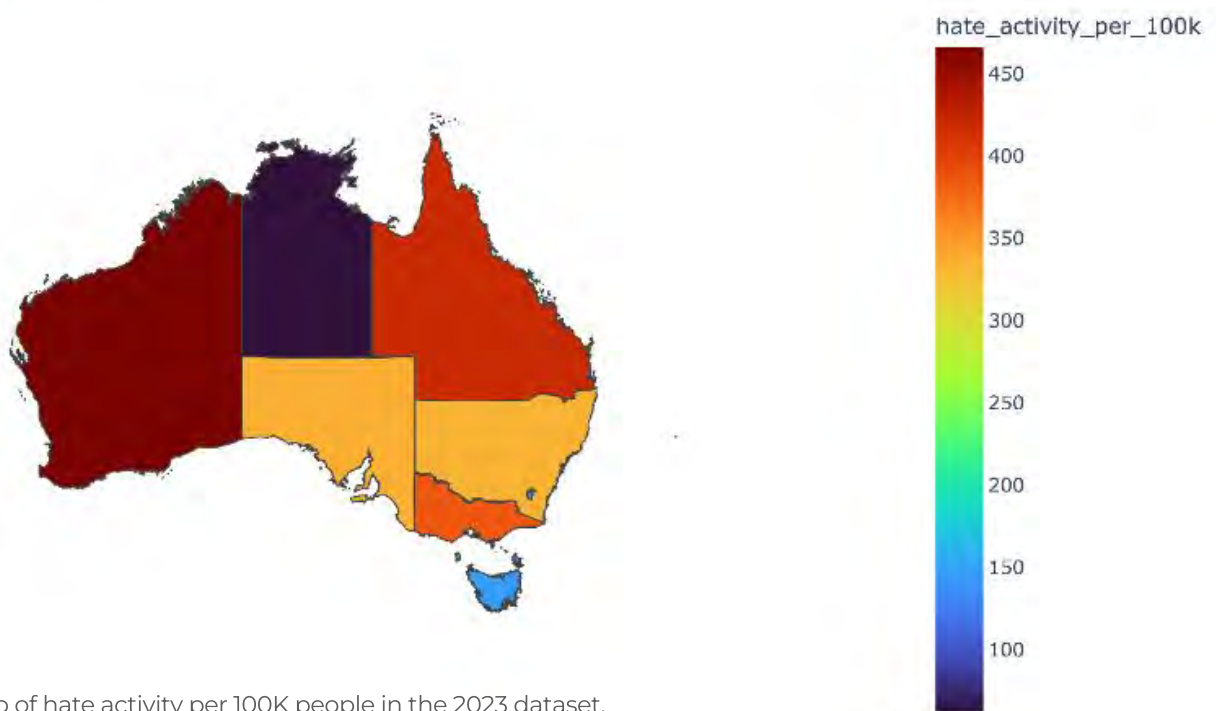


Figure 7. Heatmap of hate activity per 100K people in the 2023 dataset.

The highest relative levels are observed in Western Australia, Queensland, and Victoria, followed by South Australia and New South Wales. Lower levels are evident in the Australian Capital Territory and Tasmania, with the Northern Territory consistently recording the lowest relative activity. This pattern indicates that, within the available geolocated data, hate during the referendum period was geographically concentrated rather than evenly distributed. In 2025 (Figure 8),

overall levels decline substantially across all states, and the distribution becomes more compressed. Differences between states are less pronounced, with most jurisdictions clustering within a similar range. Western Australia and Queensland remain among the higher-activity states, while the Australian Capital Territory and Northern Territory remain lower.

State-level Map

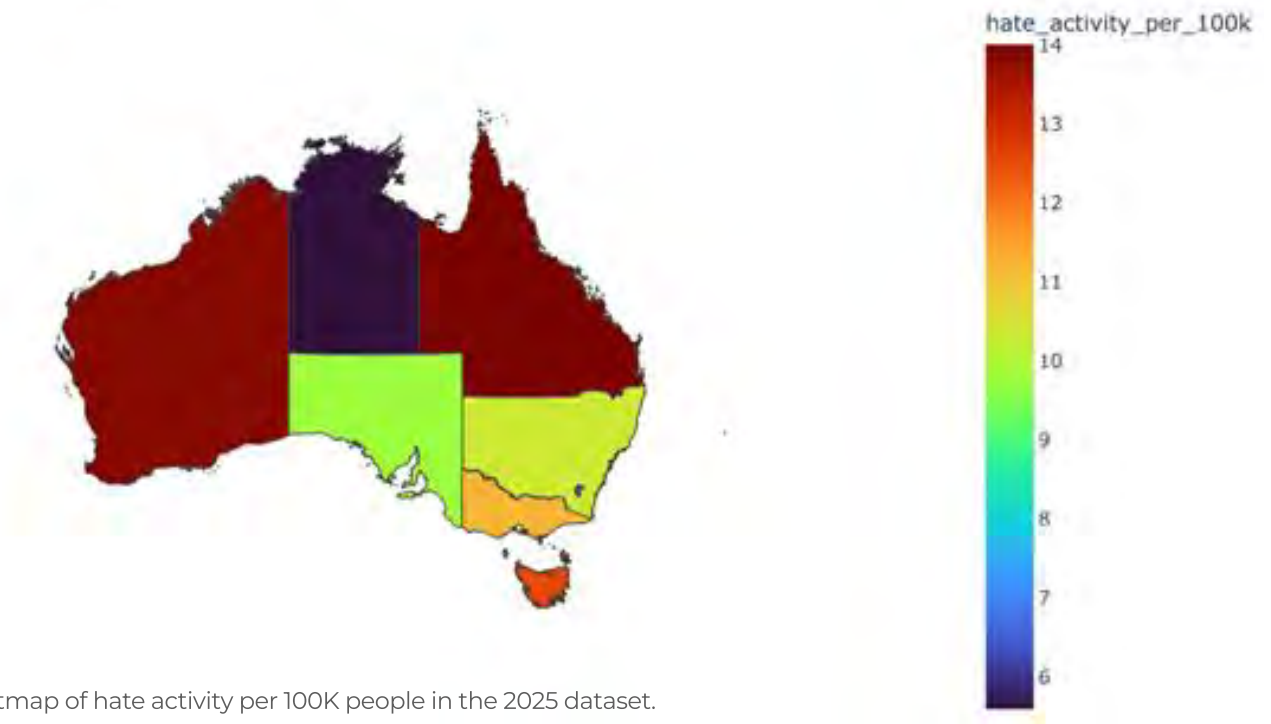


Figure 8. Heatmap of hate activity per 100K people in the 2025 dataset.





CONCLUSION

This report finds that anti-First Nations hate in Australia is highly context-dependent, varying in type, structure, and intensity across different political and social environments. During the 2023 referendum campaign, hate was sustained and cumulative, driven by ongoing mobilisation and media attention. By contrast, the 2025 cases demonstrate more episodic dynamics, with sharp spikes linked to specific events, followed in some instances by short periods of persistence.

A key shift emerges in the type of hate expressed. While the referendum period was dominated by voice-based hate focused on policy and representation, the 2025 dataset is overwhelmingly characterised by racial identity hate. This suggests that event-driven contexts are associated with underlying racialised narratives, moving beyond policy contestation to direct attacks on identity.

These dynamics are closely tied to narrative formation and circulation. Misinformation plays multiple roles, amplifying visibility, intensifying hostility, and reinforcing existing prejudices. At the same time, hate is unevenly distributed across actors and platforms, with First Nations public figures disproportionately targeted and high-amplification environments playing a central role.

5.1. Policy recommendations

The findings point to four priority areas for policy action. First, Australia needs a permanent observatory of online hate to monitor anti-First Nations hate and other forms of targeted hostility across platforms, especially around predictable high-risk political, commemorative, and protest events. Such an observatory should support community-led classification of hate, early identification of escalating narratives, and clear referral pathways when online hate moves towards threats, harassment, or mobilisation.

Second, responses to misinformation should prioritise claims most likely to intensify harm. This includes threat-based and identity-delegitimising narratives that frame First Nations recognition as fraud, unfair advantage, national dispossession, or material risk. Fact-checking should be paired with culturally informed information campaigns distributed through trusted community, media, and platform channels.

Third, institutions should strengthen protections for First Nations public figures, leaders, and community advocates, who experience disproportionate targeting during public debate. This requires practical support, including reporting assistance, threat assessment, legal referral, wellbeing support, and rapid response mechanisms.

Finally, platforms should be held accountable for their specific amplification environments. Transparency, enforcement, and product design reforms should address the reposting, recommendation, and reply dynamics that extend the reach and persistence of anti-First Nations hate.

REFERENCES

Allison, F., Cunneen, C., Whyman, A., Lewis, B. and Selcuk, A. (2025) 'Everywhere I go no matter where, if it's around non-Indigenous people I feel a hate vibe. It feels like I'm being watched'. The Call It Out Racism Register 2024-2025, Sydney: Jumbunna Institute for Indigenous Education and Research, University of Technology Sydney. Available at: <https://callitout.com.au/resources/>

Carlson, B., & Frazer, R. (2021). Indigenous digital life: The practice and politics of being Indigenous on social media (pp. 85-86). Cham: Palgrave Macmillan.

Moreton-Robinson, A. (2015). The white possessive: Property, power, and indigenous sovereignty. U of Minnesota Press.

Paradies, Y. (2016). Colonisation, racism and indigenous health. Journal of population Research, 33(1), 83-96.

AUTHORS BIO

Matteo Vergani

Matteo Vergani is an Associate Professor at Deakin University and Director of the Tackling Hate Lab. His research examines prejudice, hate crime, extremism, and social cohesion, combining social science, psychology, and data science to develop new methods for understanding and addressing complex social problems. He has led numerous large research projects in collaboration with governments and civil society organisations and is Associate Editor of the journal *Terrorism and Political Violence*.

Stephanie Zi Xin Ng

Stephanie Zi Xin Ng Holds a PhD in Engineering at Deakin University's Institute for Intelligent Systems Research and Innovation. Her research focuses on Natural Language Processing, Large Language Models, and computational methods to address social challenges, including online extremism, policy framing, and parliamentary debates. Her recent work also explores multimodal analysis, video and image processing, and computer vision applications.

Michael Lavery

Michael Lavery is a Yamatji man from Geraldton, Western Australia, currently in his final year of law studies. Over the past decade, he has worked with Indigenous communities across Australia, contributing to legal education initiatives and supporting legislative change, including work across Victoria.

Andrea Giovannetti

Andrea Giovannetti is an Assistant Professor in Economics at the Australian Catholic University and a member of the Violence Research Centre at the Institute of Criminology, University of Cambridge. His research applies statistical modelling and computational methods to study the causes and consequences of complex social problems, particularly the diffusion of behaviours across human networks. He has collaborated extensively with police forces and research organisations across the United Kingdom.

Kewen Liao

Kewen Liao is an Associate Professor at Deakin University specialising in data science, machine learning, and algorithm design. His research spans graph mining, streaming analytics, image and text analysis, and artificial intelligence applications with social impact. He has led major research projects in Australia and holds a US patent in computer vision developed with Canon Inc.

Dan Goodhardt

Dan Goodhardt is a researcher at Deakin University with experience in community responses to antisemitism, hate incidents, and prejudice-motivated threats. He previously worked with the South East Asian Network of Civil Society Organisations and has extensive experience supporting victims and documenting hate incidents at a community level. He holds a Master's degree in Counter-Terrorism from Monash University.



www.tacklinghate.org