Accepted for publication in The Sociological Review. Pre-copy edited version.

Learning from online hate speech and digital racism: From automated to diffractive methods in social media analysis

Eva Haifa Giraud (University of Sheffield), Elizabeth Poole (Keele University), Ed de Quincey (Keele University), John E. Richardson (University of Liverpool)

Abstract:

There has been a dramatic surge of big data analytics and automated methods to detect and remove hate speech from social media, with these methods deployed both by platforms themselves and within academic research. At the same time, recent social scientific scholarship has accused social media data analytics of decontextualising complex sociological issues and reducing them linguistic problems that can be straight-forwardly mapped and removed. Intervening in these debates, this article draws on findings from two interdisciplinary projects, spanning five years in total, which generated comparative datasets from Twitter (X). Focusing on three issues that we identified and negotiated in our own analysis – which we characterise as problems of context, classification, and reproducibility – we build on existing critiques of automated methods, while also charting methodological pathways forward. Informed by theoretical debates in feminist science studies and STS, we set out a diffractive approach to engaging with large datasets from social media, which centralises tensions rather than correlations between computational, quantitative, and qualitative data.

Introduction:

Amidst widespread concern about the use of social media to circulate digital racism and xenophobia, there has been a rise in engagements with automated methods to trace, and intervene in, discrimination online. Automated hate speech detection has become central to the moderation of social media platforms (Gorwa, Banns and Katzenbach, 2020). In academic contexts, similarly, automated methods have been used to process 'big data' harvested from social media, as a means of identifying patterns in online racism (for a critical overview, see Nikunen, 2021). Indeed, tools and frameworks used by social media companies are often the product of academic collaborations (X Developer Platform, ND). However, the widespread use of automation has been contentious. As Matamoros-Fernández and Farkas (2021) argue, uncritical uses of automated big data analytics risk reducing complex social phenomena such as racialization and racism to linguistically bounded hate speech that can be easily identified, mapped, and removed. These problems speak to wider debates about the limitations and potentials of using methods originating in computer science to research sociological issues (Gangneux, 2019).

Intervening in these debates, this article draws on findings from two interdisciplinary projects – spanning five years in total – which generated comparative datasets from Twitter (X) of 4,075,153 tweets, to examine the circulation and contestation of Islamophobia. Rather than discussing our findings in and of themselves, however, our focus in this article is methodological, advancing the agenda of 'interface methods' (Marres and Gerlitz, 2016) to explore how tools and techniques from fields such as computer science can be turned to sociological ends. The questions we address in the

paper are two-fold: firstly, what are the limitations of automated methods for researching complex social phenomena such as digital racism and, secondly, what overarching methodological approach could help to overcome these limitations? We answer these questions by drawing on our longitudinal, comparative datasets to identify methodological challenges in using automated methods to identify racialized Islamophobia. However, we do not seek to dismiss automated methods altogether, as some forms of automation are often the only means of organising vast datasets. To this end, we couple our critique with a framework for negotiating these problems that we characterise as a 'diffractive' approach to social media research.

In the main body of the article we, firstly, set out fraught debates about uses of automated methods in social research, particularly research about digital racism. We then turn to our own project, delineating our aims and methods and why we moved away from what we term a 'fractal' approach to analysis (which emphasises similarities between datasets) to a 'diffractive' approach that foregrounds differences and new patterns that emerge when bringing datasets together. The second half of the article elucidates our methodological arguments by identifying three problems encountered in our research, which we describe as challenges of context, classification, and reproducibility that complicate social scientific engagements with automated social media analysis. We conclude by reflecting on the way that a diffractive emphasis, informed by insights from feminist Science and Technology Studies (STS) (Barad, 2007; Thylstrup et al, 2022; see also Marres and Gerlitz, 2015; Marres, 2017), offers a way of critically thinking with and against automated methods, to resist reducing complex social phenomena such as racism to decontextualized linguistic problems.

Automated hate speech detection and digital racism

From the vantage point of contemporary debates, it seems hard to believe that in 2016 – in the wake of events such as Brexit and the US electoral victory of Donald Trump – media scholars were voicing concern about a *lack* of research into far-right uses of digital media (e.g. Mercea, Ianelli and Loader, 2016). Scholarship about far-right media practices did exist prior to this period (see Atton, 2006). Valuable research was produced about elements of digital culture that were a precursor to the media logics of the contemporary populist right, including separatist militias (Castells, 1997), the conservative blogosphere (Shaw and Benkler, 2012), far-right online networks (Caiani and Wagemann, 2009), and trolling cultures (Phillips, 2015). A slightly different body of scholarship focused on far-right communication ecologies beyond the digital, from the circulation of political pamphlets to letter-writing in newspapers (Richardson and Franklin, 2003; Richardson, 2008).

More recently, however, there has been a rise in research about far-right media and online racism that needs to be contextualized both socio-historically and methodologically. Not only has there been commitment to redressing a lack of research into ethno-nationalism in the wake of its resurgence, but – for a brief window – far-right media use had extended beyond digital enclaves to commercial platforms (Marwick and Lewis, 2017). These developments are methodologically significant in light of the sociological turn to big data (Savage and Burrows, 2007), which was itself intensified by the popularization of platforms such as Twitter as data sources (Tinati et al, 2014). As a result of these complex, mutually-reinforcing, developments, it became easier to research formerly hard-to-access far-right communities using data from commercial social media platforms, rather than requiring backstage ethnographic access.

Matamoros-Fernández and Farkas's (2021) systematic review of hate speech scholarship, for instance, traces a surge of research about online racism from 2016 alongside shifts in methodology and focus: from qualitative research with social movements, to automated analysis of large social media datasets that focuses on linguistic content. In particular, there has been a rise in data science

research that deploys machine learning to detect online racism, such as automated sentiment analysis which draws on pre-existing dictionaries that attribute value to particular words that are then combined to calculate overall semantic meaning (Ligthart et al, 2021). Matamoros-Fernández and Farkas are highly critical of these shifts, arguing that identifying hate speech via keywords 'tends to reduce racism to just overt abusive expression to be quantified and removed' (2021: 216).

Nikunen (2021), likewise, draws attention to the risks of assuming automated methods are a neutral tool for identifying racist content, drawing worrying parallels with Zuberi and Bonilla-Silva's (2008) critique of 'white methods' in sociology. That is, social scientific methodologies that present themselves as neutral data collection methods while occluding (and reproducing) racial biases that underpin the origins of these methods and hence the gathering, analysis, and framing of data. Nikunen underlines that '[t]his does not mean that big data is automatically "white method"'. Instead, dangers can arise when these methods 'adhere to contexts (technology studies and industry) and practices (de-contextualization, correlation) that can consolidate existing ideology of colour-blindness and be incapable of addressing issues of race and racial inequality' (Nikunen, 2021: 3). When automation is treated as a neutral black box, for instance, this obscures the laborious annotation, decision-making, and sorting processes that constitute training datasets for machine learning tools (Thylstrup et al, 2022). As scholarship from critical race studies underlines, dataset composition is never neutral, thus, as machine learning techniques are applied to new datasets, they replicate any biases and exclusions inherent in their training data (Benjamin, 2019).

The implications of these biases for research about digital racism are that, if keywords are labelled in a manner that lacks context, automated keyword analysis might inaccurately label (or indeed not label) something as racist. Siapera (2019), for instance, illustrates that when social media companies themselves have relied on automated keyword analysis to identify hate speech, phrases such as 'white man' have often been flagged as hateful while coded racism persists as 'legitimate debate' (Siapera, 2019). Siapera contends, however, that challenges facing automated methods are not simply technical problems of classification, but speak to wider epistemological debates about what constitutes digital racism. The problem is that a large volume of racist online content does not consist of overt abuse, explicit far-right content, ethnic slurs, or hate speech, but everyday 'banal' expressions of racism that are difficult to detect through linguistic or semantic analysis (Siapera, 2019).

It is the everydayness of large volumes of digital racism that lead Sharma and Brooker (2016) to argue that it cannot be treated as purely a linguistic phenomenon and instead is an assemblage of 'the human (social media users), social phenomena (race and racism), and the nonhuman (digital technologies and devices)' (2016: 3-4). What this means, in the case of Twitter, is that racism is not necessarily a property of the language of individual tweets; instead, 'modes of racialization emerge within and across tweets' (30). In other words, even if there are no obvious linguistic markers of racism in a particular tweet, user practices – such as hashtags, quote tweets or shared jokes – can situate a post within a particular set of relations and social contexts that materialize racist meanings. Thus, dealing with bias in automated language detection might not get to the root of digital racism as conceived as a more complex, situated, and relational phenomenon.

In addition to identifying the limitations of automated methods, critical literature has also made recommendations for ameliorating these limitations. These recommendations include grounding analysis in concepts from critical race studies (Matamoros-Fernández and Farkas, 2020) and recognising cultural biases in determining what constitutes 'hate' (Thylstrup and Talat, 2020). Perhaps the most prominent recommendation, however, has been to embrace mixed-methods approaches and combine computational analysis with qualitative research. Indeed, the value of

mixed methods approaches is well established in social media research more broadly, as with influential scholarship on feminist and anti-racist activism, which has combined large-scale data analytics with qualitative analysis of specific tweets to understand the nuances of online counter-publics (Jackson, Bailey, and Foucault Welles, 2020).

Insights about how to navigate the limitations of automated methods informed our initial research design, where we sought to trouble the primacy of computational analysis by ensuring automated methods were always triangulated with qualitative research. Yet, this approach still risked over-emphasising similarities between datasets and treating qualitative analysis simply as a layer of verification. In other words, it risked pressing qualitative methods into the service of big data rather than asking what novel insights they could provide in their own right and how these insights might enrich or even complicate automated analysis. Drawing on our own data, here we elucidate problems that arise if qualitative methods are used purely to corroborate automated data analytics and argue for the value of a diffractive approach for overcoming these problems. However, before doing so it is necessary to provide further detail about our research itself.

Background to our own research

The research that underpins this article is a collaboration between colleagues from media studies, computer science and sociology, which consisted of a smaller pilot project and larger multi-year study that focused on the emergence of online counternarratives against Islamophobia. Our primary aim was to understand the conditions that enabled successful contestation of Islamophobic hate speech, and how these conditions had evolved over time. To this end, we combined automated computational analysis with quantitative and qualitative content analysis of smaller data samples to examine the circulation – and contestation – of Islamophobia online. While this was interdisciplinary research methodologically, and in terms of the disciplinary identity of core team members, our objectives were sociological. Our research questions, for instance, sought to (i) identify the relationship between Islamophobic narratives and counter-narratives on Twitter and who was participating in these counter-narratives; (ii) trace the transnational dynamics of (counter-)narratives and how the actors involved appropriate global events to support their perspectives; and (iii) examine the relationship between social media narratives and other media platforms (for more detail see Poole et al, 2023).

These questions informed our pilot study, which focused on counter-narratives against a single hashtag, #StopIslam (examining 302,342 tweets in total). This hashtag emerged in the wake of the 2016 Brussels bombings, an Islamist attack that led to a surge both in Islamophobia and counter-narratives to contest it. Despite happening in a European context, we found that narratives surrounding the attacks were transnational and leveraged by activists in other national contexts in support of far-right agendas; this, in turn, resulted in equally transnational counternarratives that attempted to contest Islamophobia (see references removed).

Our larger project drew on a corpus of 3,772,811 tweets purchased from Twitter, compiled from keywords we submitted to the platform. These keywords related to what Awan (2014) describes as 'trigger events', newsworthy incidents that are leveraged on social media to create visibility for Islamophobic hate speech. Purchasing these tweets, rather than scraping them independently, allowed us to access all tweets relevant to our search terms that were available on Twitter at the dates of purchase. The project focused on three high-profile events: responses to the 2019 white supremacist Christchurch attack; events leading to the day of the UK leaving the European Union (so-called Brexit Day, 31st January 2020); and flashpoints during the Covid-19 pandemic (including narratives emerging during Eid-al-Fitr and Eid-al-Adha).

We selected these events because they enabled us to answer our RQs due to possessing three characteristics that we identified as requiring further research in the pilot: the entanglement of Islamophobic narratives with counter-narratives; transnational dynamics and the politics of appropriation (wherein events in particular national contexts were appropriated for political agendas in other contexts); and flows of content between Twitter and 'mainstream' media. We considered a range of examples (e.g. the London Bridge Attacks) that were disregarded due to being smaller scale or lacking one or more of these characteristics. Our successful grant bid ultimately specified that we would focus on Brexit and the Christchurch attacks, due to possessing the three characteristics that answered our questions, and that we would select a third case after beginning the research and developing a sense of which other, more recent, events shared these traits. We selected Covid-19 in July 2020 after discovering the spread of the virus was also accompanied on Twitter by the same three characteristics: it was being appropriated in support of Islamophobic narratives that were challenged by persistent counter-narratives; tweets had notable transnational dynamics; and content from Twitter was engaged with in mainstream media outlets.

Our search queries were developed in consultation with our advisory board, which contained stakeholders from NGOs who conducted research into Islamophobic media representation as well as academics working on Islamophobia, mediated activism, digital racism, and white supremacism. The search terms included terminology related to Islam alongside alternative spellings and coded phrases that are frequently used by the far-right, such as 'religion of peace' and the deliberate misspelling of Muslim, which were combined with event-specific keywords (see fig.1). After initial automated keyword analysis to determine whether any prominent hashtags were used in relation to these events, we purchased an additional six weeks of data related to two hashtags: #hellobrother (from the Christchurch dataset) and #tablighijamat/#tablighijamaat (from the Covid data).[i]

Fig 1. List of search queries

Core search terms: Islam* OR Muslim* OR Moslem* OR 'Religion of Peace' OR mosque
AND
Event-specific terms: Brexit; Christchurch OR New Zealand; Coronavirus OR Covid; Eid
Hashtags: Christchurch OR New Zealand AND #Hellobrother; Coronavirus OR Covid and #tablighijamat OR #tablighijamaat

We purchased six weeks of tweets related to each event so that we could gain insight into what happened after short-lived trending topics had peaked and the user networks around them had dissipated (see fig.2).

Fig. 2: Sample of Tweets Generated by Keywords

Event	Date Ranges	No of Tweets	Total No of Tweets	Quantitative sample
Brexit	28 Nov 2019 – 19 Dec 2019 17 Jan 2020 – 07 Feb 2020	26,473 16,061	42, 534	1000 1000

Christchurch Terror Attack	15 Mar 2019 – 15 Apr 2019 15 Jun 2019 – 21 Jun 2019 15 Sep 2019 – 21 Sep 2019	3,099,138 8,072 2,870	3,110,080	1000 500 500
#Hellobrother	As above	25, 084	25, 084	1000
Coronavirus	19 Mar 2020 – 19 Apr 2020 19 May 2020 – 25 May 2020 29 Jul 2020 – 4 Aug 2020	433,574 119,700 28,097	581, 371	1000 500 500
#Tablighijamat/ #Tablighijamaat	As above	13, 742	13, 742	1000
Total			3,772, 811	8000

Initial data from Twitter was provided to us via the platform's API v.1.1 retrieval process as Javascript Open Notation (JSON) files but when we submitted our second file request Twitter had changed the API to v.2.0, which meant the files were in slightly different formats. To ensure parity, a computer scientist in our team created code in Python, in Jupyter Notebooks, to reformat the data before applying an analytic process that identified features including: which tweets had the highest number of re-tweets and shares; commonly used words, emojis, collocations and hashtags; biographical details about users; and most active users in the datasets. As part of this process we also converted the data that Twitter had originally sent us into Excel files that could be more easily navigated. These files included the text of tweets we purchased, contextual detail gathered from the computational analysis, and links to the original tweets. To map retweet networks, bespoke code was then developed based on Asturiano's (2022) force graphs application.

We were aware, however, that the linguistic orientation of some of the automated techniques we used to identify keywords and other recurring features was not straight-forwardly compatible with our constructivist epistemology, which understood race, racism, and racialization as emergent through complex socio-technical assemblages rather than fixed properties (Sharma, 2013; Sharma and Brooker, 2016). This understanding is especially important in the context of Islamophobia because, despite Islamophobia having a long history of being racialized through orientalist discourses that 'Other' Muslims (Poole and Williamson, 2021), the unstable distinction between 'race' and religion is often leveraged to deny that anti-Muslim statements are racist (Hafez, 2014). Thus, we had to reflect carefully on how to combine digital methods with qualitative analysis to ensure we did not undercut more complex conceptions of digital racism by reducing it to rigid models of hate speech.

To avoid treating our computational analysis as a transparent representation of counter-narratives, in the second stage of analysis we corroborated our findings through quantitative content analysis on SPSS, classifying tweets using a coding schedule developed from an initial reading. Through this method we analysed the top 1,000 retweets from our largest sets of Twitter data associated with each event, and 500 from smaller datasets (see again fig. 2). Our samples for quantitative analysis simply selected the posts in each dataset that had the highest numbers of retweets, to enable us to analyse tweets that gained the most traction in online narratives. For consistency, the analysis was conducted by one researcher, however, the whole team met following initial coding to discuss if the

coding schedule needed adjustment, whether new categories needed to be added to the schedule, and if there was consensus on how particular tweets had been coded.

Finally, three team members conducted a much smaller qualitative analysis that examined the language-use and semantic meanings of the 50 most retweeted tweets in each dataset, alongside accompanying replies. The analytic process involved following hyperlinks from the Excel files, which led to the original tweets (as they appeared on the platform). This enabled us to analyse tweets in context and examine language-use, media linked to/embedded in tweets and mode of argumentation. Our examination of comments was in recognition that our sampling strategy of focusing on highly-visible retweets for both the quantitative and qualitative analysis had privileged accounts with large follower-counts (often celebrities or politicians). In contrast, the comments included a broader range of users. Thus, to capture more everyday interactions, we examined the dynamics of debates in the comments, with a focus on whether other users agreed with or contested the original tweets, uses of evidence to support arguments, and recurring tropes/motifs in discussion. This stage of the project involved the most intensive series of meetings, where team members conferred on a fortnightly basis about how to interpret findings.

Although in this article we predominantly draw on findings from our larger project, '#ContestingIslamophobia', at times we touch on materials from our pilot study. Here, again, we combined computational analysis, quantitative content analysis, and qualitative content analysis for the purpose of triangulation (for more detail, Poole et al 2019, 2021). As we began to analyse and discuss our findings, however, we began to ask more critical questions about the relationship between our methods.

From 'fractal' to 'diffractive' conceptions of computational social science

When we had our first project meeting to discuss our full datasets, one of our team-members stated:

> It's almost as though there's a fractal relationship between our qualitative and quantitative data; the qualitative data is a microcosm of the patterns identified in the big data analysis. (Notes from project meeting, 25.5.22)

What our colleague was referring to, in metaphorically describing the data as 'fractal', was the way that our datasets appeared to say broadly the same thing. While our qualitative and quantitative analysis focused on smaller subsets of the most re-tweeted tweets, these datasets seemed to be a synecdoche of the big data that we had processed and visualized.

As we deepened our analysis further, however, the challenges we identified reshaped how we understood the relationship between methods. We realised that although our use of qualitative analysis was, in part, in recognition of the need to adopt a critical stance towards digital methods, we were ultimately still treating our content and textual analysis as methodological technofixes to overcome the limits of computational analysis. As we have outlined above, the danger of trying to resolve complex problems with technical solutions is that it can extract what is a context-specific, relational phenomenon – which is mediated by the situated affordances of particular platforms – into something that can neatly be abstracted from these relationships.

Rather than treating qualitative methods as a means of corroborating computational analysis, we suggest that a productive focus for research on digital discrimination and racism – and indeed uses of digital methods in social research more broadly – is to ask how qualitative analysis complicates big data processed through automated means.

This approach, to borrow a contrasting physics concept from STS scholar Karen Barad (2007), is more akin to a 'diffractive' as opposed to a fractal understanding of digital methods. Diffraction is a process that describes what happens when two sets of light-waves meet one another (Barad, 2007: 76-86). These waves complicate one another and create an entirely new configuration (or 'diffraction pattern'): akin to ripples in a pool that generate a new pattern when they come together (83). Arguing that diffraction can be productively applied to interdisciplinary knowledge production more broadly, Barad centres the novel dynamics that emerge when knowledges *complicate* rather than *complement* each other. In the context of big data research on hate speech, we argue, focusing on the ways that qualitative methods complicate computational analysis is important in resisting decontextualization, overly neat classification, and reductive approaches to reproducibility. More broadly, this orientation – to draw on Fitzgerald and Callard's (2015) terminology – moves beyond ebullient celebration or overly hasty dismissal and critique of interdisciplinary collaboration. Fitzgerald and Callard make this argument in relation to collaborations between neuroscience and social science, but we argue it is equally applicable to social scientific engagements with tools and techniques originating in computer science. In the second half of this article, we illustrate these arguments by drawing on our own data.

Analysis: contexts, classification, and reproducibility

We now turn to our own project(s) to elucidate challenges in using automated digital methods that we identified through the course of our own research. We begin each section by critically outlining how these challenges are currently framed in scholarship, where they are often positioned as technical problems to be solved.

Contexts

As foregrounded by scholarship that conceives digital racism as a complex socio-technical assemblage (Sharma and Brooker, 2016), an issue facing automated hate speech detection is that – although overt slurs can be straight-forwardly identified – 'banal' racism is difficult to detect, with veiled or coded language used to circumvent moderation or deny prejudice (Siapera, 2019). To an extent, these challenges can be overcome with the necessary expertise (as with our own search queries incorporating white supremacist terminology after consultation with experts in this field). Linguistic markers of discrimination can nonetheless be difficult to track because they change as socio-political environments evolve and vary across geographical contexts.

Evolutions in hate speech are illustrated by contrasting our more recent datasets with our pilot project, which focused on #stopIslam. Initially, the hashtag was utilized by far-right activists in Europe who deliberately tagged alt-right influencers from the US; these users, in turn, leveraged the bombings as evidence of the 'failure' of multiculturalism, to support anti-immigration rhetoric in Donald Trump's presidential campaign. In comparison with more recent datasets, what stands out in the pilot is the presence of overtly racist and Islamophobic content. The profiles of users sharing #stopIslam often contained symbols of US ethno-nationalism, such as bald eagles, images of the twin towers captioned with 'never forget', and stars and stripes banners. Highly retweeted posts included images of crusaders captioned 'all united against Islam', a bingo card listing Islamophobic tropes, and hashtags such as #NoRapefugees. Qualitative analysis of comment threads, moreover, illustrated coordinated attempts to post disinformation and far-right memes in response to users perceived to be Muslim, or who attempted to condemn and contest the Islamophobic intention of #stopIslam, with these users labelled (derogatorily) as 'liberals' or 'social justice warriors'.

However, the proportion of US-focused tweets and users participating in overtly Islamophobic narratives was diminished in our later datasets in the wake of changing moderation policies. After

the 2020 Capitol attacks in Washington D.C., there was a 'purge' of Twitter accounts associated with the riots, including President Donald Trump (who was deemed to have violated their 'Glorification of Violence' policy) and 70,000 Twitter accounts associated with conspiracy group Q-Anon (Twitter Safety, 2021; Twitter Inc. 2021).

This widespread deletion of users was evidenced in our datasets, as we began the process of qualitative analysis. As described in our methodology, our qualitative analysis involved creating Excel files of tweets purchased from Twitter, where we could read the text of any tweet relevant to our keywords that was available on the date of purchase (as well as other characteristics such as the location and biographical details of users), before following associated hyperlinks to view the tweets in their original context. When following these links, we found that the most extreme tweets in our files were from US-based accounts that had since been suspended. Likewise, in comment threads below counter-narrative tweets we found that large numbers of posts had been deleted; suggesting the removal of users who typically engaged in coordinated attacks to normalize Islamophobia. While these findings could be indicative of successful moderation policies, any such conclusion was complicated by two factors that emerged through our qualitative analysis: the persistence of 'banal' racism, and geopolitical unevenness in how moderation policies were implemented. Both of these issues were evident in our Covid data.

The persistence of banal racism came to the fore in July 2020, when there was a lockdown in North-West England the day before Eid-al-Adha that resulted in a dramatic spike in tweets from the UK. Two days after the announcement, a member of parliament in the region stated in a radio interview that 'sections of the community are not taking the pandemic seriously' and, when asked by the interviewer if he was referring specifically to 'the Muslim community' replied: 'of course, it is the BAME communities that are not taking this seriously enough' (Wilcox, 2020). While the interchangeable use of 'Muslim' and 'BAME' in this exchange speaks to the long history of Islam being racialized in the UK, online responses to the interview speak to context-specific and subtle ways that essentialising narratives of 'cultural difference' were normalised in debates about Covid. It is precisely the context-specific nuances of these narratives that risked being lost without reflecting on ways that our different methods complicated – rather than solely complemented – one another.

Computational analysis, which had identified the most retweeted posts, showed that counter-narrative tweets dominated our data, thus indicating 'successful' contestation of racism and Islamophobia. This conclusion seemed to be corroborated by qualitative analysis of engagement with high-profile posts. Three members of the UK opposition party posted from their personal Twitter accounts to condemn the MP's statements, with the most popular post retweeted over 1500 times and liked 10k times. The largest proportion of comments underneath this tweet (122, out of 193) agreed with the condemnation of Islamophobia. However, more careful examination of the comments beneath counter-narrative tweets forced us to shift focus away from how our datasets complemented one another to a diffractive exploration of how these comments complicated any straight-forward interpretation of this example.

The tactics used in counter-narrative posts were dominated by a homogenous series of images. Locations filled with white-majority crowds (such photographs of busy beaches and pubs, or VE-Day anniversary street parties from May 2020) were repeatedly used to contest the pandemic's racialization. The homogeneity of these images fails to contest the premise of the original narrative, which had suggested racialized minorities were somehow responsible for spreading Covid, because the only counterpoint it offers is that social distancing wasn't being adhered to by other people either.

In contrast, posts that criticised the original counter-narrative tweets while agreeing with the original radio interview constituted a minority of comments (71, out of 193), with a still smaller number including more overt expressions of racism (such as comments that asked why the lockdown cities were all in 'BAME areas' and references to 'balm people' [sic]). However, unlike counter-narrative tweets – which failed to disrupt the underlying premise of the original narrative – tweets that expressed banal racism and Islamophobia made active efforts to shift the discursive context. More specifically, these tweets echoed a trend, identified in contemporary scholarship, of everyday interactions on social media being used to expand the window of what counts as legitimate debate, as a means of normalising racism (Siapera, 2019; Titley, 2020). These tweets, for instance, praised the MP for speaking the 'truth' and being unafraid to share 'facts'. Attempts to recast racist and Islamophobic stereotypes as legitimate public health concerns were evidenced through posts using the language of 'cultural difference' or pointing to 'crowded housing' and 'busy mosques' as factors that needed consideration in mitigating the spread of Covid.

These tweets, therefore, illustrate both the challenge of identifying hate speech and the risk of conflating hate speech with racism. Here, pandemic-specific terminology (such as 'social distancing') was turned to racist and Islamophobic ends, by being yoked to longstanding tropes of cultural difference to infer that 'certain communities' were not distancing. Yet the ubiquity of phrases such as social distancing (especially during the pandemic) means that they are not hate speech in and of themselves, and only understood as such if contextualised in relation to both immediate events and longer cultural discourses of Othering. This mode of racism offers especial risks for automated detection based on keywords, if they do not incorporate socio-historical and discursive contexts.

The challenge of identifying Islamophobia, however, was not just the evolution of linguistic markers over time, but across space. Aside from the aforementioned spike in tweets from the UK prior to Eid-al-Adha, our datasets on Covid were dominated by Indian accounts (particularly the first sample from March 2020), and featured a significant number of tweets and hashtags that associated the spread of the pandemic in India with Muslim minorities. While overt expressions of Islamophobia had been removed from US accounts in the wake of more stringent moderation policies, the politics of deletion was geopolitically uneven and tweets from India within our datasets included support for Hindutva extremism and direct hate speech. The hashtag #tablighijamaat, for instance, trended in the wake of accusations that a transnational Sunni gathering had been a super-spreader event, which led to disinformation that Muslims were deliberately spreading Covid-19 (Ghasiya and Sasahara, 2022).

What was most prominent, however, was visual disinformation. Commonplace memes reinforced stigmatizing narratives, such as a cartoon of two men in Islamic dress who were hugging each other in greeting, while one said 'Eid Mubarak' and the other stated 'Covid Mubarak', or the widespread use of 'coronajihad' in tweets and memes. A prominent social justice activist, moreover, received (seemingly) coordinated and repetitive responses to a tweet where they had criticized Islamophobic statements made by a minister from the ruling Bharatiya Janata Party (BJP), who claimed that the replacement of a mosque with a Hindu temple would reduce Covid. In response to their criticism of Islamophobia, this user received 50 responses and 41 quote tweets, which included explicitly racist content (such as images of excrement alongside the slogan 'sacrifice children, not animals').

The evolution of hate speech across – and within – our datasets foregrounds the risks of automated methods. Firstly, it highlights a challenge for platforms themselves: the need to be constantly attuned to coded language that evolves across time and space. Secondly, and more pertinently to our central arguments in this paper, our findings underline the dangers of relying solely on computational analysis without sufficient context. While some of our qualitative findings correlated

neatly with computational and content analysis, others complicated understandings of how racialized discrimination was enacted and contested. Relying purely on statistics about the most retweeted posts, as corroborated by the dominance of counter-narrative posts, obscures the context-specific markers of Othering that are difficult to detect via language that is abstracted from this context. These processes of Othering, however, are precisely what extend and normalise racist tropes as legitimate debate.

Classification

When developing automated methods for detecting hate speech, the challenges we have outlined above – shifting contexts, veiled language, culturally-specific jargon, and visual disinformation – are often treated as technical problems that can be overcome through the development of more sophisticated tools. However, one of the most prominent difficulties in accurately detecting hate speech relates to a different challenge: the risk of misclassifying and deleting 'legitimate' content. For social media platforms, the risk of misclassification is an economic problem. A growing body of academic research has foregrounded how the problem of misclassification relates to wider issues of racial bias in the context of machine learning that have important implications for social research. Observations about racial bias in training datasets are well-established in fields such as media studies and STS (Noble, 2018; Benjamin, 2019). These arguments are underscored by computer science research, which has foregrounded how particular dialects – such as African American Vernacular English (AAVE) – are more likely to be flagged as hate speech by automated tools because of the incorrect labelling of training datasets, due to a lack diversity in annotator teams (Sachdeva et al, 2022; Harris et al, 2022).

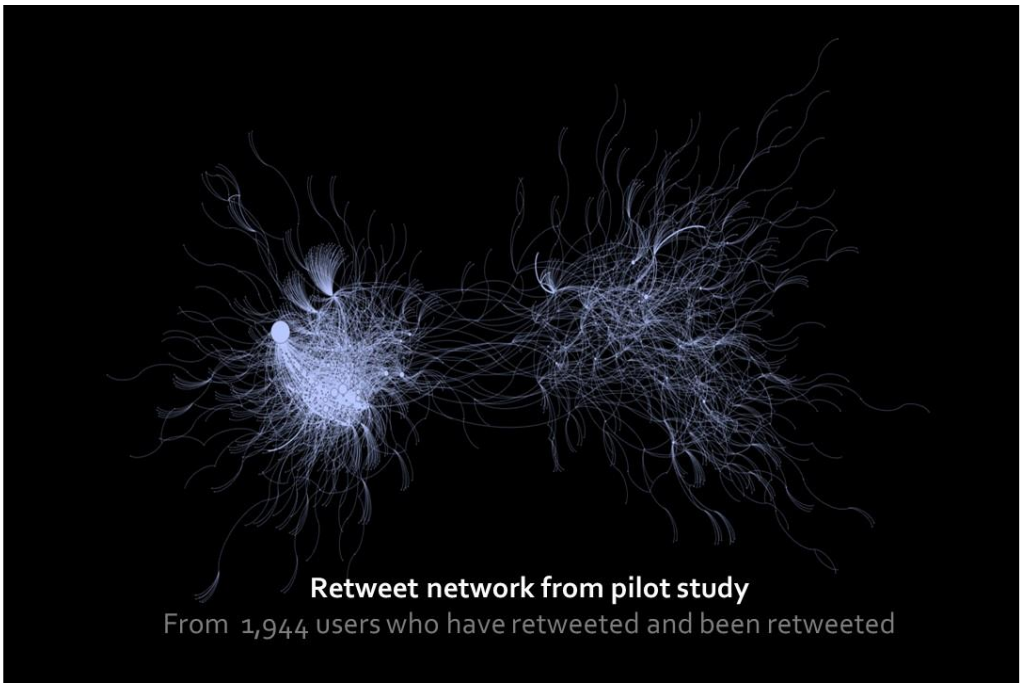
On one level, qualitative approaches could be understood as a tool to mitigate the problem of bias by adding the necessary context to nuance interpretation. As we outline below, counter-narratives against hate speech are particularly resistant to accurate classification, due to being entangled with their object of critique, but careful qualitative analysis can be invaluable in teasing out meaning. We make this point, however, not to underline the need for more accurate training data, or even to suggest that qualitative methods offer a magic bullet for ameliorating biases. Instead, we problematize the idea that neat divisions can be made between hate speech and attempts to contest it, as such distinctions rely on extracting tweets from their relationships with one another. Again, in order to reveal co-constitutive relations between narratives it is valuable to centre questions of how research methods and datasets complicate rather than just complement one another.

Counter-narratives are difficult to classify because the 'ad hoc' publics that articulate them (Bruns and Burgess, 2011) tend to coalesce in relation to affectively-charged events such as protests, political controversies, or terrorist attacks (Papacharissi, 2015; Lee and Lee, 2023). In our data, for instance, the most significant counter-narrative against hate – in terms of volume – was found in our Christchurch data, which was the largest of our datasets by a significant margin. Our first sample, taken between 15/3/19-15/4/19, contained 3,099,138 tweets about the event, the majority of which (73%) condemned the attacks. Yet, while the dominant response to Christchurch was an affective counter-narrative against Islamophobia, which contained complex expressions of solidarity (Richardson et al, 2024), this type of content is often classed as 'negative' by automated tools such as sentiment analysis. The reason why a counter-narrative against hate would be mis-classified is because typical language used to signify critique or condemnation – as represented in our dataset by language such as 'hateful', 'heartbroken' or 'horrified' and emojis depicting anger, crying faces or broken hearts – tends to be labelled 'negative' in sentiment analysis training data.

This problem of misclassification initially arose during our pilot project's examination of #StopIslam. We trialed using sentiment analysis to distinguish between Islamophobic narratives and critiques of Islamophobia but found that language classified as 'negative' was often anti-racist, such as descriptions of racialized Islamophobia as 'horrible', 'disgusting', or 'hateful', or emojis depicting anger. This is not to say that applications of sentiment analysis are not valuable in some settings, especially if care is taken to adjust word values to account for the specificity of datasets. Echoing our arguments in relation to context, however, even if it were *possible* to develop an automated means of distinguishing between narratives and counter-narratives, this approach is not necessarily *desirable*. What the presumption of neat separation misses, is that anti-racist discourse is both discursively and materially entangled with its object of critique when expressed on social media platforms. In other words, counter-narratives are always reacting against something and convey traces of what they're contesting.

The risk of obscuring the relational properties of narratives is also an issue for other methods such as keyword and network analysis. For instance, in our pilot project we cross-referenced our retweet networks of key actors circulating #StopIslam (fig.3) with keyword searches of user bios to get a sense of the types of users in each cluster.

Fig. 3: Retweet network from pilot study



Our initial impression was of a stereotypical echo chamber, with those perpetuating hate speech (the cluster of users on the left) existing as a tightly bounded networks of individuals who engaged with one another frequently. In contrast, those contesting Islamophobia were a more heterogeneous and loose-knit ad hoc public. While their counter-narrative was ephemeral and short-lived, it received far more visibility than the Islamophobic tweets circulated by the tightly bounded and insular network of individuals on the left, as illustrated by our list of most retweeted posts being entirely dominated by users in the right-hand cluster. Thus, we originally concluded that the counter-narrative had been successful in contesting Islamophobia (Poole et al, 2019). Initial qualitative analysis seemed to support this explanation by offering explanations for why counter-narrative posts gained such traction, such as the participation of celebrities in the right-hand cluster who connected user-networks and gave posts disproportionate reach.

However, other aspects of our qualitative analysis illustrated the limitations of our initial explanations and showed they only captured one layer of interaction (engagement with original posts) but failed to incorporate antagonistic responses to these tweets. On analysing comment threads we found, in contrast, that discussion was dominated by accounts that used these highly visible posts to disseminate Islamophobic memes, jokes, and disinformation beyond far-right enclaves. In other words, even as counter-narrative tweets attempted to wrest the #StopIslam hashtag away from the far-right, these tweets were, in turn, appropriated to amplify hate speech.

Overall, challenges of classification underline a key point: If counter-narratives serve to amplify hate speech, this poses questions about whether 'counter-narrative' is an accurate label and highlights the limitations of automated methods in detecting such narratives. Even mixed methods approaches can exacerbate reductive conceptions of digital racism, if they fail to understand online content in relational terms and see it instead as a discreet entity that can easily be categorized in positive or negative terms. It is important, therefore, to reflect on how qualitative methods might be used to trouble neat assumptions about how digital racism can be classified by complicating, rather than straight-forwardly verifying, automated findings.

Reproducibility

Our final challenge relates less to complications that emerged through our data and is more a problem that stems from the commonplace desire to make better training sets for reproducing the accuracy of automation. In the late 00s concern emerged in fields such as psychology and medicine that researchers often struggled to reproduce one another's findings (Baker, 2016). The so-called reproducibility crisis has spread to machine learning (e.g. Olorisade, Brereton & Andras, 2021), including research directed at hate speech detection (Brivio and Coltekin, 2022). A key explanation for the challenge in replicating machine learning analysis of big data is that researchers often fail to provide sufficient detail about the source code of algorithms, the datasets that are used in particular publications and experiments, or the parameters of machine learning techniques, to enable others to reproduce findings accurately (e.g. Hutson, 2018).

Scholars argue that social media data pose specific challenges for developing algorithms that can reliably identify hate speech and reproduce accurate detection rates across datasets (Thylstrup and Talat, 2020). For instance, Ayo and colleagues' evaluation of sentiment analysis argues that a key barrier to developing algorithms that can consistently detect hate speech is that: 'Twitter streams contain large amounts of meaningless messages, contaminated content, and rumors, which adversely affect classification algorithm performance' (2020: 2). Similarly, Ligthart et al's (2021) systematic review of automated sentiment analysis traces a shared emphasis on the need to 'clean' social media content, to construct training datasets that can be easily classified for machine learning. The inference, then, is that it is possible to develop automated tools for reliably identifying hate speech, if, firstly, sufficiently accurate and 'clean' datasets are provided for training and, secondly, sufficient technical detail is provided, to enable algorithms trained on these datasets to be tested and used by others.

However, again, what is at stake is not just whether it is possible to find solutions to technical problems, but whether technofixes are desirable. This question was brought to the fore in relation to an issue commonplace in literature on hate speech detection: decision-making about how to deal with bots and, to use Ayo et al's (2020: 2) wording, 'contaminated content'. In our pilot study, we attempted to remove bots and spam in order to provide what – at the time – we felt was a more 'accurate' picture of how users were *deliberately* spreading and contesting hate speech on Twitter. Indeed, the role of bots was verified after we cross referenced our findings with data that Twitter published from the Russian Internet Research Agency's 3,841 accounts (released as part of a greater

push for transparency to combat platform manipulation) and discovered that one of the hashtags utilized by these accounts was #stopIslam (which was tweeted 1,498 times on the day of the bombing).

In the pilot, we used these 'cleaned' datasets for computational analysis before sampling tweets for our content and qualitative analysis. However, although bot-generated tweets that hijack trending hashtags might be unhelpful for refining automated hate speech detection tools, these tweets still play a critical role in mediating narratives (Marres and Moats, 2015). In the case of #stopIslam, for instance, even if the content was not overly Islamophobic these tweets still amplified online racism in ways that shaped the social world. In our larger project, therefore, we decided to retain tweets that appeared to be bots, spam, and 'contaminated content'.

The significance of this decision is illustrated by our Covid datasets. Our second data sample from July 2020 documented socially-isolated Eid celebrations in personal residences to combat Islamophobic narratives about Covid spreading at Eid, many of which used the hashtag #EidAtHome. Qualitative analysis identified that the most retweeted examples of this narrative were posts by high-profile Nigerian politicians that documented family celebrations, several of which received thousands of likes and hundreds of comments and retweets. However, many of the users responding to these tweets did not engage with their content but appeared to be spam accounts. For instance, the most shared 'Eid at Home' tweet had 672 retweets, 16.7k likes, and 230 comments, but almost 50% of the comments were spam tweets selling products and/or asking for follow backs that did not engage with the original post at all. Despite not substantively contributing to the content of the sort of counter-narratives we were interested in, the visibility of Eid at Home was nonetheless enhanced by these seemingly 'irrelevant' tweets.

From the perspective of developing large-scale, automated attempts to divide social media content into 'acceptable' and hate speech, the contamination of datasets with spam posts are technical problems to be overcome. When considering how to incorporate automated analysis into projects with social scientific aims, however, this decision is more complex as 'contaminated' content can play an important role in relation to the affordances of a platform wherein content is rendered visible due to trending topics and retweets. Bots, in other words, are a sociologically important phenomenon to examine in terms of their role in the wider assemblage that constitutes digital racism. A key issue, therefore, is that data which is good for learning with might be less useful for capturing a qualitative sense of digital narratives as unfolding in dynamic socio-technical settings.

When datasets are stored, archived, and used for future research, if decisions are made to remove particular content for the purpose of machine learning then there is a risk, again, of treating tweets as purely linguistic entities that provide the raw material for future intervention. What this approach obscures is the role of seemingly 'irrelevant' posts and platform affordances that are integral to materialising narratives in practice. Removing these tweets from datasets, in other words, removes content that is vital for future research that seeks to develop in-depth qualitative understanding of the assemblage of digital racism in which bots play a constitutive role.

Conclusion

Amidst wider concern about uncritical positivism when big data is turned to sociological ends, there are specific dangers of using automated methods to identify the dynamics of digital racism. These methods risk reducing racism and discrimination to a linguistic phenomenon and technical problem that can be identified with the right tools. At same time, automated analysis is essential for processing large data sets, and large datasets are, in turn, valuable for gaining insight into affordances of social media platforms that have played a significant role in normalizing racism. In our

work, for instance, identifying the most retweeted posts, and frequent keywords, or most followed people in datasets, was essential in identifying which themes, issues, or network clusters needed to be excavated through subsequent qualitative methods. The initial processing of data, in other words, is what made it possible to reveal the contexts, classificatory problems, and platform-specific affordances that mediated digital racism. Yet there is a risk, even in mixed methods research, of giving the impression of a linear process wherein automated analysis is used to identify patterns for more in-depth qualitative analysis, which both corroborates and deepens computational data. It is this relationship between methods of analysis that we have sought to trouble here.

In this article we have pointed to three overlapping tensions that posed challenges for our own mixed-methods research, related to contexts, classification, and reproducibility. Our aim has been to respond to provocations offered by recent critiques of big data for sociological research, particularly in the context of racism and racialization (Matamoros-Fernández and Farkas, 2021; Nikunen, 2021). These critiques offer important admonishments for ebullient understandings of big data analytics as offering neutral tools for detecting and removing racist content. At the same time, echoing STS scholarship that has emphasized the value of diffractive and interface methods, we have avoided a totalizing critique of techniques originating in data science and maintain the value of mixed methods. We argue, however, that dialogue between methods needs to be anchored in recognition of novel patterns that arise when different knowledges are brought together.

A recurring theme throughout this article is the tension between language-centred approaches, and sociological understandings of racism and racialization as relational processes, which emerge through complex socio-technological assemblages that evolve over time (Sharma, 2013; Brooker and Sharma, 2016). In offering a framework for overcoming these problems, we have argued that it is vital to resist the reduction of qualitative analysis purely to a layer of verification and to avoid assuming overly neat relationships between qualitative and quantitative data. Instead, we have advocated a diffractive approach that foregrounds how methods complicate one another; here this approach was essential in rearticulating more complex understandings of digital racism that could not be reduced to language alone.

Acknowledgements:

We would like to thank the UK Arts and Humanities Research Council (grant AH/T004460/1) and British Academy/Leverhulme Trust small grant scheme (SG161680) for funding the projects referenced in this article. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References:

Atton, C. (2006). Far-right media on the Internet. *New Media & Society*, 8(4): 573–587.

Awan I (2014) Islamophobia on Twitter: a typology of online hate against Muslims on social media. *Policy & Internet* 6: 133–150.

Ayo, F.E. et al (2020) Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, p.100311.

Baker, M. (2016) Reproducibility crisis. *Nature*, 533(26): 353-66.

Barad, K. (2007) *Meeting the Universe Halfway*. Durham, NC: Duke University Press.

Brivio, M. and Coltekin, C., 2022, February. Reproducibility report: Hate Speech Detection based on Sentiment Knowledge Sharing. In *ML Reproducibility Challenge 2021 (Fall Edition)*.

Bruns, A. and Burgess, J. (2011) The use of Twitter hashtags in the formation of Ad hoc publics. In: *Proceedings of the 6th European Consortium for Political Research (ECPR) general conference 2011*. University of Iceland, Reykjavik. Available at: https://eprints.qut.edu.au/46515/

Caiani, M. and Wagemann, C. (2009) Online networks of the Italian and German extreme right: An explorative study with social network analysis. *Information, Communication & Society*, *12*(1), pp.66-109.

Castells, M. (1997) *The Power of Identity.* Oxford: Blackwell Publishers.

Fitzgerald, D. and Callard, F. (2015) Social science and neuroscience beyond interdisciplinarity. *Theory, Culture & Society*, 32(1): 3-32.

Gangneux, J. (2019) Rethinking social media for qualitative research. *The Sociological Review*, *67*(6), pp.1249-1264.

Ghasiya, P. and Sasahara, K. (2022) Rapid sharing of Islamophobic Hate Speech on Facebook. *Social Media + Society*, 8(4). DOI: 10.1177/2056305122112915

Gorwa, R., Binns, R. and Katzenbach, C. (2020) Algorithmic content moderation. *Big Data & Society*, 7(1), p.2053951719897945.

Harris, C. et al (2022) Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '22, June 21–24, 2022, Seoul, Republic of Korea. Available: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533144

Hutson, M. (2018) Artificial intelligence faces reproducibility crisis. *Science*, 359(6377): 725-726.

Jackson, S.J., Bailey, M. and Welles, B.F. (2020) *#HashtagActivism: Networks of race and gender justice*. Cambridge, Mass.: MIT Press.

Ligthart, A. et al (2021) Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, *54*(7): 4997-5053.

Marres, N. and Gerlitz, C. (2016) Interface methods: Renegotiating relations between digital social research, STS and sociology. *The Sociological Review*, 64(1): 21-46.

Marres, N. and Moats, D. (2015) Mapping controversies with social media: The case for symmetry. *Social Media+ Society*, *1*(2), p.2056305115604176.

Marres, N. (2017) *Digital Sociology: The reinvention of social research*. Cambridge: Polity.

Marwick, A.E. and Lewis, R. (2017) *Media Manipulation and Disinformation Online*. Available: https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Matamoros-Fernández, A. and Farkas, J. (2021) Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2): 205-224.

Mercea, D., Iannelli, L. and Loader, B.D. (2016) Protest communication ecologies. *Information, Communication & Society*, 19(3): 279-289.

Nikunen, K. (2021) Ghosts of white methods? The challenges of Big Data research in exploring racism in digital context. *Big Data & Society*, 8(2): 20539517211048964.

Olorisade, B.K., Brereton, P. and Andras, P. (2017) Reproducibility of studies on text mining for citation screening in systematic reviews. *Journal of Biomedical Informatics*, 73, pp.1-13.

Papacharissi, Z. (2015) *Affective Publics: Sentiment, technology, and politics*. Oxford: Oxford University Press.

Phillips, W. (2015) *This Is Why We Can't Have Nice Things*. Cambridge, Mass: MIT Press.

Poole, E., Giraud, E.H., Richardson, J.E. and de Quincey, E. (2023) Expedient, affective, and sustained solidarities? Mediated contestations of Islamophobia in the case of Brexit, the Christchurch terror attack, and the COVID-19 pandemic. *Social Media + Society*, 9(3). DOI: 10.1177/205630512311994

Poole, E. and Williamson, M. (2021) Disrupting or reconfiguring racist narratives about Muslims? *Journalism*, p.14648849211030129.

Poole, E., Giraud, E.H. and de Quincey, E. (2021) Tactical interventions in online hate speech: The case of# stopIslam. *New Media & Society*, 23(6): 1415-1442.

Poole, E., Giraud, E. and de Quincey, E. (2019) Contesting# StopIslam: The dynamics of a counter-narrative against right-wing populism. *Open Library of Humanities*, 5(1). DOI: 10.16995/olh.406

Richardson, J.E., Giraud, E.H., Poole, E. and de Quincey, E. (2024) 'Hypocrite!' Affective and argumentative engagement on Twitter, following the Christchurch terrorist attack. *Media, Culture & Society*, DOI: 10.1177/01634437241229322.

Richardson, J.E. (2008) "Our England": discourses of "race" and class in party election leaflets. *Social Semiotics*, 18(3): 321-335.

Richardson, J.E. and Franklin, B. (2003) 'Dear Editor': Race, readers' letters and the local press. *The Political Quarterly*, 74(2): 184-192.

Sachdeva, P.S. et al (2022) Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, June 21–24, Seoul, Republic of Korea. Available: https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533216

Savage M, Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Sharma, S. and Brooker, P. (2016) # notracist: Exploring racism denial talk on Twitter. In: J. Daniels, K. Gregory and T. M. Cottom (eds) *Digital Sociologies*. Bristol: Policy Press, 463-485.

Sharma, S. (2013) Black Twitter? Racial hashtags, networks and contagion. *New Formations*, 78(78): 46-64.

Siapera, E. (2019) Organised and ambient digital racism. *Open Library of Humanities*, 5(1). https://doi.org/10.16995/olh.405

Siapera, E. et al (2018) Refugees and network publics on Twitter. *Social Media+ Society*, 4(1): p.2056305118764437.

Thylstrup, N.B. (2022) The ethics and politics of data sets in the age of machine learning. *Media, Culture & Society*, p.01634437211060226.

Thylstrup, N. and Talat, Z. (2020) Detecting 'dirt'and 'toxicity': Rethinking content moderation as pollution behaviour. Available at SSRN 3709719.

Tinati, R. et al (2014) Big data: Methodological challenges and approaches for sociological analysis. *Sociology*, 48(4): 663-681.

Twitter (2021) Coordinated harmful activities. *Twitter Help Center*. Available: https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity

Twitter Safety (2021) An update following the riots in Washington, D.C. *Twitter Blog*, 12th January. Available: https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--

Twitter Inc. (2021) Permanent suspension of @realDonaldTrump. *Twitter Blog*, 8th January. Available: https://blog.twitter.com/en_us/topics/company/2020/suspension

Wilcox, S. (2020) Calderdale MP says Muslim and BAME communities not taking pandemic seriously. *Halifax Courier*, 31st July. Available: https://www.halifaxcourier.co.uk/news/opinion/calderdale-mp-says-muslim-and-bame-communities-are-not-taking-pandemic-seriously-2929468

X Developer Platform (ND) Success Story: Hate Lab. Available: https://developer.twitter.com/en/blog/success-stories/hatelab

Zuberi, T. and Bonilla-Silva, E. eds. (2008) *White Logic, White Methods: Racism and Methodology*. Plymouth: Rowman & Littlefield Publishers.

Accepted for publication in The Sociological Review. Pre-copy edited version.

[i] #HelloBrother trended due to being the words spoken to the white supremacist terrorist as he entered the Al Noor Mosque, by of one of the first victims, Haji-Daoud Nabi. #TablighiJamaat refers to a Sunni missionary movement and was a hashtag used to spread Islamophobic hate speech that held Muslims responsible for covid outbreaks in India.