

Racism, Hate Speech, and Social Media: A Systematic Review and Critique

Television & New Media
2021, Vol. 22(2) 205–224
© The Author(s) 2020



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1527476420982230
journals.sagepub.com/home/tvn



Ariadna Matamoros-Fernández¹ 
and Johan Farkas²

Abstract

Departing from Jessie Daniels's 2013 review of scholarship on race and racism online, this article maps and discusses recent developments in the study of racism and hate speech in the subfield of social media research. Systematically examining 104 articles, we address three research questions: Which geographical contexts, platforms, and methods do researchers engage with in studies of racism and hate speech on social media? To what extent does scholarship draw on critical race perspectives to interrogate how systemic racism is (re)produced on social media? What are the primary methodological and ethical challenges of the field? The article finds a lack of geographical and platform diversity, an absence of researchers' reflexive dialogue with their object of study, and little engagement with critical race perspectives to unpack racism on social media. There is a need for more thorough interrogations of how user practices and platform politics co-shape contemporary racisms.

Keywords

racism, hate speech, review, social media, platforms, critical race theory, whiteness

Introduction

Across the digital landscape, sociality is continuously transformed by the interplay of humans and technology (Noble 2018a). In this regard, social media companies play a particularly central role, as a handful of mainly US and Chinese corporations have grown into near-ubiquitous giants. While companies such as Facebook present

¹Queensland University of Technology, Kelvin Grove, QLD, Australia

²Malmö University, Malmö, Skåne, Sweden

Corresponding Author:

Johan Farkas, School of Art and Communication, Malmö University, Nordenskiöldsgatan 1, 211 19
Malmö, Sweden.

Email: johan.farkas@mau.se

themselves as democratizing forces, increased attention has in recent years been given to their role in mediating and amplifying old and new forms of abuse, hate, and discrimination (Noble and Tynes 2016; Matamoros-Fernández 2017; Patton et al. 2017).

In a review and critique of research on race and racism in the digital realm, Jessie Daniels (2013) identified social media platforms—specifically social network sites (SNSs)—as spaces “where race and racism play out in interesting, sometimes disturbing, ways” (Daniels 2013, 702). Since then, social media research has become a salient academic (sub-)field with its own journal (*Social Media + Society*), conference (*Social Media & Society*), and numerous edited collections (see e.g. Burgess et al. 2017). In parallel, scholars have grown increasingly concerned with racism and hate speech online, not least due to the rise of far-right leaders in countries like the US, Brazil, India, and the UK and the weaponization of digital platforms by white supremacists. This has caused a notable increase in scholarship on the topic.

As social media have come to dominate socio-political landscapes in almost every corner of the world, new and old racist practices increasingly take place on these platforms. Racist speech thrives on social media, including through covert tactics such as the weaponization of memes (Lamerichs et al. 2018) and use of fake identities to incite racist hatred (Farkas et al. 2018). Reddit gives rise to toxic subcultures (Chandrasekharan et al. 2017; Massanari 2015), YouTube to a network of reactionary right racist influencers (Murthy and Sharma 2019; Johns 2017), and coordinated harassment is pervasive on Twitter (Shepherd et al. 2015). Users also (re)produce racism through seemingly benign practices, such as the use of emoji (Matamoros-Fernández 2018) and GIFs (Jackson 2017).

Social media contribute to reshaping “racist dynamics through their affordances, policies, algorithms and corporate decisions” (Matamoros-Fernández 2018, 933). Microaggressions (Sue 2010) as well as overt discrimination can be found in platform governance and designs. Snapchat and Instagram have come under fire for releasing filters that encourage white people to perform “digital blackface” (Jackson 2017) and automatically lighten the skin of non-whites (Jerkins 2015). Facebook, by tracking user activity, enabled marketers to exclude users with what they called an African American or Hispanic “ethnic affinity” (Angwin and Parris 2016). And TikTok has faced criticism, when it suspended a viral video raising awareness of China’s persecution of Uighurs (Porter, 2019). This shows that digital technologies not only “render oppression digital” but also reshape structural oppression based on race, gender, and sexuality as well as their intersectional relationship (Bivens and Haimson 2016; Chun 2009; Nakamura 2008; Noble 2018a; Noble and Tynes 2016). Social media platforms’ policies and processes around content moderation play a significant role in this regard. Companies like Facebook and Twitter have been criticized for providing vast anonymity for harassers (Farkas et al. 2018) and for being permissive with racist content disguised in humor because it triggers engagement (Roberts 2019; Shepherd et al. 2015).

Racist discourses and practices on social media represent a vital, yet challenging area of research. With race and racism increasingly being reshaped within proprietary platforms like Facebook, WhatsApp, WeChat, and YouTube, it is timely to review publications on the subject to discuss the state of this field, particularly given the

growth in scholarly attention. This article presents a systematic literature review and critique of academic articles on racism and hate speech on social media from 2014 to 2018. Departing from Daniels's (2013) literature review, the article critically maps and discusses recent developments in the subfield, paying specific attention to the empirical breadth of studies, theoretical frameworks used as well as methodological and ethical challenges. The paper seeks to address three research questions: (1) Which geographical contexts, social media platforms and methods do researchers engage with in studies of racism and hate speech on social media? (2) To what extent does scholarship draw on resources from critical race perspectives to interrogate how systemic racism is (re)produced on social media? (3) What are the primary methodological and ethical challenges of the field?

A Note on the Importance of Critical Race Perspectives for Social Media Research

In this review, we use the term critical race perspectives to refer to theoretical frameworks that interrogate how race and power are shaped by and shape socioeconomic and legal systems and institutions. This includes critical race theory (see Delgado and Stefancic 2001; Matsuda et al. 1993), intersectionality (see Collins 2000; Crenshaw 1991), whiteness studies (see Feagin 2006; Frankenberg 1993), postcolonial theory (see Goldberg and Quayson 2002; Wolfe 2016), and critical Indigenous studies (see Moreton-Robinson 2015; Nakata, 2007). These theories are not only productive lenses to unpack race and power on social media, but they can also help scholars avoid perpetuating power imbalances in research designs since they ground ethical research best practices in the experiences of marginalized groups (Collins 2000; see also Linabary and Corple 2019 for the use of standpoint theory in online research practice).

Scholars have used critical race perspectives to examine digital technologies since the early years of internet research (Nakamura 2002; Everett 2009). However, as Daniels (2013) identified in her literature review on racism in digital media, much of this work has relied on Omi and Winant's (1994) racial formation theory to explain how racial categories are created and contested online. Daniels called for more thorough critiques of whiteness within internet studies—for example drawing on the foundational work of W. E. B. Du Bois (1903)—as a crucial precondition to combating rampant inequality in our societies. She also argued that racial formation theory is an inadequate framework for understanding the relationship “between racism, globalization and technoculture” since, she notes, Omi and Winant gave little attention to racism and focused too much on the State as the primary system of oppression (p. 710). To this can be added that US-focused theories like Omi and Winant's do not always translate easily to thinking practices, infrastructures and norms outside the US, for example in reproducing European coloniality and settler gazes.

New research is emerging on the opportunities and challenges of using social media for Indigenous people (Carlson and Dreher 2018; Raynauld et al. 2018), some of which is using Indigenous perspectives, such as the work of Torres Strait Islander

academic Martin Nakata (2007), to problematize race struggles on social media (Carlson and Frazer 2020). The use of critical Indigenous studies as lenses to interrogate racism and hate speech on social media, though, is still scarce. Regardless of the specific critical race perspective used to question digital technologies, more needs to be done to scrutinize how digital platforms reproduce, amplify, and perpetuate racist systems. As we shall see in the results of this systematic review, only a minority of academic articles are drawing on critical race perspectives to study racist practices on social media and the inner workings of digital platforms, despite sustained calls for more critical interrogations.

Research Design

The studied literature was collected in October and November 2018 through Google Scholar and the Web of Science. While the Web of Science was selected due to its emphasis on reputable and high-ranking academic journals, Google Scholar was included due its emphasis on article citations as a metric of proliferation. To limit the scope of the study, we established five sampling criteria: (1) publications had to be in the form of scholarly articles; (2) publications had to be written in English; (3) publications had to be primarily focused on racism or hate speech against ethno-cultural minorities; (4) publications had to be primarily focused on social media platforms; and (5) publications had to be published between 2014 and 2018. The time period was selected in order to capture scholarship in the 5-year period since Daniels's (2013) literature review on race, racism and Internet studies from 2013. In terms of limitations, it should be noted that the exclusion of non-English articles limits the geo-linguistic breadth of the study. Additionally, the exclusion of monographs and anthologies means the study cannot be seen as a definite account of the field. We try to compensate for the latter by including references to key books on race, racism and technology from the study period in the introduction and discussion (e.g. Noble and Tynes 2016; Noble 2018a). Nonetheless, a potential limitation of the sampling criteria is an underrepresentation of in-depth and critical perspectives, which are likely more prevalent in longer exposés (as we return to in the discussion).

After an explorative identification of productive queries, four were chosen with combined search terms: (a) "racism" "social media," (b) "racism" "digital methods," (c) "hate speech" "social media," (d) "hate speech" "digital methods." We queried "digital methods" due to our study's focus on methodological and ethical challenges of the field. From Google Scholar, 200 publications were collected (50 from each query), while 110 publications were collected from the Web of Science (all results). Of the collected texts, 40 were archived in multiple queries and/or from both websites, bringing the total number down to 270. Subsequently, all publications were evaluated based on the five sampling criteria, excluding texts that were not relevant for the study.¹ In total, 166 publications were excluded, bringing the final sample down to 104 articles. Of these, 15 were published in the first 2 years (2014–2015), while 60 were published in the latter two (2017–2018), showing a clear rise in scholarly attention.

We analyzed the final sample through two rounds of deductive content analysis followed by an in-depth qualitative analysis. In order to ensure consistency in the

coding process, we both initially coded a random subsample of 12 articles (11.54% of the sample), comparing results and revising the approach accordingly. The first round of content analysis contained eight variables: year of publication; national context(s) of study; social media platform(s) studied; language(s) of studied material; methodological approach (qualitative/quantitative or mixed methods); whether the study relies on access to social media platform's API(s) (Application Programming Interfaces); whether the article contains a positionality statement from the author(s); and whether the study contains mentions of critical race perspectives (as defined above). In the second round of deductive coding, we coded the method(s) used in each study. To ensure consistency, we first established a codebook inductively, containing nine methods in total (multiple of which could be selected when coding each article). Finally, the dataset was subject to an in-depth, open-ended qualitative analysis, noting methodological and ethical challenges described in the literature. These were grouped under overall themes and topics.

In a study and critique of academic literature on racism and hate speech, it is important to reflect on how our position as white, cisgender, heterosexual, European, middle class researchers shape our assumptions about racism on social media. There are obvious limitations to our interpretations of the dynamics of racism as a result of our own privilege and lack of lived experiences with racial discrimination. However, instead of leaving “the burden of noticing race on the internet. . . to researchers who are people of color” (Daniels 2013, 712), our aim with this paper is to contribute with a careful analysis of how whiteness informs social media studies on racism and to propose ways to avoid (re)producing racism through research designs and a lack of commitment to critical theory.

RQ1: Geography, Platforms, and Methodologies

While discrimination of ethno-cultural minorities takes place across the globe and across the digital realm, scholarship on racism, hate speech and social media remains limited to a few contexts and platforms. In terms of geographical breadth, our findings show that North America—especially the United States—is by far the most studied geographical context, with 44.23% of all studies focusing on this region ($n=46$). Europe is the second most studied region (25.96%, $n=27$), with close to half of European studies focusing on the United Kingdom ($n=12$). This is followed by Asia and Oceania (each at 5.77%, $n=6$), the Middle East (1.92%, $n=2$) and South America and Africa (each at 0.96%, $n=1$). These figures highlight a wide discrepancy between, what has been termed, the Global North and Global South (see Figure 1). These findings resonate with previous research, arguing for a grave need to “de-Westernize” media and data studies (Cunningham and Terry 2000, 210; see also Milan and Treré 2019).

Twitter is by far the most studied platform (see Figure 2), examined in 54.81% of all articles in our sample ($n=57$), followed by Facebook (34.62%, $n=36$), YouTube (8.65%, $n=9$), Reddit (3.85%, $n=4$), Whisper (2.88%, $n=3$), YikYak (1.92%, $n=2$), Tumblr (1.92%, $n=2$), Instagram (0.96%, $n=1$), and Tinder (0.96%, $n=1$). Not a

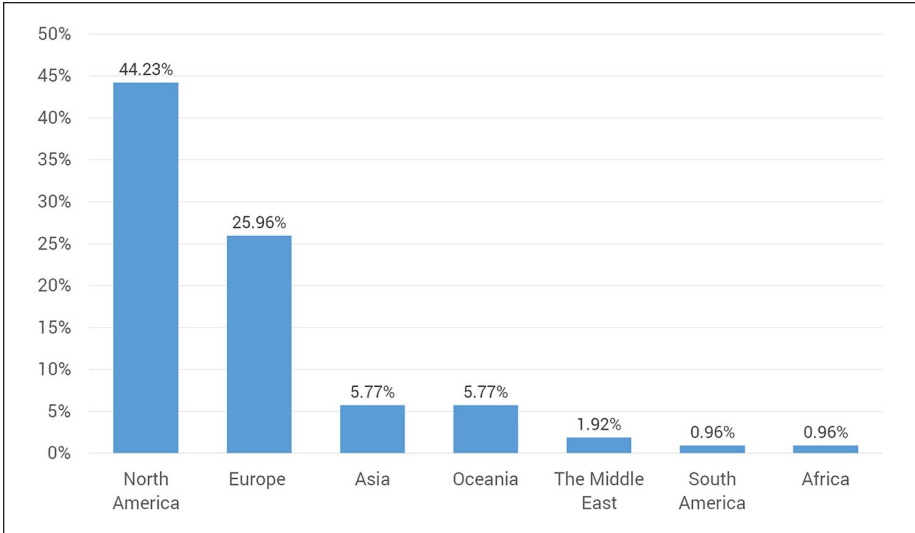


Figure 1. Percentage of studies examining different geographic regions.

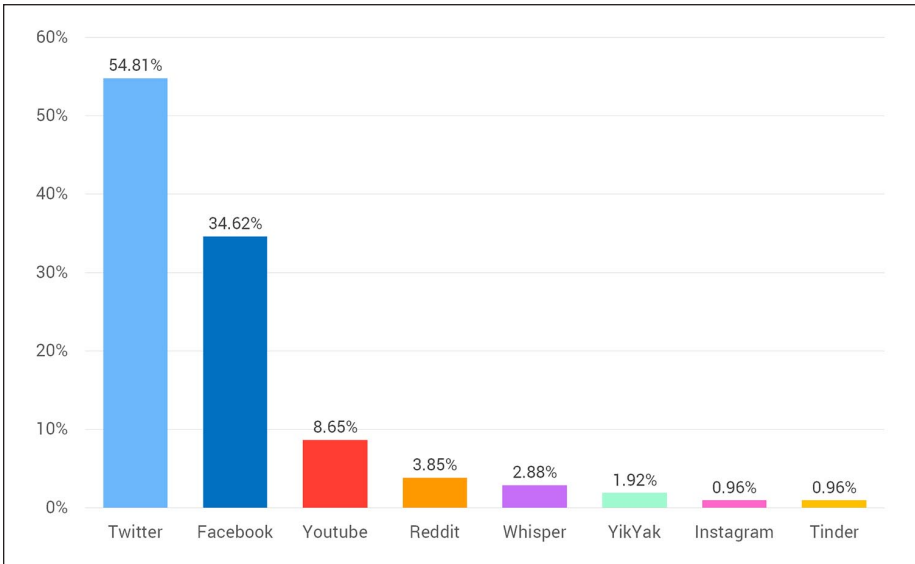


Figure 2. Percentage of studies examining different social media platforms.

single study examines major platforms like WhatsApp or WeChat. This points towards a key challenge for the field in terms of ensuring platform diversity and cross-platform analyses of racism and hate speech.

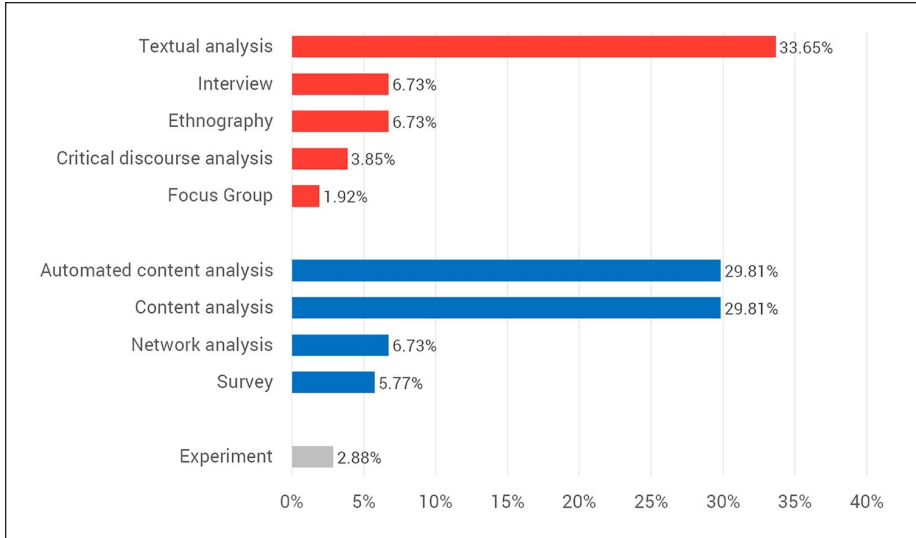


Figure 3. Use of specific qualitative methods (red) and quantitative methods (blue).

The prominence of Twitter in the academic literature is likely tied to the relative openness of the platform’s APIs. Some studies explicate this connection, stating that Twitter “differs from others such as Facebook, in that it is public and the data are freely accessible by researchers” (Williams and Burnap 2016, 218). Twitter allows researchers to collect “public” data without obtaining informed consent or talk to the communities under study, a practice that has increasingly been criticized for potentially reproducing inequalities (Florini et al. 2018; Linabary and Corple 2019; Milan and Treré 2019). In total, 41.35% of studies relied on platform APIs for data collection ($n=43$), 67.44% of which focused on Twitter ($n=29$).

Methodological Approaches and the Hate Speech/Racism Divide

Qualitative and quantitative approaches are close to equally represented in the field. While qualitative methods are used in 40.38% of studies ($n=42$), quantitative methods are used in 35.58% ($n=37$). Only 12.5% rely on mixed methods approaches ($n=13$), while 11.54% do not rely on empirical data ($n=12$). Across the corpus, we find a clear overrepresentation of text-based forms of analyses (see Figure 3), a trend already observed in previous literature (Bliuc et al. 2018). In qualitative studies, textual analysis is by far the most prevalent method, used in 33.65% of all studies (63.64% of qualitative and mixed methods studies, $n=35$). This stands in contrast to interactional forms of research, such as interviews and ethnography (each used in 6.73% of all studies, $n=7$). In quantitative studies, text-based analysis also dominates, with 29.81% of all studies using manual and automated form of content analysis respectively ($n=31$, that is 62% of quantitative and mixed methods studies). This stands in contrast to network analysis (used in 6.73% of all studies, $n=7$) and surveys (5.77%, $n=6$).

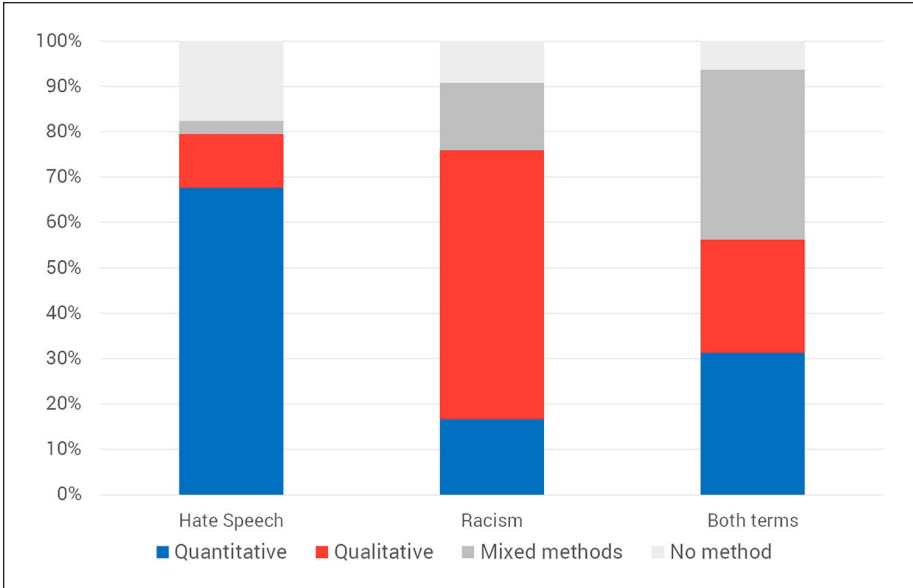


Figure 4. Methodological approaches of studies found in queries for “hate speech,” “racism,” and both.

While qualitative and quantitative research on racism, hate speech, and social media shares a preoccupation with text-based analysis, we find a clear discrepancy in the use of concepts (see Figure 4). Of the sampled articles collected solely through queries containing the term “hate speech” (as opposed to “racism”), 67.65% of studies draw on quantitative methods ($n=23$), while only 11.77% rely on qualitative methods ($n=4$). In studies archived solely through queries for “racism,” we find the opposite pattern. Here, 59.26% of studies draw on qualitative methods ($n=32$), while solely 16.67% rely on quantitative methods ($n=9$). This points to a terminological divide in the field, indicating a lack of scholarly exchange between the humanities/social sciences and computer science/data science. Our findings point to the latter group placing less emphasis on structural, ideological, and historical dimensions of racial oppression (associated with the term “racism”) than the former group and more emphasis on text-level identification and the legality of content (associated with the term “hate speech”). As we will return to, this divide has critical implications, especially due to the lack of critical reflections from quantitative scholars.

RQ2: Critical Race Perspectives and Positionality Statements

In terms of the prevalence of critical race perspectives, we find that 23.08% of articles contain mentions and/or references to these lines of research ($n=24$), while 76.92% do

not ($n=80$). This indicates that only a minority of scholars are relying on critical approaches to the study of racism and social media. We once again find a clear divide between qualitative and quantitative research, with only 5.41% of quantitative studies containing mentions of critical race perspectives ($n=2$), as opposed to 45.24% of qualitative studies ($n=19$).

From the critical literature, less than half of the papers examine how whiteness plays out on social media. Mason (2016) uses Du Bois (1903) to argue that hookup apps like Tinder secure and maintain “the color line” (p. 827). Nishi, Matias, and Montoya (2015) draw on Fanon’s and Lipsitz’s thinking on whiteness to critique how virtual white avatars perpetuate American racism, and Gantt-Shafer (2017) adopts Picca and Feagin’s (2007) “two-faced racism” theory to analyze frontstage racism on social media. Omi and Winant’s racial formation theory is still used, with authors drawing on this framework to examine racial formation in Finland during the refugee crisis in Europe 2015–2016 (Keskinen 2018) and racist discourse on Twitter (Carney 2016; Cisneros and Nakayama 2015). Research drawing on critical Indigenous studies to examine racism on social media is scarce but present in our sample. Matamoros-Fernández (2017) incorporates Moreton-Robinson’s (2015) concept of the “white possessive” to examine Australian racism across different social media platforms, and Ilmonen (2016) argues that studies interrogating social media could benefit from triangulating different critical lenses such as postcolonial studies and Indigenous modes of criticism. Echoing Daniels (2013), several scholars also call for developing “further critical inquiry into Whiteness online” (Oh 2016, 248), arguing there is a neglected “potential for critical race studies to attend to the ways that social media technologies express and reinforce cultural logics of race and the articulation of racism” (Cisneros and Nakayama 2015, 119).

In terms of positionality statements from authors, reflecting on their role as researchers in studying and contesting oppression, only 6.73% of studies contain such statements ($n=7$), making them marginal within the field. In the few statements we find, authors acknowledge how their “interpretation of the data is situated within the context of our identities, experiences, perspectives, and biases as individuals and as a research team” (George Mwangi et al. 2018, 152). Similarly, in a few ethnographic studies, authors reflect on taking part in the fight against discrimination (see Carney 2016).

RQ3: Methodological and Ethical Challenges

There are key commonalities in the methodological challenges faced by researchers in our sample. A majority of quantitative scholars note the difficulty of identifying text-based hate speech due to a lack of unanimous definition of the term; the shortcomings of only keyword-based and list-based approaches to detecting hate speech (Davidson et al. 2017; Eddington 2018; Saleem et al. 2017; Waseem and Hovy 2016); and how the intersection of multiple identities in single victims presents a particular challenge for automated identification of hate speech (see Burnap and Williams 2016). As a possible solution to these challenges, Waseem and Hovy (2016) propose the incorporation of critical race theory in n-gram probabilistic language models to detect hate speech.

Instead of using list-based approaches to detecting hate speech, the authors use Peggy McIntosh's (2003) work on white privilege to include speech that silences minorities, such as negative stereotyping and showing support for discriminatory causes (i.e. #BanIslam). Such approaches to detecting hate speech were rare in our sample, pointing to a need for further engagement among quantitative researchers with critical race perspectives.

Data limitations are a widely recognised methodological concern too. These limitations include: the non-representativeness of single-platform studies (see Brown et al. 2017; Hong et al. 2016; Puschmann et al. 2016; Saleem et al. 2017); the low and incomplete quality of API data, including the inability to access historical data and content deleted by platforms and users (see Brown et al. 2017; Chandrasekharan et al. 2017; Chaudhry 2015; ElSherief et al. 2018; Olteanu et al. 2018); and geo-information being limited (Chaudhry 2015; Mondal et al. 2017). Loss of context in data extractive methods is also a salient methodological challenge (Chaudhry 2015; Eddington 2018; Tulkens et al. 2016; Mondal et al. 2017; Saleem et al. 2017). To this, Taylor et al. (2017, 1) note that hate speech detection is a "contextual task" and that researchers need to know the racists communities under study and learn the codewords, expressions, and vernaculars they use (see also Eddington 2018; Magu et al. 2017).

The qualitative and mixed methods studies in our sample also describe methodological challenges associated with a loss of context, difficulty of sampling, slipperiness of hate speech as a term, and data limitations such as non-representativeness, API restrictions and the shortcomings of keyword and hashtag-based studies (Black et al. 2016; Bonilla and Rosa 2015; Carney 2016; Johnson 2018; Miškolci et al. 2020; Munger 2017; Murthy and Sharma 2019; George Mwangi et al. 2018; Oh 2016; Petray and Collin 2017; Sanderson et al. 2016; Shepherd et al. 2015).

Researchers in the field note how the ambivalence of social media communication and context collapse poses serious challenges to research, since racism and hate speech can wear multiple cloaks on social media including humor, irony, and play (Cisneros and Nakayama 2015; Farkas et al. 2018; Gantt-Shafer 2017; Lamerichs et al. 2018; Matamoros-Fernandez 2018; Page et al. 2016; Petray and Collin 2017; Shepherd et al. 2015). This invites researchers to look into new pervasive racist practices on social media, for example as part of meme culture.

In terms of the ethics of racism and hate speech research on social media, particularly qualitative studies raise important points. To avoid processes of amplification, researchers make explicit their choice of not including the name of hateful sites under scrutiny (Tulkens et al. 2016). Noble (2018b) warns about oversharing visual material on social media that denounces police brutality by questioning whether videos of Black people dying serve as anything but a spectacle, while McCosker and Johns (2014) note that the sharing of videos of racist encounters raises issues of privacy. Ethical reflections among quantitative studies are conspicuously absent, which is an important reminder of Leurs' observation: "What often gets silenced in the methods sections of journal articles is how gathering digital data is a context-specific and power-ridden process similar to doing fieldwork offline" (Leurs 2017, 140). Reflections on ethical challenges of studying far-right groups also largely remain absent in the

literature, despite clear ethical challenges regarding risk of attacks on researchers, emotional distress and difficult questions of respecting the privacy of abusers versus protecting victims.

Discussion: The Intersectional Relationship Between Place, Race, Sex, and Gender

Based on our findings, this section draws on an intersectional lens and critical understandings of whiteness to discuss the overall patterns observed in our review and propose ways to move forward in the field. Specifically, following Linabary and Corple (2019), we consider that key intersectional concepts such as ethics of care and standpoint theories, which “inform the enactment of the principles of context, dialogue, and reflexivity” (1459), are fruitful when thinking about best practices within research in the (sub-)field of social media research on racism and hate speech.

Starting with the skewed representation of geographical regions, platforms, and methods in the field—our first research question - our study finds that the United States is more studied than the rest of the world combined. It should, of course, be noted in this regard that our sampling strategy focuses on English-language articles only, which likely plays a role in amplifying this pattern. That said, we still see our results as substantial, considering how English has increasingly become a global linguistic standard in academic publishing over the past decades (Jenkins 2013). Media studies is in general largely Western-centric (Cunningham and Terry 2000; Milan and Treré 2019) and white scholars also dominate within the US and Europe (Chakravartty et al. 2018). Our review points towards a grave need for further studies exploring racism outside the US, including discrimination against minorities such as the Roma (see Miškolci et al. 2020). There is also a lack of acknowledgement from scholars within the US—particularly quantitative researchers—of leading researchers of colour studying how race, gender, and sex intersect online (see Brock 2012, Nakamura 2008, Noble 2018a).

Turning to the social media platforms in the literature, the dominance of Twitter is substantial and problematic. This platform is far overrepresented, especially considering its relatively small user base as compared to for example Facebook, YouTube, WeChat, WhatsApp, and Instagram. Daniels (2013) noted that there were substantive areas missing in her review, such as “literature about race, racism and Twitter” (711). Studies of Twitter have since mushroomed, making all other platforms seem marginal in the field. Moving beyond Twitter is important, as social media platforms’ specific designs and policies play a key role in shaping racism and hate speech online (Pater et al. 2016; Noble 2018a). Digital interfaces, algorithms and user options “play a crucial role in determining the frequency of hate speech” (Miškolci et al. 2020, 14), for example by enabling anonymity to harassers and algorithmically suggesting racist content (Awan 2014; Gin et al. 2017). Platforms also attract different demographics, with Twitter being known for its usage by political elites and journalists (Gantt-Shafer, 2017), activists (Bosch 2017; Puschmann et al. 2016; Keskinen, 2018), and racial minorities (most notably in the US with what been dubbed “Black Twitter,” see Bock

2017). Accordingly, ensuring platform diversity and cross-platform analyses in empirical studies of racism, hate speech and social media—from TikTok and WeChat to WhatsApp, YouTube, Tumblr, and Tinder—is crucial for understanding and contesting how different technologies (re)shape racisms.

Regarding methodological approaches in the field, it is positive to find qualitative and quantitative methods close to equally represented. It is significant to note, however, the striking differences in the conceptual vocabularies used across quantitative and qualitative studies, with the former predominantly using the term “hate speech” and the latter using “racism.” This indicates a disciplinary split between the humanities/social sciences and computer science/data science, with researchers in the former tradition placing greater emphasis on histories, ideologies and structures of oppression. A majority of the quantitative articles focus on surface-level detection of hate speech without drawing connections to wider systems of oppressions and without engaging with critical scholarship. While hate speech identification is a legitimate research problem, this literature tends to reduce racism to *just* overt abusive expression to be quantified and removed, ignoring how racism is defined as social and institutional power plus racial prejudice (Bonilla-Silva 2010), which in social media translates to the power platforms exert on historically marginalised communities through their design and governance as well as user practices (Matamoros-Fernández 2017). Accordingly, computer scientists and data scientists have to start reflecting more on the connection between online expressions of bigotry and systemic injustice.

In terms of methods used in the field, the predominant focus on written text in both quantitative and qualitative studies neglects important factors involved in racism on social media, such as the ways in which discrimination is increasingly mediated through visual content (Lamerichs et al. 2018; Mason 2016; Matamoros-Fernández 2018). More scholarly work is needed on user practices and lived experiences with racism on social media, a gap in the literature also observed in previous reviews (Bliuc et al. 2018). Such research, combined with a broader focus on semi-closed platforms such as WhatsApp and ephemeral content such as Instagram Stories, which have become increasingly prevalent, might uncover new aspects of online racism.

In relation to our second research question, we find that only a minority of studies draw on critical race perspectives to examine racism and hate speech on social media. From these, critical race theory/whiteness lenses are mainly used to examine texts as well as the experience of users, and less to explain the implications of materiality in relation to racism and social media. In their edited collection on race, sex, class, and culture online, Nobles and Tynes (2016) propose intersectionality as a lens for “thinking critically about the Internet as a way that *reflects*, and a site that *structures*, power and values” (2, original emphasis). Our review indicates that more work is needed in this area, especially more studies using intersectionality as a framework for understanding how capitalism (class), white supremacy (race), and heteropatriarchy (gender) reflect and structure social media designs and practices. A lack of engagement with critical perspectives risks reproducing ideological investments in color-blindness, neglecting how power operates on social media. For example, only two studies in our sample focused on how white people inadvertently perpetuate racism on social media

(see Cisneros and Nakayama 2015; Mason 2016). In terms of limitations of these findings, it should be noted that the exclusion of monographs and anthologies from our sample likely amplifies the lack of criticality in the field, as several important critical monographs and collections have come out in the study period (see e.g. Noble and Tynes, 2016, Noble, 2018a). Even if this is the case, however, this does not resolve the identified problem of a lack of critical engagement in academic articles. Instead, it should spark a discussion about how to overcome these limitations.

Based on our findings, and attaining to the importance of context when interrogating race and racism, Indigenous perspectives are missing in the literature. While there is growing research on how social media “is providing the means whereby Indigenous people can ‘reterritorialise’ and ‘Indigenise’ the information and communication space” (Wilson et al. 2017, 2; see also O’Carroll 2013), Indigenous ontologies and epistemologies are yet to be foregrounded as lenses to interrogate the politics of social media. For example, in a manifesto on how to rethink humans’ relationship with AI—which could be also applied to rethink humans’ relationship with Silicon Valley-developed social media platforms—Indigenous scholars explain how “relationality is rooted in context and the prime context is place” (Lewis et al. 2018, 3). In this regard, the authors argue that the country to which AI (or social media) currently belongs “excludes the multiplicity of epistemologies and ontologies that exist in the world” (Lewis et al. 2018, 14). Building on Indigenous perspectives that acknowledge kinship networks that extend to non-humans, the authors propose to “make kin with the machine” as an alternative to escape western epistemology that “does not account for all members of the community and has not made it possible for all members of the community to survive let alone flourish” (Lewis et al. 2018, 10). There is potential in exploring Indigenous frameworks to rethink the design and governance of social media platforms. This exploration should happen without romanticizing Indigenous knowledges, as Milan and Treré (2019) warn, but instead “exploring it in all its contradictory aspects” in order to allow for diverse ways of understanding the productions of meaning making on social media (Milan and Treré 2019, 325-326).

The third and last research question in our study focused on methodological and ethical challenges. The capacity to extract large amounts of data from “public” social media platforms have led to malpractices in the field, which is especially problematic in studies involving vulnerable communities. Digital media scholars have increasingly critiqued the overuse of Twitter’s “easy data” available through standard API access in social media research (Burgess and Burns 2015), including in studies of racism. Digital methods research should avoid perpetuating historical processes of dispossession through nonconsensual data extraction from marginalised communities, embrace user privacy by not synonymising user acceptance of platform ToS with informed consent, and pay attention to power, vulnerability, and subjectivity (Florini et al. 2018; Leurs 2017; Linabary and Corple 2019; Milan and Treré 2019).

Another point of critique regarding the literature is a tendency in the qualitative works to reproduce posts verbatim, which can easily lead to identification even though the users are anonymised. The exception that proves the rule is one study, in which the

researchers asked the Twitter users identified in their dataset, whether they could include their tweets in their analysis (Petray and Collin 2017). Sanderson et al. (2016) also note that it is preferable to contact people when assessing intention on social media. Some justifications observed in our sample as to why informed consent was not obtained seemed a bit flawed, like justifying reporting on data extracted from a private social media space because “with tens of thousands of members”, private Facebook groups “cannot be considered a private space in any meaningful sense” (Allington 2018, 131). From a feminist approach to privacy on social media research, Linabary and Corple (2019) note the importance of informed consent and invite researchers to think carefully about how data collection and analyses can put social media users at risk. As a solution to the impracticability of obtaining informed consent in big data studies, Linabary and Corple (2019) suggest: “Individuals who scrape data from websites, forums, or listservs can use these same platforms for posting about their work and eliciting participant feedback” (p. 1458).

Conclusion

This article has provided a review and critique of scholarly research on racism, hate speech, and social media, focusing in particular on methodological, theoretical, and ethical challenges of the field and critically discussing their implications for future research. Departing from Daniels’s literature review from 2013, the article has focused on developments in the years 2014 to 2018 in the subfield of social media research. Scholarly work on racism and social media has come a long way since Daniels’s article, which only briefly touched upon social media as novel spaces. There are new insights coming out of our review. *First*, while studies of social media and racism have certainly become more prominent, as Daniels forecasted, there is a dire need for a broader range of research, going beyond text-based analyses of overt and blatant racist speech, Twitter, and the United States and into the realm of wider geographical contexts, further platforms, multiplatform analyses, and thorough examinations of how racism on social media is ordinary, everyday, and often mediated through the visual. *Second*, we echo Daniels’s concern about the need for more scholarly work that pays attention to the structural nature of racism by interrogating how race is baked into social media technologies’ design and governance rather than just focusing on racist expression in these spaces. *Third*, we argue that a factor that contributes to neglecting the role of race in the subfield is the lack of reflexivity in research designs. There is a preponderance of research on racism, hate speech, and social media done by white scholars that rarely acknowledges the positionality of the authors, which risks reinforcing colour-blind ideologies within the field. To this, ethical malpractices within social media research can inadvertently reproduce historical power imbalances. *Fourth*, there are clear limitations in centring “hate speech” to approach the moderation and regulation of racist content. Not only is “hate speech” a contested term in a definitional sense, but a focus on illegal hate speech risks conceptualising racism on social media as something external to platforms that can be simply fought through technical fixes such as machine learning. *Last*, although we found that some authors followed Daniels’s call to explore the ideas of critical authors, such

as DuBois (1903) and Feagin (2006) for more robust understandings of how whiteness contributes to perpetuating racist systems, this work is still a minority in the field. We double down on Daniels and other scholars' call for a commitment to critical race perspectives to interrogate the inner workings of social media platforms. In this regard, we suggest that scholars interested in advancing the field could benefit from exploring new emerging work that is using Indigenous critical perspectives to explore race struggles on social media. We hope this review and critique will inform future research on the complex topic of racism on social media and best practices on how to study it.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ariadna Matamoros-Fernández  <http://orcid.org/0000-0003-2149-3820>

Note

1. Of these, 17 were excluded based on sampling criteria (1), five were excluded based on criteria (2), 82 were excluded based on criteria (3), 61 were excluded based on criteria (4), and one was excluded based on criteria (5).

References

(Papers in the corpus are marked by *)

- *Allington, Daniel. 2018. "Hitler Had a Valid Argument against Some Jews": Repertoires for the Denial of Antisemitism in Facebook Discussion of a Survey of Attitudes to Jews and Israel." *Discourse, Context and Media* 24: 129–36.
- Angwin, Julia, and Terry Parris Jr. 2016. "Facebook Lets Advertisers Exclude Users by Race." *ProPublica*, October 28, 2016. <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>.
- *Awan, Imran. 2014. "Islamophobia and Twitter: A Typology of Online Hate against Muslims on Social Media." *Policy and Internet* 6 (2): 133–50.
- Bivens, Rena, and Oliver L. Haimson. 2016. "Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers." *Social Media+ Society* 2 (4): 2056305116672486.
- *Black, Erik W., Kelsey Mezzina, and Lindsay A. Thompson. 2016. "Anonymous Social Media - Understanding the Content and Context of Yik Yak." *Computers in Human Behavior* 57: 17–22.
- Bliuc, Ana-Maria, Nicholas Faulkner, Andrew Jakubowicz, and Craig McGarty. 2018. "Online Networks of Racial Hate: A Systematic Review of 10 Years of Research on Cyber-Racism." *Computers in Human Behavior* 87: 75–86.

- *Bock, Sheila. 2017. "Ku Klux Kasserole and Strange Fruit Pies: A Shouting Match at the Border in Cyberspace." *Journal of American Folklore* 130 (516): 142–65.
- Bonilla-Silva, Eduardo. 2010. *Racism Without Racists: Colorblind Racism and the Persistence of Racial Inequality in the United States*. Lanham: Rowman & Littlefield Publishers Inc.
- *Bonilla, Yarimar, and Jonathan Rosa. 2015. "#Ferguson: Digital Protest, Hashtag Ethnography, and the Racial Politics of Social Media in the United States." *American Ethnologist* 42 (1): 4–17.
- *Bosch, Tanja. 2017. "Twitter Activism and Youth in South Africa: The Case of #RhodesMustFall." *Information Communication and Society* 20 (2): 221–32.
- Brock, André. 2012. "From the Blackhand Side: Twitter as a Cultural Conversation." *Journal of Broadcasting and Electronic Media* 56 (4): 529–49.
- *Brown, Melissa, Rashawn Ray, Ed Summers, and Neil Fraistat. 2017. "#SayHerName: A Case Study of Intersectional Social Media Activism." *Ethnic and Racial Studies* 40 (11): 1831–46.
- Burgess, Jean, and Axel Bruns. 2015. "Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research After the Computational Turn." In *Compromised Data: From Social Media to Big Data*, edited by Ganaele Langlois, Joanna Redden, and Greg Elmer, 93–111. New York, NY: Bloomsbury.
- Burgess, Jean, Alice Marwick, and Thomas Poell, eds. 2017. *The Sage Handbook of Social Media*. London: SAGE Publications.
- *Burnap, Pete, and Matthew L. Williams. 2016. "Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics." *EPJ Data Science* 5 (1): 11.
- Carlson, Bronwyn, and Ryan Frazer. 2020. "They Got Filters": Indigenous Social Media, the Settler Gaze, and a Politics of Hope." *Social Media + Society* 6 (2): 2056305120925261.
- Carlson, Bronwyn, and Tanja Dreher. 2018. "Introduction: Indigenous Innovation in Social Media." *Media International Australia* 169 (1): 16–20.
- *Carney, Nikita. 2016. "All Lives Matter, but so Does Race: Black Lives Matter and the Evolving Role of Social Media." *Humanity & Society* 40 (2): 180–99.
- Chakravartty, Paula, Rachel Kuo, Victoria Grubbs, and Charlton McIlwain. 2018. "#CommunicationSoWhite." *Journal of Communication* 68: 254–66.
- *Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech." Proceedings of the ACM on Human-Computer Interaction, New York. <https://doi.org/10.1145/3134666>.
- *Chaudhry, Ifran. 2015. "#Hashtagging Hate: Using Twitter to Track Racism Online." *First Monday* 20 (2). doi:10.5210/fm.v20i2.5450.
- Chun, Wendy Hui Kyong. 2009. "Introduction: Race and/as Technology; or, How to Do Things to Race." *Camera Obscura* 24 (1): 7–35.
- *Cisneros, J. David, and Thomas K. Nakayama. 2015. "New Media, Old Racisms: Twitter, Miss America, and Cultural Logics of Race." *Journal of International and Intercultural Communication* 8 (2): 108–27.
- Collins, Patricia Hill. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. 1st ed. New York: Routledge.
- Crenshaw, Kimberle. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." *Stanford Law Review* 43 (6): 1241–99.
- Cunningham, Stuart, and Flew Terry. 2000. "De-Westernising Australia: Media Systems and Cultural Coordinates." In *De-Westernizing Media Studies*, edited by James Curran and Myung-Jin Park, 210–220. New York: Routledge.

- Daniels, Jessie. 2013. "Race and Racism in Internet Studies: A Review and Critique." *New Media & Society* 15 (5): 695–719.
- *Davidson, Thomas, Dana Warmlesley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM 2017, Montreal, Québec, Canada. <https://arxiv.org/pdf/1703.04009.pdf>
- Delgado, Richard, and Jean Stefancic. 2001. *Critical Race Theory: An Introduction*. New York, NY: New York University Press.
- Du Bois, William Edward B. 1903. *The Souls of Black Folk*. Chicago, IL: A.C. McClurg Press.
- *Eddington, Sean M. 2018. "The Communicative Constitution of Hate Organizations Online: A Semantic Network Analysis of 'Make America Great Again.'" *Social Media + Society* 4 (3): 205630511879076.
- *Elsherief, Mai, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. "Peer to Peer Hate: Hate Speech Instigators and Their Targets." <https://arxiv.org/abs/1804.04649>
- *Farkas, Johan, Jannick Schou, and Christina Neumayer. 2018. "Cloaked Facebook Pages: Exploring Fake Islamist Propaganda in Social Media." *New Media & Society* 20 (5): 1850–67.
- Feagin, Joe R. 2006. *Systemic Racism: A Theory of Oppression*. New York: Routledge.
- Florini, Sarah, André Brock, Catherine Knight Steele, Kishonna Gray, and Miriam Sweeney. 2018. "Digital Critical Race Mixtape." Montreal QC, Canada. <https://aoir2018.sched.com/event/HP50/digital-critical-race-mixtape>.
- Frankenberg, Ruth. 1993. *White Women, Race Matters: The Social Construction of Whiteness*. Minneapolis, MN: University of Minnesota Press.
- *Gantt-Shafer, Jessica. 2017. "Donald Trump's 'Political Incorrectness': Neoliberalism as Frontstage Racism on Social Media." *Social Media and Society* 3 (3): 1–10 doi:10.1177/2056305117733226.
- *George Mwangi, Chrystal A., Genia M. Bettencourt, and Victoria K. Malaney. 2018. "Collegians Creating (Counter)Space Online: A Critical Discourse Analysis of the I, Too, Am Social Media Movement." *Journal of Diversity in Higher Education* 11 (2): 146–63.
- *Gin, Kevin J., Ana M. Martínez-Alemán, Heather T. Rowan-Kenyon, and Derek Hottell. 2017. "Racialized Aggressions and Social Media on Campus." *Journal of College Student Development* 58 (2): 159–74.
- Goldberg, David Theo, and Ato Quayson, eds. 2002. *Relocating Postcolonialism*. 1st ed. Oxford, UK; Malden, MA: Wiley-Blackwell.
- *Hong, Lingzi, Weiwei Yang, Philip Resnik, and Vanessa Frias-Martinez. 2016. "Uncovering Topic Dynamics of Social Media and News: The Case of Ferguson." In *SocInfo 2016: Social Informatics*, edited by E. Spiro and Y. Y. Ahn, 240–56. Cham, Switzerland: Springer.
- *Ilmonen, Kaisa. 2016. "Beyond the Postcolonial, but Why Exactly? Ten Steps towards a New Enthusiasm for Postcolonial Studies." *Journal of Literary Theory* 10 (2): 345–365. <https://doi.org/10.1515/jlt-2016-0013>.
- Jackson, Lauren Michele. 2017. "We Need to Talk About Digital Blackface in Reaction GIFs." *Teen Vogue*, August 2. <https://www.teenvogue.com/story/digital-blackface-reaction-gifs>.
- Jenkins, Jennifer. 2013. *English as a Lingua Franca in the International University: The Politics of Academic English Language Policy*. London: Routledge.
- Jerkins, Megan. 2015. "The Quiet Racism of Instagram Filters." *Racked.Com*, July 7. <https://www.racked.com/2015/7/7/8906343/instagram-racism>.
- *Johns, A. 2017. "Flagging White Nationalism 'After Cronulla': From the Beach to the Net." *Journal of Intercultural Studies* 38 (3): 349–64.

- *Johnson, Brett G. 2018. "Tolerating and Managing Extreme Speech on Social Media." *Internet Research* 28 (5): 1275–91.
- *Keskinen, S. 2018. "The 'Crisis' of White Hegemony, Neonationalist Femininities and Antiracist Feminism." *Women's Studies International Forum* 68: 157–63.
- *Lamerichs, Nicolle, Dennis Nguyen, Mari Carmen, Puerta Melguizo, Radmila Radojevic, and Anna Lange-Böhmer. 2018. "Elite Male Bodies: The Circulation of Alt-Right Memes and the Framing of Politicians on Social Media." *Participations* 15 (1): 180–206.
- Leurs, Koen. 2017. "Feminist Data Studies: Using Digital Methods for Ethical, Reflexive and Situated Socio-Cultural Research." *Feminist Review* 115 (1): 130–54.
- Lewis, Jason Edward, Noelani Arista, Archer Pechawis, and Suzanne Kite. 2018. "Making Kin with the Machines." *Journal of Design and Science*. Published electronically July 16, 2018. doi:10.21428/bfefd97b.
- Linabary, Jasmine R., and Danielle J. Corple. 2019. "Privacy for Whom?: A Feminist Intervention in Online Research Practice." *Information Communication and Society* 22 (10): 1447–63.
- *Magu, Rijul, Kshitij Joshi, and Jiebo Luo. 2017. "Detecting the Hate Code on Social Media." Eleventh International AAAI Conference on Weblogs and Social Media (ICWSM) 2017. <http://arxiv.org/abs/1703.05443>.
- Montréal, Québec, Canada. <http://arxiv.org/abs/1703.05443>.
- *Mason, Corinne Lysandra. 2016. "Tinder and Humanitarian Hook-Ups: The Erotics of Social Media Racism." *Feminist Media Studies* 16 (5): 822–37.
- Massanari, Adrienne. 2015. "# Gamergate and The Fapping: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." *New Media & Society* 19 (3): 329–46.
- *Matamoros-Fernández, Ariadna. 2017. "Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube." *Information, Communication & Society* 20 (6): 930–46.
- *Matamoros-Fernández, Ariadna. 2018. "Inciting Anger through Facebook Reactions in Belgium: The Use of Emoji and Related Vernacular Expressions in Racist Discourse." *First Monday* 23 (9): 1–20.
- Matsuda, Mari J., Charles R. Lawrence III, Richard Delgado, and Kimberle Williams Crenshaw. 1993. *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment*. 1st ed. Boulder, Colo: Westview Press.
- *McCosker, Anthony, and Amelia Johns. 2014. "Contested Publics: Racist Rants, Bystander Action and Social Media Acts of Citizenship." *Media International Australia* 151: 66–72.
- McIntosh, Peggy. 2003. "White Privilege: Unpacking the Invisible Knapsack." In *Understanding Prejudice and Discrimination*, edited by Scott Plous, 191–6. New York: McGraw-Hill.
- Milan, Stefania, and Emiliano Treré. 2019. "Big Data from the South(s): Beyond Data Universalism." *Television and New Media* 20 (4): 319–35.
- Miškolci, Jozef, Lucia Kováčová, and Edita Rigová. 2020. "Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia." *Social Science Computer Review* 38 (2): 128–46.
- *Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. "A Measurement Study of Hate Speech in Social Media." Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 85–94. <https://doi.org/10.1145/3078714.3078723>.
- Moreton-Robinson, Aileen. 2015. *The White Possessive*. Minneapolis, MN; London: University of Minnesota Press.

- *Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39 (3): 629–49.
- *Murthy, Dhiraj, and Sanjay Sharma. 2019. "Visualizing YouTube's Comment Space: Online Hostility as a Networked Phenomena." *New Media and Society* 21 (1): 191–213.
- Nakamura, Lisa. 2008. *Digitizing Race: Visual Cultures of the Internet*. Minneapolis, MN: University of Minneapolis Press.
- Nakata, Martin N. 2007. *Disciplining the Savages: Savaging the Disciplines*. Canberra, Australia: Aboriginal Studies Press.
- Nishi, Naomi W., Cheryl E. Matias, and Roberto Montoya. 2015. "Exposing the White Avatar: Projections, Justifications, and the Ever-Evolving American Racism." *Social Identities* 21 (5): 459–73. <https://doi.org/10.1080/13504630.2015.1093470>
- Noble, Safiya Umoja. 2018a. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- *Noble, Safiya Umoja. 2018b. "Critical Surveillance Literacy in Social Media: Interrogating Black Death and Dying Online." *Black Camera* 9 (2): 147.
- Noble, Safiya Umoja, and Brendesha M. Tynes, eds. 2016. *The Intersectional Internet: Race, Sex, Class, and Culture Online*. New York: Peter Lang Inc., International Academic Publishers.
- O'carroll, Acushla Deanne. 2013. "Virtual Whanaungatanga: Māori Utilizing Social Networking Sites to Attain and Maintain Relationships." *AlterNative: An International Journal of Indigenous Peoples* 9 (3): 230–45. <https://doi.org/10.1177/117718011300900304>.
- *Oh, David C. 2016. "'Payback for Pearl Harbor': Racist Ideologies Online of Karmic Retribution for White America and Postracial Resistance." *Journal of Communication Inquiry* 40 (3): 247–66.
- *Olteanu, Alexandra, Carlos Castillo, and Jeremy Boy. 2018. "The Effect of Extremist Violence on Hateful Speech Online." <https://arxiv.org/abs/1804.05704>
- Omi, Michael, and Howard Winant. 1994. *Racial Formation in the United States*. 2nd edition. New York: Routledge.
- *Page, Janis Teruggi, Margaret Duffy, Cynthia Frisby, and Gregory Perreault. 2016. "Richard Sherman Speaks and Almost Breaks the Internet: Race, Media, and Football." *Howard Journal of Communications* 27 (3): 270–89.
- *Pater, Jessica A., Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. "Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms." GROUP '16, ACM, 369–74. <https://dl.acm.org/doi/10.1145/2957276.2957297>
- *Patton, Desmond Upton, Douglas-Wade Brunton, Andrea Dixon, Reuben Jonathan Miller, Patrick Leonard, and Rose Hackman. 2017. "Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use of Social Media for Identification of Criminal Conduct and Associations." *Social Media and Society* 3 (3): 205630511773334.
- *Petray, Theresa L., and Rowan Collin. 2017. "Your Privilege Is Trending: Confronting Whiteness on Social Media." *Social Media and Society* 3 (2): 205630511770678.
- Picca, Leslie Houts, and Joe R. Feagin. 2007. *Two-Faced Racism: Whites in the Backstage and Frontstage*. New York, NY: Routledge.
- Porter, Jon. 2019. "TikTok Unblocks US Teen Who Slammed China for Uighur Treatment." *The Verge*, November 28. <https://www.theverge.com/2019/11/28/20986867/tiktok-unblock-us-teen-china-criticism-muslim-minority-terrorist-imagery-moderation-guidelines>.
- *Puschmann, Cornelius, Julian Ausserhofer, Noura Maan, and Markus Hametner. 2016. "Information Laundering and Counter-Publics: The News Sources of Islamophobic Groups on Twitter." *The Workshops of the Tenth International AAAI Conference on Web and Social Media Social Media in the Newsroom: Technical Report WS-16-19, May 17-20, 2016*,

- Cologne, Germany, 143–50. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13224/12858>.
- Raynauld, Vincent, Emmanuelle Richez, and Katie Boudreau Morris. 2018. "Canada Is #IdleNoMore: Exploring Dynamics of Indigenous Political and Civic Protest in the Twittersverse." *Information, Communication & Society* 21 (4): 626–42.
- Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. 1st ed. New Haven, CT: Yale University Press.
- *Saleem, Haji Mohammad, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. "A Web of Hate: Tackling Hateful Speech in Online Social Spaces." <https://arxiv.org/abs/1709.10159>.
- *Sanderson, Jimmy, Evan Frederick and Mike Stocz. 2016. "When Athlete Activism Clashes With Group Values: Social Identity Threat Management via Social Media." *Mass Communication and Society* 19 (3): 301–22.
- *Shepherd, Tamara, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. 2015. "Histories of Hating." *Social Media and Society* 1 (2): 1–10. doi:10.1177/2056305115603997.
- Sue, Derald Wing. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Hoboken, NJ: Wiley.
- *Taylor, Jherez, Melvyn Peignon, and Yi-Shin Chen. 2017. "Surfacing Contextual Hate Speech Words within Social Media." <http://arxiv.org/abs/1711.10093>.
- *Tulkens, Stéphan, Lisa Hilde, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. "The Automated Detection of Racist Discourse in Dutch Social Media." *Computational Linguistics in the Netherlands Journal* 6 (March): 3–20.
- *Waseem, Zeerak, and Dirk Hovy. 2016. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." Proceedings of the NAACL Student Research Workshop, San Diego, California, 88–93. <https://doi.org/10.18653/v1/N16-2013>.
- *Williams, Matthew L., and Pete Burnap. 2016. "Cyberhate on Social Media In The Aftermath of Woolwich: A Case Study In Computational Criminology And Big Data." *British Journal of Criminology* 56 (2): 211–38.
- Wilson, Alex, Bronwyn Carlson, and Acushla Sciascia. 2017. "Reterritorialising Social Media: Indigenous People Rise Up." *Australasian Journal of Information Systems* 21 (Kamelamelala 2016): 1–4.
- Wolfe, Patrick. 2016. *Traces of History: Elementary Structures of Race*. London; New York: Verso.

Author Biographies

Ariadna Matamoros-Fernández is a Lecturer in Digital Media at the Queensland University of Technology, Chief Investigator at the Digital Media Research Centre (DMRC) and Associate Investigator at the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S). She is setting up a research agenda around platform governance in relation to memes and other controversial humorous content, and has extensive experience in building digital methods to study digital platforms. Her research has been published in *Information, Communication & Society*, *Convergence*, and other international peer-reviewed journals.

Johan Farkas is a PhD fellow at Malmö University, Sweden. His research engages with the intersection of digital media platforms, disinformation, oppression, politics, and democracy. Farkas has published on these topics in international journals such as *New Media & Society*, *Social Media + Society*, and *Critical Discourse Studies*. In 2019, his debut book, *Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood* (with Jannick Schou) was published by Routledge.