Extreme speech

Udupa, Sahana

Erstveröffentlichung / Primary Publication

Sammelwerksbeitrag / collection article

Mitglied der

Leibniz-Gemeinschaft

gesis
Leibniz-Institut
für Sozialwissenschaften

Abstract: Extreme speech is a critical conceptual framework that aims to uncover vitriolic online cultures through comparative and ethnographic excavations of digital practices. It is not one more new definition or a term replaceable with *extremist speech.* Rather, it is a conceptual framework developed to foreground historical awareness, critical deconstruction of existing categories, and a grounded understanding of evolving practices in online communities, in ways to holistically analyze the contours and consequences of contemporary digital hate cultures. This framework suggests that the *close contextualization* of proximate contexts—of media affordances in use or situated speech cultures—should accompany *deep contextualization*, which accounts for grave historical continuities and technopolitical formations unfolding on a planetary scale. Through such elaborate forays into everyday practices and deeper histories, extreme speech theory proposes to nuance normative and regulatory efforts to classify and isolate hate speech and disinformation.

Sahana Udupa

Extreme Speech

1   Introduction

Extreme speech is a critical conceptual framework that has drawn attention to online vitriolic cultures by ethnographically analyzing digital practices and online user communities from a comparative, historically sensitive perspective.

The concept of *extreme speech* departs, in particular, from the dominant legal-normative definitions of *hate speech* and the discourse of securitization around *terrorism* and *political extremism*. These definitions approach *hate speech* primarily as a discourse of pathology by predetermining the effects of online volatile speech as vilifying, polarizing, or lethal. *Extreme speech* instead stresses the importance of holistic comprehension over classification by placing such practices in a broader context of contestations over power and allowing normative approaches and mitigation efforts to emerge from a grounded, historically aware analysis (Pohjonen & Udupa 2017; Udupa, 2015, 2017, 2019, 2020; Udupa et al., 2021; Udupa & Pohjonen, 2019). In this sense, the contributions of extreme speech research *qualify*, rather than seek to replace, the existing repertoire by highlighting areas that hate speech research has insufficiently explored as well as by drawing attention to the political consequences of hate speech discourse.

In terms of its definitional scope, extreme speech analysis focuses on derogatory speech forms aimed at any group (including groups that hold power) and

exclusionary discourses with hateful language and expressions dressed as *facts* that implicitly or explicitly exclude or harm a person or group on the basis of their group belonging. Derogatory extreme speech is particularly ambivalent since it represents online discourses that challenge the protocols of polite language to speak back to power, but it also constitutes a volatile slippery ground on which what is comedic and merely insulting could quickly slide down to downright abuse and threat. For content moderation, such derogatory expressions can serve as the earliest cultural cues to brewing and more hardboiled antagonisms. The analysis of exclusionary extreme speech builds on existing definitional standards of *hate speech* set up by the United Nations and Wardle and Derakhshan's (2017) distinction between *disinformation* ("when false information is knowingly shared to cause harm") and *malinformation* ("when genuine information is shared to cause harm") (p. 5). Extreme speech analysis covers *misinformation* (spreading false information without an intention to cause harm) so far as it is part of the social fields, where deliberate efforts to spread hate activate a variety of actors and networks, which ultimately spread hateful language that could harm vulnerable groups. The purpose of extreme speech analysis is, therefore, to exceed the legal focus on culpability and instead analyze—with ethnographic and historical depth—how different actors and actions animate one another and how new interventions must be crafted to address not only actors who deliberately engineer hateful language and disinformation but also those who succumb to it or do it to earn a livelihood. This approach allows researchers and policymakers to chart new analytical pathways and diverse fields of action, beyond intentionality-based investigations.

The focus on cultural practice is especially important for extreme speech analysis. In particular, it calls for ethnographic explorations of media cultural practice—that is, what people do that relates to media (Couldry, 2012) within particular structural conditions that shape and are shaped by such practices. The practice approach emphasizes that political configurations of discourses and inherited dispositions prefigure mediated action inasmuch as users' situated practices alter political discourse.

These analytical moves require situating the contemporary moment of online volatile speech within regional and historical contexts—ranging from the micro-contexts of online user cultures and the contradictory pulls of realpolitik to macro-historical formations of colonial imperialism—as a necessary corrective to the seeming universality of the normative basis of the hate speech discourse.

Coloniality is conceptualized as a global process that institutionalized and legitimized three sets of relations—market relations, nation state relations and racial relations—that constitute a composite structure of oppression with impacts beyond the actual geographies that the Empire colonized. For this reason, coloniality is a relevant critical framework to understand contemporary formations of inequality and repression, including those articulated through digital mediations and how they are especially shaped by digital capitalist logics that facilitate and exacerbate vectors of difference. Such a historically contextualized understanding calls for a comparative analysis that looks beyond the West, extending its focus into the rapidly expanding online worlds of the Global South. The comparative approach here is not based on a model with quantitative metrics that are tested across selected case studies; rather, it is rooted in ethnography of practice and historical anthropology (Van der Veer, 2016).

2   The limits of hate speech

Key interventions of the conceptual framework of extreme speech have emerged from highlighting the limits of the hate speech discourse while also recognizing its significance as a regulatory concept. The legal-regulatory terminology of *hate speech* draws on longer legal debates over speech restrictions (Nockleby, 2000; Udupa et al., 2020; Warner & Hirschberg, 2012). Although legal traditions and scholarly discussions differ, a common element throughout this discourse is the assumption that hate speech involves the disparagement of other groups, based on their belonging to a particular group with a collective identity. Waldron (2012) argues that this kind of speech has two key characteristics: the first is to dehumanize members who belong to another group, and the second is to reinforce the boundaries of the in-group against the out-group by attacking the other group's members. Hate speech discourse predefines the effects of hate speech as *negative* and *damaging*, and its regulatory rationale is, thus, of control and containment. The state is the largest actor in this effort, but internet intermediaries also increasingly monitor and restrict speech on their platforms. Responding to civil society concerns, governmental injunctions, and international conventions on hate speech, online forums and social networking sites have developed their own terms of service to detect, regulate, and prohibit hate speech.

As it jostles between state regulation, the capitalist market, and political fields, hate speech has become what Brubaker and Cooper (2000) would describe as a *thick concept* with a "tangle of meanings" and an evaluative load (p. 14). Moreover, these concepts become empirical objects in themselves; the researcher's task would be merely to discover the degree of variance or agreement between different kinds of online speech from this ideal object type. Extreme speech calls such contextual flattening into question.

Furthermore, as a form of power, the discourse of hate speech is inextricably tied to the state and its political economies of violence. Historically, it emerged from the projects of civility that coincided with (and partly constituted) the state's monopolization of violence (Giddens, 1987; Thirangama et al., 2018). The moral claims of liberal thought require that hate speech regulation protects substantive virtues, such as sympathy and understanding (at least in the procedural terms of decorum), in the interest of a common good. Liberal understandings premised on abstract principles of equality conceal multifarious and, at times, manipulative political agendas that have grown around the regulatory discourse of hate speech.

Moreover, the liberal moral principle of civility that partly informs the rationale of hate speech is "intimately tied up with class and race privilege," which consolidated the colonial and postcolonial state (Thirangama et al., 2018, pp. 153–155). Colonial histories have cemented the self-righteous schema of the liberal center (the self-understanding of the West) and extreme periphery (the rendering of the non-West), which is now manifest in diverse forms of political grandstanding and control not only between the (former) metropole and colony but also within the nation-states where similar structures of speech restriction, based on moral self-understandings, have taken root.

Under these conditions, the pressure to speak the *polite* language has been an act of domination—moral injunctions linked to assertions of privilege. Civility, thus, is an "effect of political recognition and of a responsive structure of authority" (Mitchell, 2018, p. 217). In other words, the implications of incivility—or the extremeness of speech more broadly—cannot be apprehended without analyzing particular forms of recognition and responsiveness to demands that are available to diverse groups.

The thick concept of *hate speech* comes with an evaluative load aimed at immediate action, raising the risk of glossing over historical trajectories, as well

as the ambivalence of extremeness within particular contexts of power. This is not merely a fine grained theoretical objection but also, more gravely, a political problem. Both historically and in the contemporary moment, the ambivalence of extreme speech is closed off when political actors who are pressured to do something about hate invoke the label of *hate speech* (Pohjonen, 2019), at times brutally using force to target marginalized groups. Examples abound of regimes misusing the hate speech discourse to squash dissent or target minoritized groups. In the context of India, currently ruled by a Hindu nationalist regime, selective application of state restrictions on online speech has cited the "law and order" rationale, invoking the legally imprecise term of *hate speech* in conjunction with colonial legislations around sedition or outraging religious feelings (Modh, 2015). Such restrictions on the national level have sought to quell dissenting voices, while regional governments with diverse ideological agendas, set in a multiparty system of competitive electoral politics, have mobilized similar efforts to frame political opposition as *hate speech*.

In Kenya's context, Katiambo (2021) has argued that "the polysemy of extreme speech is removed when *incivility* becomes known as hate speech, blocking us from ever knowing its alternative possibilities" (p. 49). In everyday conversational contexts, *hate speech* is often used as a charge or an accusation that closes off, rather than opens up, avenues for change and dialogue (Boromisza-Habashi, 2013).

Recognizing the limits of hate speech both as a regulatory value and a concept-in-use in everyday interactions, ethnographic sensibility advocated by extreme speech research insists that the moral charge around vitriol and disinformation should come from lived concepts and situated contexts, rather than frameworks imposed from the outside. This shift requires a critical approach that is sensitive to cultural variations in speech, including sanctioned forms of disrespect; political contexts where hate, as an order value of regulation, is assigned to speech acts; and historical conditions that implicate extreme speech with particular forms of power—subversive in some contexts and repressive in others.

3   Extreme speech as methodology

The conceptual framework of extreme speech comes with a set of methodological perspectives.

3.1 Comparative Practice

Extreme speech research proposes to map a critical typology of vitriol based on historical, cultural, and political variations, and a focus on media-cultural practice described in the preceding section. This methodological approach might be described as *comparative practice*, where interlocking factors in different national and regional scenarios are studied for their specificities and in relation to one another.

3.2 Everydayness and emic categories

Drawing from an anthropological emphasis on everyday cultures, extreme speech research draws attention to *emic* categories (i.e., categories derived from the perspectives of research participants than the observer), through which the complex use of language operates. Methodologically, it involves exploring the meanings that online users, as historical actors, attach to *vitriol* and the diverse practices that congeal around them.

Online *gaali*, in the Indian context, might illustrate such an emic category (Udupa, 2017). *Gaali* is a Hindi term for a complex amalgam of abrasive, abusive, or unabashed language seen as joking and disrespectful at the same time. It is a commonly invoked term to define the aggressive styles of online debating cultures. Online gaali has provided new avenues of participation for politically savvy internet users, especially among the educated middle-class groups in urban India and diverse class groups with access to mobile media who feel confident that they can trump legacy media and political authorities by engaging in social media discussions. While anti-establishment *gaali* does not always articulate progressive politics, gaali's performative spread has, nonetheless, brought new political voices to the fore of public debate. By online actors' own account, gaali—as rancorous rabble-rousing—has helped them thrust their voices into the public domain hitherto dominated by the state and organized commercial media. Consequently, gaali has sparked voluminous online contestations around the developmental, representational, and economic issues facing contemporary India.

At the same time, the blurred arena of online comedy, insult, and abuse that gaali represents has facilitated the perpetuation of religious majoritarian nationalism and

exclusionary discourses centering on assertions of Hindu-first India. Often, online gaali grows into a full-blown shaming punishment, articulating nationalism through the gendered trope of regulating sexuality and what Irvine (1993) calls "evaluative talk" (p. 106). Online gaali as gendered abuse has led to severe cases of intimidation and harassment against female online commentators.

Nested in digital culture but drawing on longer histories, gaali has spawned the interlocking practices of insult, comedy, shame, and abuse that unfold in a blurred arena of online speech. On this slippery ground of shifting practices, comedy stops and insult begins or insult morphs into abuse in mutually generative ways. Contextually sensitive analysis reveals, in this case, gaali's Janus-faced status as performance; while its routine detoxification opens up new lines of participation in political discourse to online users, it takes a menacing edge when they instantiate gendered discursive relations of nationalism.

3.3   *Empathy and reflexivity*

Other key methodological approaches of extreme speech research include reflexivity and empathy. It is difficult to develop access to complex ground realities that are rife with contradictions without sustained ethnographic engagement among communities even when such communities harbor despicable or less than ideal political views. Sustained engagement comes with a commitment to extend the same principles of honesty and openness that inform a sound ethnographic practice.   Arguably, the foremost ethical principle in advancing such an ethnographic sensibility is empathy, which is guided by a commitment to learn and see insider views as a *working morality.* Empathy as a practical or working morality in ethnographic practice does not, in itself, entail an endorsement of the views expressed by online actors or claims to moral equivalence between different ideological positions. As researchers explore the political implications of digital practice in each case to its fullest possible detail, empathy as an ethical stance allows them to avoid a tendency for critique to precede understanding or for a moral-evaluative framework to predetermine what to expect. Empathy demands active dispositions on the researcher's part, foremost a commitment to listen to actors who inhabit the digital world through myriad expressions, aspirations, habits, and tactics, including those aimed at advancing politically problematic

ideologies. As we, as researchers, navigate our interlocuters' diverse narratives and life worlds, anthropological reflexive praxis is especially pertinent since our own positionalities are intricately interwoven with digital discourses, and our material, social, and political circumstances shape the ethical, affective, and political terms with which we approach online speech as problematic or otherwise (Udupa & Dattatreyan, 2023).

4   Global conjuncture and deep contextualization

These methodological moves are important in advancing a *conjunctural analysis* of varied forces, rather than assessing social and political worlds based on predefined normative categories (Mankekar, 1999). By the same token, inasmuch as extreme speech stresses the analytical value of highlighting ambiguity in online speech, it is methodologically equipped to examine the diverse factors that shape particular political formations. In this sense, extreme speech avows ethnographic specificities—but in ways that connect contexts and situate them within socio-technological transformations that are unfolding on a global scale and in relation to long-standing historical processes.

Gleaning from cases around the world, extreme speech analysis has highlighted that, over the last two decades, vitriolic cultures have precipitated a condition of violent exclusion based on "exacerbated fracture lines of difference that include race, gender, sexuality, religion, nation and class" in a context where "computational capital has built itself and its machines out of those capitalized and technologized social differentiations" (Beller, 2017).

We define this condition as the global conjuncture of affects, actors, and affordances that is driving contemporary forms of exclusionary extreme speech. The socio-technological mediations of internet-based media are particularly significant in this conjuncture; we argue that they constitute a context in themselves, rather than acting as mere channels for discourses external to them. In particular, exclusionary extreme speech rides on digital affordances of peer-to-peer mobilizations, continuous exchange, platform migration, and layered anonymity.

Through the lens of Ahmed's (2004) semiotic analysis of affect, it is possible to see digital mediation as mechanisms that materialize the surfaces of hateful bodies through association, alignment, displacement, and "stickiness" (p. 89). If

hate is part of the "production of the ordinary" (Ahmed, 2004, p. 56), digital exchange has realized hate by bringing hateful expressions closer to one's everyday conversational realities. Tagging on to the small-screen intimacy of digital exchange, hate passes to the ordinary in continuous loops, powered by the systematic channeling of affect—of anger, glee, envy, and the transgressive pleasures of vitriolic online exchanges—within the participatory condition of digital capitalism (Udupa, 2020). I have argued that fun is a particularly significant affective infrastructure in ramping up online extreme speech among right-wing ideological communities in digital environments (Udupa, 2019). From quasi-public forums such as Twitter to image boards such as 4Chan, hate sticks to bodies through signs that are constantly innovated upon in *creative funny* ways, allowing the affective economy of hate to spread laterally between peers in solidarity.

Yet, far from a media-centric argument and claims that online affordances have let loose humankind's most primal animosities, extreme speech analysis highlights interconnections and continuities underwritten by longer historical processes. Exclusionary online extreme speech is shaped by the longer *global* process of colonial modern relations that unfolds as both internal and external forces in different societies. Colonial relations could be traced along three interconnected lines: nation-state relations established by colonial power, which frames the boundaries of minority-versus-majority and inside-versus-outside; market relations institutionalized by colonial power, which now manifest as uneven data relations; and racial relations naturalized by colonial power, which dispose people as objects of hatred (Udupa, 2020).

This analysis is a corrective to not only liberal moral panics about digital communication but also certain strands of Western left intellectualism that anxiously term ongoing digital turbulences as a "strange brew of bellicosity, disinhibition and rancor" among people who have been pushed to the wrong side of economic liberalization (Brown, 2019, p. 61). Such analysis elides the grave history of systematic violence that installed unequal racialized relations through actions—past and present—that are orchestrated, directed, and economic inasmuch as they are helpless reactions of backbiting revenge.

Following De Genova (2010), these historical conditions could be defined as postcolonial *metastasis.* Assertions of aggrieved power, common among White supremacists, emanate not only from structural subordination under oppressive market conditions but also through a sense of dethronement—a product of

far-reaching global imperial legacies. Crucially, through nation-state relations canonized by colonialism, this aggression wrought by imaginary wounds unfolds *within* different national and subnational contexts as racialized relations of majoritarian belligerence. Hindu nationalists in India, Sinhalese nationalists in Sri Lanka (Aguilera-Carnerero & Azeez, 2016), Han nationalists online in China (de Seta, 2021), the Sunni majoritarian politics around blasphemy in Pakistan (Schaflechner, 2021), Duterte's trolls in the Philippines (Ong & Cabanes, 2018), and online nationalists in Nepal (Dennis, 2017) are some examples, and so are the meme makers in northern Chile who seize internet memes' mashup cultures to portray migrants from Bolivia and Peru as backward, dirty, uneducated plunderers of limited resources and contributors to cultural degradation (Haynes, 2019). Such exclusionary discourses against *immigrants* (a category that emerged from the nation-state distinction between inside and outside) and *minorities* (a category that emerged from the nation-state distinction between a majority and a minority) are rife with racialized portrayals. Colonialism reproduced hierarchy and difference as intrinsic features of the modern nation-state, and this process of racialization of social relations within the newly stabilized structure of the nation-state alongside market relations was *global* in scope (Shankar, 2020; Treitler, 2013).

The framework of extreme speech has, thus, emphasized that longer historical processes should be examined in relation to proximate contemporary contexts of digital circulation and practice—a kind of dual analysis that might be described as *decolonial thinking*. This kind of analysis is not a macrohistorical glossing of diverse power dynamics. Without doubt, affective energies that emanate from and animate internet spaces should be analyzed in relation to specific structures of animosities and interlocking systems of coercion and power along various axes—including race, class, gender, religion, caste, nationality and ethnicity—that have precipitated the current global conjuncture of exclusionary extreme speech. Intersectionality invites attention to structures of power that predated, comingled or remained rather independent of colonial occupation. However, conceptualizing colonialism as a *set of relations* (market, nation-state and race) is important in tracking the overarching frameworks and historical continuities that undergird contemporary forms of exclusionary extreme speech. We might call this analysis *deep contextualization*. Decolonial thinking suggests that the *close contextualization* of proximate contexts—of media affordances in use or situated speech cultures—

should accompany *deep contextualization* that accounts for grave historical continuities and technopolitical formations unfolding on a planetary scale.

5   People-centric models of moderation

Through such elaborate forays into everyday practices and deeper histories, extreme speech theory proposes to nuance normative and regulatory efforts at classifying and isolating hate speech and disinformation. In this regard, regulatory and policy approaches honed by extreme speech perspectives call for people-centric models that can account for cultural variation, ambiguities, and dynamic forms of vitriolic online exchange.

An illustrative case might be the AI4Dignity project, a European Research Council-funded project that I run as the principle investigator. The project has partnered with independent fact-checkers from the Global North and the Global South as critical community intermediaries in developing artificial intelligence-assisted models for speech moderation. Recognizing that human supervision is critical, the project has devised ways to connect, support, and mobilize existing communities who have gained reasonable access to the meaning and context of speech because of their involvement in online speech moderation of some kind. Building spaces of direct dialogue and collaboration between artificial intelligence (AI) developers and relatively independent fact-checkers who are not part of a large media corporation, political party, or social media company is a key component of AI4Dignity. Moreover, this dialogue has involved academic researchers specialized in particular regions as facilitators. Through this triangulation, AI4Dignity's process model has aimed to stabilize a more encompassing collaborative structure in which *hybrid* models of human-machine filters can incorporate dynamic reciprocity between critical communities, such as independent fact-checkers, AI developers, and academic researchers. These efforts offer pathways to ground big data and computational methods with the extreme speech framework's emphasis on critical sensibility to cultural difference, historical contexts, local practices, and meanings drawn by users themselves in everyday lived environments.

Importantly, such efforts offer ways to bring inclusive training data sets to AI models. These datasets are more inclusive because they are based on culturally coded, linguistically diverse, and dynamic expressions that critical communities—

such as fact-checkers—can locate, rather than based on corporate social media definitions or annotations that natural langauge processing (NLP) experts develop within their professional fields. AI4Dignity's labeling process has involved reflexive and active iterations between ethnographers, communities, and AI developers. These iterations have, at times, led to confusing twists and turns in the annotation process, but they have also strengthened efforts to bring cultural nuance to data sets. For instance, at the beginning of the annotation process, confusion arose around the distinction between the three labels *derogatory extreme speech* (defined as expressions that do not conform to accepted norms of civility within specific local or national contexts and targeted at any group but not explicitly excluding vulnerable and historically disadvantaged groups; it includes derogatory jokes and sobriquets; Udupa, 2020), *exclusionary extreme speech* (defined as expressions that call for or imply excluding disadvantaged and vulnerable groups; Udupa, 2020) and *dangerous speech* (defined as expressions that have reasonable chances to trigger or catalyze harm and violence against target groups; Benesch, 2013). We had drawn this distinction based on published work and after some internal discussions with team members, but when we invited collaborating fact-checkers to categorize social media passages under one of these labels, several questions came up. A partnering fact-checker remarked that all extreme passages they encountered were indeed *dangerous* in the broadest sense of negatively affecting society. This opinion was, indeed, completely legitimate, but I requested that he appreciate efforts to keep the categories more precise because, once machine learning (ML) models begin to categorize, these mapped data sets could have regulatory implications. In the next round of discussions, we observed more clarity around the term *dangerous speech*, but fact-checkers found the distinction between *derogatory extreme speech* and *exclusionary extreme speech* rather slippery and difficult to operationalize. These questions led us to clarify the definitions by listing target groups. (For *derogatory extreme speech*, we listed protected categories such as gender, caste, ethnicity, and national origin, as well as racialized categories, but also the state, legacy media, politicians and civil society representatives advocating for inclusive societies). Our objective was to capture the cultural patterns of speech forms that are seen as uncivil within specific linguistic, cultural contexts but also express diverse and ambivalent forms of political contestation, as mentioned at the beginning of this article. We did not include the state, legacy media and politicians as target groups under *exclusionary extreme speech* since this label was meant

to capture expressions that exclude marginalized and vulnerable groups. AI developers were keen to keep the labels as precise as possible, while participating fact-checkers were keen to see more target groups added to the list. After several iterations, the project has received annotated passages for the three categories in the English, Hindi, Swahili, German, and Portuguese languages from partnering fact-checkers. These fact-checkers have brought—with their keen understanding and involvement in the political discourses of the region and its lifeworlds—linguistically diverse, contextually rich datasets to the ML pipeline, allowing the automated detection of problematic online speech to acquire some degree of cultural knowledge and contextualization.

Aside from its efforts to bring contextually sensitive, inclusive datasets to ML models, AI4Dignity aims to develop a tool for fact-checkers, expanding the access to AI-related technological resources for communities who are actively involved in grounding digital discourse in democratic values in different regions of the world. AI4Dignity's collaborative process model and policy engagements around AI-assisted content moderation have directly emerged from the extreme speech framework and its emphasis on comparative ethnographic excavations of the complex politics surrounding online speech.

Thus, as a critical framework, extreme speech offers methodological, policy, and theoretical perspectives rooted in ethnographic sensibility and historical awareness, toward envisioning a (digital) world of dignity.

Sahana Udupa is Professor of Media Anthropology at the University of Munich, Germany. https://orcid.org/0000-0003-3647-9570

Acknowledgments

References

Aguilera-Carnerero, C., & Azeez, A. H. (2016). Islamonausea, not Islamophobia: The many faces of cyber hate speech. *Journal of Arab & Muslim Media Research*, *9*(1), 21–40. https://doi.org/10.1386/jammr.9.1.21_1

Ahmed, S. (2004). *The cultural politics of emotion.* Edinburgh University Press.

Beller, J. (2017). *The fourth determination.* E-Flux. https://www.e-flux.com/journal/85/156818/the-fourth-determination/

Benesch, S. (2013). Dangerous Speech: A Proposal to Prevent Group Violence. *Dangerous Speech Project.* Available at: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf

Boromisza-Habashi, D. (2013). *Speaking hatefully: Culture, communication, and political action in Hungary.* The Pennsylvania State University Press.

Brown, W. (2019). Neoliberalism's Frankenstein: Authoritarian freedom in twenty-first century "democracies." *Critical Times, 1*(1), *60–79.* https://doi.org/10.1215/26410478-1.1.60

Brubaker, R., & Cooper, F. (2000). Beyond "identity." *Theory and Society*, *29*(1), 1–47. https://doi.org/10.1023/A:1007068714468

Couldry, N. (2012). *Media, society, world: Social theory and digital media practice.* Polity.

De Genova, N. (2010). Migration and race in Europe: The trans-Atlantic metastases of a post-colonial cancer. *European Journal of Social Theory*, *13*(3), 405–419. https://doi.org/10.1177/1368431010371767

de Seta, G. (2021). The politics of Muhei: Ethnic humor and Islamophobia on Chinese social media. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech.* Indiana University Press.

Dennis, D. (2017). Mediating claims to Buddha's birthplace and Nepali national identity. In S. Udupa & S. McDowell (Eds.), *Media as politics in South Asia* (pp. 176–189). Routledge.

Giddens, A. (1987). *The nation state and violence.* Polity Press.

Haynes, N. (2019). Writing on the walls: Discourses on Bolivian immigrants in Chilean meme humor. *International Journal of Communication*, *13*, 3122–3142.

Irvine, J. T. (1993). Insult and responsibility: Verbal abuse in a Wolof village. In J. H. Hill & J. T. Irvine (Eds.), *Responsibility and evidence in oral discourse* (pp. 105–134). Cambridge University Press.

Katiambo, D. (2021). It is incivility, not hate sppeech: Application of Laclau and Mouffe's discourse theory to analysis of non-anthropocentric agency. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech.* Indiana University Press.

Mankekar, P. (1999). *Screening culture, viewing politics: An ethnography of television, womanhood and nation in postcolonial India.* Duke University Press.

Mitchell, L. (2018). Civility and collective action: Soft speech, loud roars, and the politics of recognition. *Anthropological Theory*, *18*(2–3), 217–247. https://doi. org/10.1177/1463499618782792

Modh, Ketan. (2015). *Controlling hate speech on the Internet: The Indian perspective.* SSRN. https://ssrn.com/abstract=2783447

Nockleby, J. T. (2000). Hate speech. In W. L. Levy, K. L. Karst, & A. Winkler (Eds.), *Encyclopedia of the American Constitution* (pp. 1277–1279). Macmillan.

Ong, J. C., & Cabanes, J. V. (2018). *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.* The Newton Tech4Dev Network. https://newtontechfordev.com/wp-content/ uploads/2018/02/ARCHITECTS-OF-NETWORKED-DISINFORMATION-FULL-REPORT.pdf

Pohjonen, M., and Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication*, *11*, 1173–1191.

Pohjonen, M. (2019). A comparative approach to social media extreme speech: Online hate speech as media commentary. *International Journal of Communication*, *13*, 3088–3103.

Schaflechner, J. (2021). Blasphemy accusations as extreme speech acts in Pakistan. In S. Udupa, I. Gagliardone, & P. Hervik (Eds.), *Digital hate: The global conjuncture of extreme speech.* Indiana University Press.

Shankar, A. (2020). Primitivism and race in ethnographic film: A decolonial re-visioning. *Oxford Bibliographies in Anthropology.* Oxford University Press.

Thirangama, S., Kelly, T., & Forment, C. (2018). Introduction: Whose civility? *Anthropological Theory*, *18*(2–3), 153–174. https://doi.org/10.1177/ 1463499618780870

Treitler, V. B. (2013). *The Ethnic Project: Transforming racial fiction into ethnic factions.* Standford University Press.

Udupa, S. (2015). Abusive exchange on social media: The politics of online gaali cultures in India. Media Anthropology Network 52nd E-seminar, European Association of Social Anthropologists, July.

Udupa, S. (2017). Gaali cultures: The politics of abusive exchange on social media. *New Media and Society*, *20*(4), 1506–1522. https://doi.org/10.1177/1461444817698776

Udupa, S. (2019). Nationalism in the digital age: Fun as a metapractice of extreme speech. *International Journal of Communication*, *13*, 3143–3163.

Udupa, S. (2020). *Decoloniality and extreme speech.* Media Anthropology Network 65th E-Seminar, European Association of Social Anthropologists. https://www.easaonline.org/downloads/networks/media/65p.pdf

Udupa, S., & Dattatreyan, E.G. (2023). *Digital Unsettling: Decoloniality, Disposession, Rupture.* New York University Press.

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures. *International Journal of Communication*, *13*, 3049–3067.

Udupa, S., Gagliardone, I., & Hervik, P. (2021). *Digital hate: The global conjuncture of extreme speech.* Indiana University Press.

Udupa, S., Gagliardone, I., Deem, A., & Csuka, L. (2020). *Field of disinformation, democratic processes and conflict prevention.* Social Science Research Council. https://www.ssrc.org/publications/view/the-field-of-disinformation-democratic-processes-and-conflict-prevention-a-scan-of-the-literature/

Van der Veer, P. (2016). *The value of comparison.* Duke University Press.

Waldron, J. (2012). *The harm in hate speech.* Harvard University Press.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe. https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Langauge in Social Media* (pp. 19–26). Association for Computational Linguistics. https://www.aclweb.org/anthology/W12-2103